

source. This is the same kind of calculation that we're used to in electromagnetism. The resulting spacetime fields are given by

$$\begin{aligned} ds^2 &= f(r)^{-1} (-dt^2 + dX_1^2) + \sum_{i=2}^{25} dX_i^2 \\ B &= (f(r)^{-1} - 1) dt \wedge dX_1 , \quad e^{2\Phi} = f(r)^{-1} \end{aligned} \quad (7.23)$$

The function $f(r)$ depends only on the transverse direction $r^2 = \sum_{i=2}^{25} X_i^2$ and is given by

$$f(r) = 1 + \frac{g_s^2 N l_s^{22}}{r^{22}}$$

Here N is some constant which we will shortly demonstrate counts the number of strings which source the background. The string length scale in the solutions is $l_s = \sqrt{\alpha'}$. The function $f(r)$ has the property that it is harmonic in the space transverse to the string, meaning that it satisfies $\nabla_{\mathbf{R}^{24}}^2 f(r) = 0$ except at $r = 0$.

Let's compute the B -field charge of this solution. We do exactly what we do in electromagnetism: we integrate the total flux through a sphere which surrounds the object. The string lies along the X^1 direction so the transverse space is \mathbf{R}^{24} . We can consider a sphere \mathbf{S}^{23} at the boundary of this transverse space. We should be integrating the flux over this sphere. But what is the expression for the flux?

To see what we should do, let's look at the action for $H_{\mu\nu\rho}$ in the presence of a string source. We will use form notation since this is much cleaner and refer to $H_{\mu\nu\rho}$ simply as H_3 . Schematically, the action takes the form

$$\frac{1}{g_s^2} \int_{\mathbf{R}^{26}} H_3 \wedge {}^*H_3 + \int_{\mathbf{R}^2} B_2 = \frac{1}{g_s^2} \int_{\mathbf{R}^{26}} H_3 \wedge {}^*H_3 + g_s^2 B_2 \wedge \delta(\omega)$$

Here $\delta(\omega)$ is a delta-function source with support on the 2d worldsheet of the string. The equation of motion is

$$d{}^*H_3 \sim g_s^2 \delta(\omega)$$

From this we learn that to compute the charge of a single string we need to integrate

$$\frac{1}{g_s^2} \int_{\mathbf{S}^{23}} {}^*H_3 = 1$$

After these general comments, we now return to our solution (7.23). The above discussion was schematic and no attention was paid to factors of 2 and π . Keeping in this spirit, the flux of the solution (7.23) can be checked to be

$$\frac{1}{g_s^2} \int_{\mathbf{S}^{23}} {}^*H_3 = N$$

This is telling us that the solution (7.23) describes the background sourced by N coincident, parallel fundamental strings. Another way to check this is to compute the ADM mass per unit length of the solution: it is $NT \sim N/\alpha'$ as expected.

Note as far as the low-energy effective action is concerned, there is nothing that insists $N \in \mathbf{Z}$. This is analogous to the statement that nothing in classical Maxwell theory requires e to be quantized. However, in string theory, as in QED, we know the underlying sources of the microscopic theory and N must indeed take integer values.

Finally, notice that as $r \rightarrow 0$, the solution becomes singular. It is not to be trusted in this regime where higher order α' corrections become important.

7.4.3 Magnetic Branes

We've already seen that string theory is not just a theory of strings; there are also D-branes, defined as surfaces on which strings can end. We'll have much more to say about D-branes in Section 7.5. Here, we will consider a third kind of object that exists in string theory. It is again a brane — meaning that it is extended in some number of spacetime directions — but it is not a D-brane because the open string cannot end there. In these lectures we will call it the *magnetic brane*.

Electric and Magnetic Charges

You're probably not used to talking about magnetically charged objects in electromagnetism. Indeed, in undergraduate courses we usually don't get much further than pointing out that $\nabla \cdot \vec{B} = 0$ does not allow point-like magnetic charges. However, in the context of quantum field theory, much of the interesting behaviour often boils down to understanding how magnetic charges behave. And the same is true of string theory. Because this may be unfamiliar, let's take a minute to discuss the basics.

In electromagnetism in $d = 3 + 1$ dimensions, we measure electric charge q by integrating the electric field \vec{E} over a sphere \mathbf{S}^2 that surrounds the particle,

$$q = \int_{\mathbf{S}^2} \vec{E} \cdot d\vec{S} = \int_{\mathbf{S}^2} {}^*F_2 \quad (7.24)$$

In the second equality we have introduced the notation of differential forms that we also used in the previous example to discuss the string solutions.

Suppose now that a particle carries magnetic charge g . This can be measured by integrating the magnetic field \vec{B} over the same sphere. This means

$$g = \int_{\mathbf{S}^2} \vec{B} \cdot d\vec{S} = \int_{\mathbf{S}^2} F_2 \quad (7.25)$$

In $d = 3+1$ dimensions, both electrically and magnetically charged objects are particles. But this is not always true in any dimension! The reason that it holds in $4d$ is because both the field strength F_2 and the dual field strength *F_2 are 2-forms. Clearly, this is rather special to four dimensions.

In general, suppose that we have a p -brane that is electrically charged under a suitable gauge field. As we discussed in Section 7.2.1, a $(p+1)$ -dimensional object naturally couples to a $(p+1)$ -form gauge potential C_{p+1} through,

$$\mu \int_W C_{p+1}$$

where μ is the charge of the object, while W is the worldvolume of the brane. The $(p+1)$ -form gauge potential has a $(p+2)$ -form field strength

$$G_{p+2} = dC_{p+1}$$

To measure the electric charge of the p -brane, we need to integrate the field strength over a sphere that completely surrounds the object. A p -brane in D -dimensions has a transverse space \mathbf{R}^{D-p-1} . We can integrate the flux over the sphere at infinity, which is \mathbf{S}^{D-p-2} . And, indeed, the counting works out nicely because, in D dimensions, the dual field strength is a $(D-p-2)$ -form, ${}^*G_{p+2} = \tilde{G}_{D-p-2}$, which we can happily integrate over the sphere to find the charge sitting inside,

$$q = \int_{\mathbf{S}^{D-p-2}} {}^*G_{p+2}$$

This equation is the generalized version of (7.24)

Now let's think about magnetic charges. The generalized version of (7.25) suggest that we should compute the magnetic charge by integrating G_{p+2} over a sphere \mathbf{S}^{p+2} . What kind of object sits inside this sphere to emit the magnetic charge? Doing the sums backwards, we see that it should be a $(D-p-4)$ -brane.

We can write down the coupling between the $(D-p-4)$ -brane and the field strength. To do so, we first need to introduce the magnetic gauge potential defined by

$${}^*G_{p+2} = \tilde{G}_{D-p-2} = d\tilde{C}_{D-p-3} \tag{7.26}$$

We can then add the magnetic coupling to the worldvolume \tilde{W} of a $(D-p-4)$ -brane simply by writing

$$\tilde{\mu} \int_{\tilde{W}} \tilde{C}_{D-p-3}$$

where $\tilde{\mu}$ is the magnetic charge. Note that it's typically not possible to write down a Lagrangian that includes both magnetically charged object and electrically charged objects at the same time. This would need us to include both C_{p+1} and \tilde{C}_{D-p-3} in the Lagrangian, but these are not independent fields: they're related by the rather complicated differential equations (7.26).

The Magnetic Brane in Bosonic String Theory

After these generalities, let's see what it means for the bosonic string. The fundamental string is a 1-brane and, as we saw in Section 7.2.1, carries electric charge under the 2-form B . The appropriate object carrying magnetic charge under B is therefore a $(D - p - 4) = (26 - 1 - 4) = 21$ -brane.

To stress a point: neither the fundamental string, nor the magnetic 21-brane are D-branes. They are not surfaces where strings can end. We are calling them *branes* only because they are extended objects.

The magnetic 21-brane of the bosonic string can be found as a solution to the low-energy equations of motion. The solution can be written in terms of the dual potential \tilde{B}_{22} such that $d\tilde{B}_{22} = {}^*dB_2$. It is

$$\begin{aligned} ds^2 &= \left(-dt^2 + \sum_{i=1}^{21} dX_i^2 \right) + h(r) (dX_{22}^2 + \dots + dX_{25}^2) \\ \tilde{B}_{22} &= (1 - h(r)^{-2}) dt \wedge dX_1 \wedge \dots \wedge dX_{21} \\ e^{2\Phi} &= h(r) \end{aligned} \tag{7.27}$$

The function $h(r)$ depends only on the radial direction in \mathbf{R}^4 transverse to the brane: $r^2 = \sum_{i=22}^{25} X_i^2$. It is a harmonic function in \mathbf{R}^4 , given by

$$h(r) = 1 + \frac{Nl_s^2}{r^2}$$

The role of this function in the metric (7.27) is to warp the transverse \mathbf{R}^4 directions. Distances get larger as you approach the brane and the origin, $r = 0$, is at infinite distance.

It can be checked that the solution carried N units of magnetic charge and has tension

$$T \sim \frac{N}{l_s^{22}} \frac{1}{g_s^2}$$

Let's summarize how the tension of different objects scale in string theory. The powers of $\alpha' = l_s^2$ are entirely fixed on dimensional grounds. (Recall that the tension is mass per spatial volume, so the tension of a p -brane has $[T_p] = p + 1$). More interesting is the dependence on the string coupling g_s . The tension of the fundamental string does not depend on g_s , while the magnetic brane scales as $1/g_s^2$. This kind of $1/g^2$ behaviour is typical of solitons in field theories. The D-branes sit between the two: their tension scales as $1/g_s$. Objects with this behaviour are somewhat rarer (although not unheard of) in field theory.

In the perturbative limit, $g_s \rightarrow 0$, both D-branes and magnetic branes are heavy. The coupling of an object with tension T to gravity is governed by $T\kappa^2$ where the gravitational coupling scales as $\kappa \sim g_s^2$ (7.20). This means that in the weak coupling limit, the gravitational backreaction of the string and D-branes can be neglected. However, the coupling of the magnetic brane to gravity is always of order one.

The Magnetic Brane in Superstring Theory

Superstring theories also have a brane magnetically charged under B . It is a $(D - p - 4) = (10 - 1 - 4) = 5$ -brane and is usually referred to as the NS5-brane. The solution in the transverse \mathbf{R}^4 again takes the form (7.27).

The NS5-brane exists in both type II and heterotic string. In many ways it is more mysterious than D-branes and its low-energy effective dynamics is still poorly understood. It is closely related to the 5-brane of M-theory.

7.4.4 Moving Away from the Critical Dimension

The beta function equations provide a new view on the critical dimension $D = 26$ of the bosonic string. To see this, let's look more closely at the dilaton beta function $\beta(\Phi)$ defined in (7.15): it takes the same form as the Weyl anomaly that we discussed back in Section 4.4.2. This means that if we consider a string propagating in $D \neq 26$ then the Weyl anomaly simply arises as the leading order term in the dilaton beta function. So let's relax the requirement of the critical dimension. The equations of motion arising from $\beta_{\mu\nu}(G)$ and $\beta_{\mu\nu}(B)$ are unchanged, while the dilaton beta function equation becomes

$$\beta(\Phi) = \frac{D - 26}{6} - \frac{\alpha'}{2} \nabla^2 \Phi + \alpha' \nabla_\mu \Phi \nabla^\mu \Phi - \frac{\alpha'}{24} H_{\mu\nu\lambda} H^{\mu\nu\lambda} = 0 \quad (7.28)$$

The low-energy effective action in string frame picks up an extra term which looks like a run-away potential for Φ ,

$$S = \frac{1}{2\kappa_0^2} \int d^D X \sqrt{-G} e^{-2\Phi} \left(\mathcal{R} - \frac{1}{12} H_{\mu\nu\lambda} H^{\mu\nu\lambda} + 4\partial_\mu \Phi \partial^\mu \Phi - \frac{2(D - 26)}{3\alpha'} \right)$$

This sounds quite exciting. Can we really get string theory living in $D = 4$ dimensions so easily? Well, yes and no. Firstly, with this extra potential term, flat D -dimensional Minkowski space no longer solves the equations of motion. This is in agreement with the analysis in Section 2 where we showed that full Lorentz invariance was preserved only in $D = 26$.

Another, technical, problem with solving the string equations of motion this way is that we're playing tree-level term off against a one-loop term. But if tree-level and one-loop terms are comparable, then typically all higher loop contributions will be as well and it is likely that we can't trust our analysis.

The Linear Dilaton CFT

In fact, there is one simple solution to (7.28) which we can trust. It is the solution to

$$\partial_\mu \Phi \partial^\mu \Phi = \frac{26 - D}{6\alpha'}$$

Recall that we're working in signature $(-, +, +, \dots)$, meaning that Φ takes a spacelike profile if $D < 26$ and a timelike profile if $D > 26$,

$$\begin{aligned} \Phi &= \sqrt{\frac{26 - D}{6\alpha'}} X^1 & D < 26 \\ \Phi &= \sqrt{\frac{D - 26}{6\alpha'}} X^0 & D > 26 \end{aligned}$$

This gives a dilaton which is linear in one direction. This can be compared to the study of the path integral for non-critical strings that we saw in 5.3.2. There are two ways of seeing the same physics.

The reason that we can trust this solution is that there is an exact CFT underlying it which we can analyze to all orders in α' . It's called, for obvious reasons, the *linear dilaton CFT*. Let's now look at this in more detail.

Firstly, consider the worldsheet action associated to the dilaton coupling. For now we'll consider an arbitrary dilaton profile $\Phi(X)$,

$$S_{\text{dilaton}} = \frac{1}{4\pi} \int d^2\sigma \sqrt{g} \Phi(X) R^{(2)} \quad (7.29)$$

Although this term vanishes on a flat worldsheet, it nonetheless changes the stress-energy tensor $T_{\alpha\beta}$ because this is defined as

$$T_{\alpha\beta} = -4\pi \left. \frac{\partial S}{\partial g^{\alpha\beta}} \right|_{g_{\alpha\beta}=\delta_{\alpha\beta}}$$

The variation of (7.29) is straightforward. Indeed, the term is akin to the Einstein-Hilbert term in general relativity but things are simpler in 2d because, for example $R_{\alpha\beta} = \frac{1}{2} g_{\alpha\beta} R$. We have

$$\delta(\sqrt{g} g^{\alpha\beta} R_{\alpha\beta}) = \sqrt{g} g^{\alpha\beta} \delta R_{\alpha\beta} = \sqrt{g} \nabla^\alpha v_\alpha$$

where

$$v_\alpha = \nabla^\beta \delta g_{\alpha\beta} - g^{\gamma\delta} \nabla_\alpha \delta g_{\gamma\delta}$$

Using this, the variation of the dilaton term in the action is given by

$$\delta S_{\text{dilaton}} = \frac{1}{4\pi} \int d^2\sigma \sqrt{g} (\nabla^\alpha \nabla^\beta \Phi - \nabla^2 \Phi g^{\alpha\beta}) \delta g_{\alpha\beta}$$

which, restricting to flat space $g_{\alpha\beta} = \delta_{\alpha\beta}$, finally gives us the stress-energy tensor of a theory with dilaton coupling

$$T_{\alpha\beta}^{\text{dilaton}} = -\partial_\alpha \partial_\beta \Phi + \partial^2 \Phi \delta_{\alpha\beta}$$

Note that this stress tensor is not traceless. This is to be expected because, as we described above, the dilaton coupling is not Weyl invariant at tree-level. In complex coordinates, the stress tensor is

$$T^{\text{dilaton}} = -\partial^2 \Phi \quad , \quad \bar{T}^{\text{dilaton}} = -\bar{\partial}^2 \Phi$$

Linear Dilaton OPE

The stress tensor above holds for any dilaton profile $\Phi(X)$. Let's now restrict to a linear dilaton profile for a single scalar field X ,

$$\Phi = QX$$

where Q is some constant. We also include the standard kinetic terms for D scalar fields, of which X is a chosen one, giving the stress tensor

$$T = -\frac{1}{\alpha'} : \partial X \partial X : -Q \partial^2 X$$

It is a simple matter to compute the TT OPE using the techniques described in Section 4. We find,

$$T(z) T(w) = \frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots$$

where the central charge of the theory is given by

$$c = D + 6\alpha' Q^2$$

Note that Q^2 can be positive or negative depending on the whether we have a timelike or spacelike linear dilaton. In this way, we see explicitly how a linear dilaton gradient can absorb central charge.

7.4.5 The Elephant in the Room: The Tachyon

We've been waxing lyrical about the details of solutions to the low-energy effective action, all the while ignoring the most important, relevant field of them all: the tachyon. Since our vacuum is unstable, this is a little like describing all the beautiful pictures we could paint if only that damn paintbrush would balance, unaided, on its tip.

Of course, the main reason for discussing these solutions is that they all carry directly over to the superstring where the tachyon is absent. Nonetheless, it's interesting to ask what happens if the tachyon is turned on. Its vertex operator is simply

$$V_{\text{tachyon}} \sim \int d^2\sigma \sqrt{g} e^{ip \cdot X}$$

where $p^2 = 4/\alpha'$. Piecing together a general tachyon profile $V(X)$ from these Fourier modes and exponentiating, results in a potential on the worldsheet of the string

$$S_{\text{potential}} = \int d^2\sigma \sqrt{g} \alpha' V(X)$$

This is a relevant operator for the worldsheet CFT. Whenever such a relevant operator turns on, we should follow the RG flow to the infra-red until we land on another CFT. The c-theorem tells us that $c_{IR} < c_{UV}$, but in string theory we always require $c = 26$. The deficit, at least initially, is soaked up by the dilaton in the manner described above. The end point of the tachyon RG flow for the bosonic string is not understood. It may be that there is no end point and the bosonic string simply doesn't make sense once the tachyon is turned on. Or perhaps we haven't yet understood the true ground state of the bosonic string.

7.5 D-Branes Revisited: Background Gauge Fields

Understanding the constraints of conformal invariance on the closed string backgrounds led us to Einstein's equations and the low-energy effective action in spacetime. Now we would like to do the same for the open string. We want to understand the restrictions that consistency places on the dynamics of D-branes.

We saw in Section 3 that there are two types of massless modes that arise from the quantization of an open string: scalars, corresponding to the fluctuation of the D-brane, and a $U(1)$ gauge field. We will ignore the scalar fluctuations for now, but will return to them later. We focus initially on the dynamics of a gauge field A_a , $a = 0, \dots, p$ living on a D p -brane

The first question that we ask is: how does the end of the string react to a background gauge field? To answer this, we need to look at the vertex operator associated to the photon. It was given in (5.10)

$$V_{\text{photon}} \sim \int_{\partial\mathcal{M}} d\tau \zeta_a \partial^\tau X^a e^{ip \cdot X}$$

which is Weyl invariant and primary only if $p^2 = 0$ and $p^a \zeta_a = 0$. Exponentiating this vertex operator, as described at the beginning of Section 7, gives the coupling of the open string to a general background gauge field $A_a(X)$,

$$S_{\text{end-point}} = \int_{\partial\mathcal{M}} d\tau A_a(X) \frac{dX^a}{d\tau}$$

But this is a very familiar coupling — we've already mentioned it in (7.9). It is telling us that the end of the string is charged under the background gauge field A_a on the brane.

7.5.1 The Beta Function

We can now perform the same type of beta function calculation that we saw for the closed string⁹. To do this, it's useful to first use conformal invariance to map the open string worldsheet to the Euclidean upper-half plane as we described in Section 4.7. The action describing an open string propagating in flat space, with its ends subject to a background gauge field on the D-brane splits up into two pieces

$$S = S_{\text{Neumann}} + S_{\text{Dirichlet}}$$

where S_{Neumann} describes the fluctuations parallel to the Dp-brane and is given by

$$S_{\text{Neumann}} = \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \partial^\alpha X^a \partial_\alpha X^b \delta_{ab} + i \int_{\partial\mathcal{M}} d\tau A_a(X) \dot{X}^a \quad (7.30)$$

Here $a, b = 0, \dots, p$. The extra factor of i arises because we are in Euclidean space. Meanwhile, the fields transverse to the brane have Dirichlet boundary conditions and take range $I = p+1, \dots, D-1$. Their dynamics is given by

$$S_{\text{Dirichlet}} = \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \partial^\alpha X^I \partial_\alpha X^J \delta_{IJ}$$

⁹We'll be fairly explicit here, but if you want to see more details then the best place to look is the original paper by Abouelsaood, Callan, Nappi and Yost, “*Open Strings in Background Gauge Fields*”, Nucl. Phys. B280 (1987) 599.

The action $S_{\text{Dirichlet}}$ describes free fields and doesn't play any role in the computation of the beta-function. The interesting part is S_{Neumann} which, for non-zero $A_a(X)$, is an interacting quantum field theory with boundary. Our task is to compute the beta function associated to the coupling $A_a(X)$. We use the same kind of technique that we earlier applied to the closed string. We expand the fields $X^a(\sigma)$ as

$$X^a(\sigma) = \bar{x}^a(\sigma) + \sqrt{\alpha'} Y^a(\sigma)$$

where $\bar{x}^a(\sigma)$ is taken to be some fixed background which obeys the classical equations of motion,

$$\partial^2 \bar{x}^a = 0$$

(In the analogous calculation for the closed string we chose the special case of \bar{x}^a constant. Here we are more general). However, we also need to impose boundary conditions for this classical solution. In the absence of the gauge field A_a , we require Neumann boundary conditions $\partial_\sigma X^a = 0$ at $\sigma = 0$. However, the presence of the gauge field changes this. Varying the full action (7.30) shows that the relevant boundary condition is supplemented by an extra term,

$$\partial_\sigma \bar{x}^a + 2\pi\alpha' i F^{ab} \partial_\tau \bar{x}_b = 0 \quad \text{at } \sigma = 0 \quad (7.31)$$

where the F_{ab} is the field strength

$$F_{ab}(X) = \frac{\partial A_b}{\partial X^a} - \frac{\partial A_a}{\partial X^b} \equiv \partial_a A_b - \partial_b A_a$$

The fields $Y^a(\sigma)$ are the fluctuations which are taken to be small. Again, the presence of $\sqrt{\alpha'}$ in the expansion ensures that Y^a are dimensionless. Expanding the action S_{Neumann} (which we'll just call S from now on) to second order in fluctuations gives,

$$\begin{aligned} S[\bar{x} + \sqrt{\alpha'} Y] &= S[\bar{x}] + \frac{1}{4\pi} \int_{\mathcal{M}} d^2\sigma \partial Y^a \partial Y^b \delta_{ab} \\ &\quad + i\alpha' \int_{\partial\mathcal{M}} d\tau \left(\partial_a A_b Y^a \dot{Y}^b + \frac{1}{2} \partial_a \partial_b A_c Y^a Y^b \dot{x}^c \right) + \dots \end{aligned}$$

where all expressions involving the background gauge fields are now evaluated on the classical solution \bar{x} . We can rearrange the boundary terms by splitting the first term up into two halves and integrating one of these pieces by parts,

$$\int d\tau (\partial_a A_b) Y^a \dot{Y}^b = \frac{1}{2} \int d\tau \partial_a A_b Y^a \dot{Y}^b - \partial_a A_b \dot{Y}^a Y^b - \partial_c \partial_a A_b Y^a Y^b \dot{x}^c$$

Combining this with the second term means that we can write all interactions in terms of the gauge invariant field strength F_{ab} ,

$$\begin{aligned} S[\bar{x} + \sqrt{\alpha'} Y] &= S[\bar{x}] + \frac{1}{4\pi} \int_{\mathcal{M}} d^2\sigma \partial Y^a \partial Y^b \delta_{ab} \\ &\quad + \frac{i\alpha'}{2} \int_{\partial\mathcal{M}} d\tau \left(F_{ab} Y^a \dot{Y}^b + \partial_b F_{ac} Y^a Y^b \dot{x}^c \right) + \dots \end{aligned} \quad (7.32)$$

where the $+ \dots$ refer to the higher terms in the expansion which come with higher derivatives of F_{ab} , accompanied by powers of α' . We can neglect them for the purposes of computing the one-loop beta function.

The Propagator

This Lagrangian describes our interacting boundary theory to leading order. We can now use this to compute the beta function. Firstly, we should determine where possible divergences arise. The offending term is the last one in (7.32). This will lead to a divergence when the fluctuation fields Y^a are contracted with their propagator

$$\langle Y^a(z, \bar{z}) Y^b(w, \bar{w}) \rangle = G^{ab}(z, \bar{z}; w, \bar{w})$$

We should be used to these free field Green's functions by now. The propagator satisfies

$$\partial \bar{\partial} G^{ab}(z, \bar{z}) = -2\pi \delta^{ab} \delta(z, \bar{z}) \quad (7.33)$$

in the upper half plane. But now there's a subtlety. The Y^a fields need to satisfy a boundary condition at $\text{Im } z = 0$ and this should be reflected in the boundary condition for the propagator. We discussed this briefly for Neumann boundary conditions in Section 4.7. But we've also seen that the background field strength shifts the Neumann boundary conditions to (7.31). Correspondingly, the propagator $G(z, \bar{z}; w, \bar{w})$ must now satisfy

$$\partial_\sigma G^{ab}(z, \bar{z}; w, \bar{w}) + 2\pi\alpha' i F_c^a \partial_\tau G^{cb}(z, \bar{z}; w, \bar{w}) = 0 \quad \text{at } \sigma = 0 \quad (7.34)$$

In Section 4.7, we showed how Neumann boundary conditions could be imposed by considering an image charge in the lower half plane. A similar method works here. We extend $G^{ab} \equiv G^{ab}(z, \bar{z}; w, \bar{w})$ to the entire complex plane. The solution to (7.33) subject to (7.34) is given by

$$G^{ab} = -\delta^{ab} \ln |z - w| - \frac{1}{2} \left(\frac{1 - 2\pi\alpha' F}{1 + 2\pi\alpha' F} \right)^{ab} \ln(z - \bar{w}) - \frac{1}{2} \left(\frac{1 + 2\pi\alpha' F}{1 - 2\pi\alpha' F} \right)^{ab} \ln(\bar{z} - w)$$

The Counterterm and Beta Function

Let's now return to the interacting theory (7.32) and see what counterterm is needed to remove the divergence. Since all interactions take place on the boundary, we should evaluate our propagator on the boundary, which means $z = \bar{z}$ and $w = \bar{w}$. In this case, all the logarithms become the same and, in the limit that $z \rightarrow w$, gives the leading divergence $\ln|z - w| \rightarrow \epsilon^{-1}$. We learn that the UV divergence takes the form,

$$-\frac{1}{\epsilon} \left[\delta^{ab} + \frac{1}{2} \left(\frac{1 - 2\pi\alpha' F}{1 + 2\pi\alpha' F} \right)^{ab} + \frac{1}{2} \left(\frac{1 + 2\pi\alpha' F}{1 - 2\pi\alpha' F} \right)^{ab} \right] = -\frac{2}{\epsilon} \left(\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right)^{ab}$$

It's now easy to determine the necessary counterterm. We simply replace $Y^a Y^b$ in the final term with $\langle Y^a Y^b \rangle$. This yields

$$-\frac{i2\pi\alpha'^2}{\epsilon} \int_{\partial\mathcal{M}} d\tau \partial_b F_{ac} \left[\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right]^{ab} \dot{x}^c$$

For the open string theory to retain conformal invariance, we need the associated beta function to vanish. This gives us the condition on the field strength F_{ab} : it must satisfy the equation

$$\partial_b F_{ac} \left[\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right]^{ab} = 0 \quad (7.35)$$

This is our final equation governing the equations of motion that F_{ab} must satisfy to provide a consistent background for open string propagation.

7.5.2 The Born-Infeld Action

Equation (7.35) probably doesn't look too familiar! Following the path we took for the closed string, we wish to write down an action whose equations of motion coincide with (7.35). The relevant action was actually constructed many decades ago as a non-linear alternative to Maxwell theory: it goes by the name of the *Born-Infeld action*:

$$S = -T_p \int d^{p+1}\xi \sqrt{-\det(\eta_{ab} + 2\pi\alpha' F_{ab})} \quad (7.36)$$

Here ξ are the worldvolume coordinates on the brane and T_p is the tension of the D p -brane (which, since it multiplies the action, doesn't affect the equations of motion). The gauge potential is to be thought of as a function of the worldvolume coordinates: $A_a = A_a(\xi)$. It actually takes a little work to show that the equations of motion that we derive from this action coincide with the vanishing of the beta function (7.35). Some hints on how to proceed are provided on Example Sheet 4.

For small field strengths, $F_{ab} \ll 1/\alpha'$, the action (7.36) coincides with Maxwell's action. To see this, we need simply expand to get

$$S = -T_p \int d^{p+1}\xi \left(1 + \frac{(2\pi\alpha')^2}{4} F_{ab}F^{ab} + \dots \right)$$

The leading order term, quadratic in field strengths, is the Maxwell action. Terms with higher powers of F_{ab} are suppressed by powers of α' .

So, for small field strengths, the dynamics of the gauge field on a D-brane is governed by Maxwell's equations. However, as the electric and magnetic field strengths increase and become of order $1/\alpha'$, non-linear corrections to the dynamics kick in and are captured by the Born-Infeld action.

The Born-Infeld action arises from the one-loop beta function. It is the exact result for constant field strengths. If we want to understand the dynamics of gauge fields with large gradients, ∂F , then we will have determine the higher loop contributions to the beta function.

7.6 The DBI Action

We've understood that the dynamics of gauge fields on the brane is governed by the Born-Infeld action. But what about the fluctuations of the brane itself. We looked at this briefly in Section 3.2 and suggested, on general grounds, that the action should take the Dirac form (3.6). It would be nice to show this directly by considering the beta function equations for the scalar fields ϕ^I on the brane. Turning these on corresponds to considering boundary conditions where the brane is bent. It is indeed possible to compute something along the lines of beta-function equations and to show directly that the fluctuations of the brane are governed by the Dirac action¹⁰.

More generally, one could consider both the dynamics of the gauge field and the fluctuation of the brane. This is governed by a mixture of the Dirac action and the Born-Infeld action which is usually referred to as the *DBI action*,

$$S_{DBI} = -T_p \int d^{p+1}\xi \sqrt{-\det(\gamma_{ab} + 2\pi\alpha' F_{ab})}$$

As in Section (3.2), γ_{ab} is the pull-back of the spacetime metric onto the worldvolume,

$$\gamma_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} \eta_{\mu\nu}$$

¹⁰A readable discussion of this calculation can be found in the original paper by Leigh, *Dirac-Born-Infeld Action from Dirichlet Sigma Model*, Mod. Phys. Lett. A4: 2767 (1989).

The new dynamical fields in this action are the embedding coordinates $X^\mu(\xi)$, with $\mu = 0, \dots, D - 1$. This appears to be D new degrees of freedom while we expect only $D - p - 1$ transverse physical degrees of freedom. The resolution to this should be familiar by now: the DBI action enjoys a reparameterization invariance which removes the longitudinal fluctuations of the brane.

We can use this reparameterization invariance to work in static gauge. For an infinite, flat D p -brane, it is useful to set

$$X^a = \xi^a \quad a = 0, \dots, p$$

so that the pull-back metric depends only on the transverse fluctuations X^I ,

$$\gamma_{ab} = \eta_{ab} + \frac{\partial X^I}{\partial \xi^a} \frac{\partial X^J}{\partial \xi^b} \delta_{IJ}$$

If we are interested in situations with small field strengths F_{ab} and small derivatives $\partial_a X$, then we can expand the DBI action to leading order. We have

$$S = -(2\pi\alpha')^2 T_p \int d^{p+1}\xi \left(\frac{1}{4} F_{ab} F^{ab} + \frac{1}{2} \partial_a \phi^I \partial^a \phi^I + \dots \right)$$

where we have rescaled the positions to define the scalar fields $\phi^I = X^I / 2\pi\alpha'$. We have also dropped an overall constant term in the action. This is simply free Maxwell theory coupled to free massless scalar fields ϕ^I . The higher order terms that we have dropped are all suppressed by powers of α' .

7.6.1 Coupling to Closed String Fields

The DBI action describes the low-energy dynamics of a D p -brane in flat space. We could now ask how the motion of the D-brane is affected if it moves in a background created by closed string modes $G_{\mu\nu}$, $B_{\mu\nu}$ and Φ . Rather than derive this, we'll simply write down the answer and then justify each term in turn. The answer is:

$$S_{DBI} = -T_p \int d^{p+1}\xi e^{-\tilde{\Phi}} \sqrt{-\det(\gamma_{ab} + 2\pi\alpha' F_{ab} + B_{ab})}$$

Let's start with the coupling to the background metric $G_{\mu\nu}$. It's actually hidden in the notation in this expression: it appears in the pull-back metric γ_{ab} which is now given by

$$\gamma_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} G_{\mu\nu}$$

It should be clear that this is indeed the natural place for it to sit.

Next up is the dilaton. As in (7.17), we have decomposed the dilaton into a constant piece and a varying piece: $\Phi = \Phi_0 + \tilde{\Phi}$. The constant piece governs the asymptotic string coupling, $g_s = e^{\Phi_0}$, and is implicitly sitting in front of the action because the tension of the D-brane scales as

$$T_p \sim 1/g_s$$

This, then, explains the factor of $e^{-\tilde{\Phi}}$ in front of the action: it simply reunites the varying part of the dilaton with the constant piece. Physically, it's telling us that the tension of the D-brane depends on the local value of the dilaton field, rather than its asymptotic value. If the dilaton varies, the effective string coupling at a point X in spacetime is given by $g_s^{eff} = e^{\Phi(X)} = g_s e^{\tilde{\Phi}(X)}$. This, in turn, changes the tension of the D-brane. It can lower its tension by moving to regions with larger g_s^{eff} .

Finally, let's turn to the $B_{\mu\nu}$ field. This is a 2-form in spacetime. The function B_{ab} appearing in the DBI action is the pull-back to the worldvolume

$$B_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} B_{\mu\nu}$$

Its appearance in the DBI action is actually required on grounds of gauge invariance alone. This can be seen by considering an open string, moving in the presence of both a background $B_{\mu\nu}(X)$ in spacetime and a background $A_a(X)$ on the worldvolume of a brane. The relevant terms on the string worldsheet are

$$\frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu B_{\mu\nu} + \int_{\partial\mathcal{M}} d\tau A_a \dot{X}^a$$

Under a spacetime gauge transformation

$$B_{\mu\nu} \rightarrow B_{\mu\nu} + \partial_\mu C_\nu - \partial_\nu C_\mu \quad (7.37)$$

the first term changes by a total derivative. This is fine for a closed string, but it doesn't leave the action invariant for an open string because we pick up the boundary term. Let's quickly look at what we get in more detail. Under the gauge transformation (7.37), we have

$$\begin{aligned} S_B &= \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu B_{\mu\nu} \\ &\longrightarrow S_B + \frac{1}{2\pi\alpha'} \int_{\mathcal{M}} d\sigma d\tau \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu \partial_\mu C_\nu \\ &= S_B + \frac{1}{2\pi\alpha'} \int_{\mathcal{M}} d\sigma d\tau \epsilon^{\alpha\beta} \partial_\alpha (\partial_\beta X^\nu C_\nu) \\ &= S_B + \frac{1}{2\pi\alpha'} \int_{\partial\mathcal{M}} d\tau \dot{X}^\nu C_\nu = S_B + \frac{1}{2\pi\alpha'} \int_{\partial\mathcal{M}} d\tau \dot{X}^a C_a \end{aligned}$$

where, in the last line, we have replaced the sum over all directions X^ν with the sum over those directions obeying Neumann boundary conditions X^a , since $\dot{X}^I = 0$ at the end-points for any directions with Dirichlet boundary conditions.

The result of this short calculation is to see that the string action is not invariant under (7.37). To restore this spacetime gauge invariance, this boundary contribution must be canceled by an appropriate shift of A_a in the second term,

$$A_a \rightarrow A_a - \frac{1}{2\pi\alpha'} C_a \quad (7.38)$$

Note that this is not the usual kind of gauge transformation that we consider in electrodynamics. In particular, the field strength F_{ab} is not invariant. Rather, the gauge invariant combination under (7.37) and (7.38) is

$$B_{ab} + 2\pi\alpha' F_{ab}$$

This is the reason that this combination must appear in the DBI action. This is also related to an important physical effect. We have already seen that the string in spacetime is charged under $B_{\mu\nu}$. But we've also seen that the end of the string is charged under the gauge field A_a on the D-brane. This means that the open string deposits B charge on the brane, where it is converted into A charge. The fact that the gauge invariant field strength involves a combination of both F_{ab} and B_{ab} is related to this interplay of charges.

7.7 The Yang-Mills Action

Finally, let's consider the case of N coincident D-branes. We discussed this in Section 3.3 where we showed that the massless fields on the brane could be naturally packaged as $N \times N$ Hermitian matrices, with the element of the matrix telling us which brane the end points terminate on. The gauge field then takes the form

$$(A_a)_n^m$$

with $a = 0, \dots, p$ and $m, n = 1, \dots, N$. Written this way, it looks rather like a $U(N)$ gauge connection. Indeed, this is the correct interpretation. But how do we see this? Why is the gauge field describing a $U(N)$ gauge symmetry rather than, say, $U(1)^{N^2}$?

The quickest way to see that coincident branes give rise to a $U(N)$ gauge symmetry is to recall that the end point of the string is charged under the $U(1)$ gauge field that inhabits the brane it's ending on. Let's illustrate this with the simplest example. Suppose that we have two branes. The diagonal components $(A_a)_1^1$ and $(A_a)_2^2$ arise

from strings which begin and end on the same brane. Each is a $U(1)$ gauge field. What about the off-diagonal terms $(A_a)_2^1$ and $(A_a)_1^2$? These come from strings stretched between the two branes. They are again massless gauge bosons, but they are charged under the two original $U(1)$ symmetries; they carry charge $(+1, -1)$ and $(-1, +1)$ respectively. But this is precisely the structure of a $U(2)$ gauge theory, with the off-diagonal terms playing a role similar to W-bosons. In fact, the only way to make sense of massless, charged spin 1 particles is through non-Abelian gauge symmetry.

So the massless excitations of N coincident branes are a $U(N)$ gauge field $(A_a)_n^m$, together with scalars $(\phi^I)_n^m$ which transform in the adjoint representation of the $U(N)$ gauge group. We saw in Section 3 that the diagonal components $(\phi^I)_m^m$ have the interpretation of the transverse fluctuations of the m^{th} brane. Can we now write down an action describing the interactions of these fields?

In fact, there are several subtleties in writing down a non-Abelian generalization of the DBI action and such an action is not known (if, indeed, it makes sense at all). However, we can make progress by considering the low-energy limit, corresponding to small field strengths. The field strength in question is now the appropriate non-Abelian expression which, neglecting the matrix indices, reads

$$F_{ab} = \partial_a A_b - \partial_b A_a + i[A_a, A_b]$$

The low-energy action describing the dynamics of N coincident Dp-branes can be shown to be (neglecting an overall constant term),

$$S = -(2\pi\alpha')^2 T_p \int d^{p+1}\xi \text{ Tr} \left(\frac{1}{4} F_{ab} F^{ab} + \frac{1}{2} \mathcal{D}_a \phi^I \mathcal{D}^a \phi^I - \frac{1}{4} \sum_{I \neq J} [\phi^I, \phi^J]^2 \right) \quad (7.39)$$

We recognize the first term as the $U(N)$ Yang-Mills action. The coefficient in front of the Yang-Mills action is the coupling constant $1/g_{YM}^2$. For a Dp-brane, this is given by $\alpha'^2 T_p$, or

$$g_{YM}^2 \sim l_s^{p-3} g_s$$

The kinetic term for ϕ^I simply reflects the fact that these fields transform in the adjoint representation of the gauge group,

$$\mathcal{D}_a \phi^I = \partial_a \phi^I + i[A_a, \phi^I]$$

We won't derive this action in these lectures: the first two terms basically follow from gauge invariance alone. The potential term is harder to see directly: the quick ways to derive it use T-duality or, in the case of the superstring, supersymmetry.

A flat, infinite D p -brane breaks the Lorentz group of spacetime to

$$S(1, D - 1) \rightarrow SO(1, p) \times SO(D - p - 1) \quad (7.40)$$

This unbroken group descends to the worldvolume of the D-brane where it classifies all low-energy excitations of the D-brane. The $SO(1, p)$ is simply the Lorentz group of the D-brane worldvolume. The $SO(D - p - 1)$ is a global symmetry of the D-brane theory, rotating the scalar fields ϕ^I .

The potential term in (7.39) is particularly interesting,

$$V = -\frac{1}{4} \sum_{I \neq J} \text{Tr} [\phi^I, \phi^J]^2$$

The potential is positive semi-definite. We can look at the fields that can be turned on at no cost of energy, $V = 0$. This requires that all ϕ^I commute which means that, after a suitable gauge transformation, they take the diagonal form,

$$\phi^I = \begin{pmatrix} \phi_1^I \\ \vdots \\ \phi_N^I \end{pmatrix} \quad (7.41)$$

The diagonal component ϕ_n^I describes the position of the n^{th} brane in transverse space \mathbf{R}^{D-p-1} . We still need to get the dimensions right. The scalar fields have dimension $[\phi] = 1$. The relationship to the position in space (which we mentioned before in 3.2) is

$$\vec{X}_n = 2\pi\alpha' \vec{\phi}_n \quad (7.42)$$

where we've swapped to vector notation to replace the I index.

The eigenvalues ϕ_n^I are not quite gauge invariant: there is a residual gauge symmetry — the Weyl group of $U(N)$ — which leaves ϕ^I in the form (7.41) but permutes the entries by S_N , the permutation group of N elements. But this has a very natural interpretation: it is simply telling us that the D-branes are indistinguishable objects.

When all branes are separated, the vacuum expectation value (7.41) breaks the gauge group from $U(N) \rightarrow U(1)^N$. The W-bosons gain a mass M_W through the Higgs mechanism. Let's compute this mass. We'll consider a $U(2)$ theory and we'll separate

the two D-branes in the direction $X^D \equiv X$. This means that we turn on a vacuum expectation value for $\phi^D = \phi$, which we write as

$$\phi = \begin{pmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{pmatrix} \quad (7.43)$$

The values of ϕ_1 and ϕ_2 are the positions of the first and second brane. Or, more precisely, we need to multiply by the conversion factor $2\pi\alpha'$ as in (7.42) to get the position X_m of the $m = 1^{\text{st}}, 2^{\text{nd}}$ brane,

Let's compute the mass of the W-boson from the Yang-Mills action (7.39). It comes from the covariant derivative terms $\mathcal{D}\phi$. We expand out the gauge field as

$$A_a = \begin{pmatrix} A_a^{11} & W_a \\ W_a^\dagger & A_a^{22} \end{pmatrix}$$

with A^{11} and A^{22} describing the two $U(1)$ gauge fields and W the W-boson. The mass of the W-boson comes from the $[A_a, \phi]$ term inside the covariant derivative which, using the expectation value (7.43), is given by

$$\frac{1}{2} \text{Tr} [A_a, \phi]^2 = -(\phi_2 - \phi_1)^2 |W_a|^2$$

This gives us the mass of the W-boson: it is

$$M_W^2 = (\phi_2 - \phi_1)^2 = T^2 |X_2 - X_1|^2$$

where $T = 1/2\pi\alpha'$ is the tension of the string. But this has a very natural interpretation. It is precisely the mass of a string stretched between the two D-branes as shown in the figure above. We see that D-branes provide a natural geometric interpretation of the Higgs mechanism using adjoint scalars.

Notice that when branes are well separated, and the strings that stretch between them are heavy, their positions are described by the diagonal elements of the matrix given in (7.41). However, as the branes come closer together, these stretched strings become light and are important for the dynamics of the branes. Now the positions of the branes should be described by the full $N \times N$ matrices, including the off-diagonal elements. In this manner, D-branes begin to see space as something non-commutative at short distances.

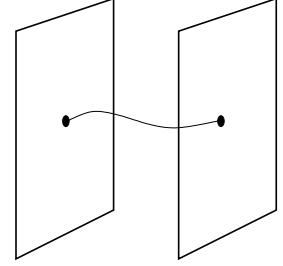


Figure 44:

In general, we can consider N D-branes located at positions \vec{X}_m , $m = 1, \dots, N$ in transverse space. The string stretched between the m^{th} and n^{th} brane has mass

$$M_W = |\vec{\phi}_n - \vec{\phi}_m| = T|\vec{X}_n - \vec{X}_m|$$

which again coincides with the mass of the appropriate W-boson computed using (7.39).

7.7.1 D-Branes in Type II Superstring Theories

As we mentioned previously, D-branes are ingredients of the Type II superstring theories. Type IIA has Dp -branes with p even, while Type IIB is home to Dp -branes with p odd. The D-branes have a very important property in these theories: they preserve half the supersymmetries.

Let's take a moment to explain what this means. We'll start by returning to the Lorentz group $SO(1, D - 1)$ now, of course, with $D = 10$. We've already seen that an infinite, flat Dp -brane is not invariant under the full Lorentz group, but only the subgroup (7.40). If we act with either $SO(1, p)$ or $SO(D - p - 1)$ then the D-brane solution remains invariant. We say that these symmetries are preserved by the solution.

However, the role of the preserved symmetries doesn't stop there. The next step is to consider small excitations of the D-brane. These must fit into representations of the preserved symmetry group (7.40). This ensures that the low-energy dynamics of the D-brane must be governed by a theory which is invariant under (7.40) and we have indeed seen that the Lagrangian (7.39) has $SO(1, p)$ as a Lorentz group and $SO(D - p - 1)$ as a global symmetry group which rotates the scalar fields.

Now let's return to supersymmetry. The Type II string theories enjoy a lot of supersymmetry: 32 supercharges in total. The infinite, flat D-branes are invariant under half of these; if we act with one half of the supersymmetry generators, the D-brane solutions don't change. Objects that have this property are often referred to as *BPS* states. Just as with the Lorentz group, these unbroken symmetries descend to the worldvolume of the D-brane. This means that the low-energy dynamics of the D-branes is described by a theory which is itself invariant under 16 supersymmetries.

There is a unique class of theories with 16 supersymmetries and a non-Abelian gauge field and matter in the adjoint representation. This class is known as maximally supersymmetric Yang-Mills theory and the bosonic part of the action is given by (7.39). Supersymmetry is realized only after the addition of fermionic fields which also live on the brane. These theories describe the low-energy dynamics of multiple D-branes.

As an illustrative example, consider D3-branes in the Type IIB theory. The theory describing N D-branes is $U(N)$ Yang-Mills with 16 supercharges, usually referred to as $U(N) \mathcal{N} = 4$ super-Yang-Mills. The bosonic part of the action is given by (7.39), where there are $D - p - 1 = 6$ scalar fields ϕ^I in the adjoint representation of the gauge group. These are augmented with four Weyl fermions, also in the adjoint representation.

8. Compactification and T-Duality

In this section, we will consider the simplest compactification of the bosonic string: a background spacetime of the form

$$\mathbf{R}^{1,24} \times \mathbf{S}^1 \quad (8.1)$$

The circle is taken to have radius R , so that the coordinate on \mathbf{S}^1 has periodicity

$$X^{25} \equiv X^{25} + 2\pi R$$

We will initially be interested in the physics at length scales $\gg R$ where motion on the \mathbf{S}^1 can be ignored. Our goal is to understand what physics looks like to an observer living in the non-compact $\mathbf{R}^{1,24}$ Minkowski space. This general idea goes by the name of *Kaluza-Klein compactification*. We will view this compactification in two ways: firstly from the perspective of the spacetime low-energy effective action and secondly from the perspective of the string worldsheet.

8.1 The View from Spacetime

Let's start with the low-energy effective action. Looking at length scales $\gg R$ means that we will take all fields to be independent of X^{25} : they are instead functions only on the non-compact $\mathbf{R}^{1,24}$.

Consider the metric in Einstein frame. This decomposes into three different fields on $\mathbf{R}^{1,24}$: a metric $\tilde{G}_{\mu\nu}$, a vector A_μ and a scalar σ which we package into the $D = 26$ dimensional metric as

$$ds^2 = \tilde{G}_{\mu\nu} dX^\mu dX^\nu + e^{2\sigma} (dX^{25} + A_\mu dX^\mu)^2 \quad (8.2)$$

Here all the indices run over the non-compact directions $\mu, \nu = 0, \dots, 24$ only.

The vector field A_μ is an honest gauge field, with the gauge symmetry descending from diffeomorphisms in $D = 26$ dimensions. To see this recall that under the transformation $\delta X^\mu = V^\mu(X)$, the metric transforms as

$$\delta G_{\mu\nu} = \nabla_\mu \Lambda_\nu + \nabla_\nu \Lambda_\mu$$

This means that diffeomorphisms of the compact direction, $\delta X^{25} = \Lambda(X^\mu)$, turn into gauge transformations of A_μ ,

$$\delta A_\mu = \partial_\mu \Lambda$$

We'd like to know how the fields $G_{\mu\nu}$, A_μ and σ interact. To determine this, we simply insert the ansatz (8.2) into the $D = 26$ Einstein-Hilbert action. The $D = 26$ Ricci scalar $\mathcal{R}^{(26)}$ is given by

$$\mathcal{R}^{(26)} = \mathcal{R} - 2e^{-\sigma}\nabla^2 e^\sigma - \frac{1}{4}e^{2\sigma}F_{\mu\nu}F^{\mu\nu}$$

where \mathcal{R} in this formula now refers to the $D = 25$ Ricci scalar. The action governing the dynamics becomes

$$S = \frac{1}{2\kappa^2} \int d^{26}X \sqrt{-\tilde{G}^{(26)}} \mathcal{R}^{(26)} = \frac{2\pi R}{2\kappa^2} \int d^{25}X \sqrt{-\tilde{G}} e^\sigma \left(\mathcal{R} - \frac{1}{4}e^{2\sigma}F_{\mu\nu}F^{\mu\nu} + \partial_\mu\sigma\partial^\mu\sigma \right)$$

The dimensional reduction of Einstein gravity in D dimensions gives Einstein gravity in $D - 1$ dimensions, coupled to a $U(1)$ gauge theory and a single massless scalar. This illustrates the original idea of Kaluza and Klein, with Maxwell theory arising naturally from higher-dimensional gravity.

The gravitational action above is not quite of the Einstein-Hilbert form. We need to again change frames, absorbing the scalar σ in the same manner as we absorbed the dilaton in Section 7.3.1. Moreover, just as for the dilaton, there is no potential dictating the vacuum expectation value of σ . Changing the vev of σ corresponds to changing R , so this is telling us that nothing in the gravitational action fixes the radius R of the compact circle. This is a problem common to all Kaluza-Klein compactifications¹¹: there are always massless scalar fields, corresponding to the volume of the internal space as well as other deformations. Massless scalar fields, such as the dilaton Φ or the volume σ , are usually referred to as *moduli*.

If we want this type of Kaluza-Klein compactification to describe our universe — where we don't see massless scalar fields — we need to find a way to “fix the moduli”. This means that we need a mechanism which gives rise to a potential for the scalar fields, making them heavy and dynamically fixing their vacuum expectation value. Such mechanisms exist in the context of the superstring.

Let's now also look at the Kaluza-Klein reduction of the other fields in the low-energy effective action. The dilaton is easy: a scalar in D dimensions reduces to a scalar in $D - 1$ dimensions. The anti-symmetric 2-form has more structure: it reduces to a 2-form $B_{\mu\nu}$, together with a vector field $\tilde{A}_\mu = B_{\mu 25}$.

¹¹The description of compactification on more general manifolds is a beautiful story involving aspects of differential geometry and topology. This story is told in the second volume of Green, Schwarz and Witten.

In summary, the low-energy physics of the bosonic string in $D - 1$ dimensions consists of a metric $G_{\mu\nu}$, two $U(1)$ gauge fields A_μ and \tilde{A}_μ and two massless scalars Φ and σ .

8.1.1 Moving around the Circle

In the above discussion, we assumed that all fields are independent of the periodic direction X^{25} . Let's now look at what happens if we relax this constraint. It's simplest to see the resulting physics if we look at the scalar field Φ where we don't have to worry about cluttering equations with indices. In general, we can expand this field in Fourier modes around the circle

$$\Phi(X^\mu; X^{25}) = \sum_{n=-\infty}^{\infty} \Phi_n(X^\mu) e^{inX^{25}/R}$$

where reality requires $\Phi_n^* = \Phi_{-n}$. Ignoring the coupling to gravity for now, the kinetic terms for this scalar are

$$\int d^{26}X \partial_\mu \Phi \partial^\mu \Phi + (\partial_{25}\Phi)^2 = 2\pi R \int d^{25}X \sum_{n=-\infty}^{\infty} \left(\partial_\mu \Phi_n \partial^\mu \Phi_{-n} + \frac{n^2}{R^2} |\Phi_n|^2 \right)$$

This simple Fourier decomposition is telling us something very important: a single scalar field on $\mathbf{R}^{1,D-1} \times \mathbf{S}^1$ splits into an infinite number of scalar fields on $\mathbf{R}^{1,D-2}$, indexed by the integer n . These have mass

$$M_n^2 = \frac{n^2}{R^2} \tag{8.3}$$

For R small, all particles are heavy except for the massless zero mode $n = 0$. The heavy particles are typically called Kaluza-Klein (KK) modes and can be ignored if we're probing energies $\ll 1/R$ or, equivalently, distance scales $\gg R$.

There is one further interesting property of the KK modes Φ_n with $n \neq 0$: they are charged under the gauge field A_μ arising from the metric. The simplest way to see this is to look at the appropriate gauge transformation which, from the spacetime perspective, is the diffeomorphism $X^{25} \rightarrow X^{25} + \Lambda(X^\mu)$. Clearly, this shifts the KK modes

$$\Phi_n \rightarrow \exp\left(\frac{in\Lambda}{R}\right) \Phi_n$$

This tells us that the n^{th} KK mode has charge n/R . In fact, one usually rescales the gauge field to $A'_\mu = A_\mu/R$, under which the charge of the KK mode Φ_n is simply $n \in \mathbf{Z}$.

8.2 The View from the Worldsheet

We now consider the Kaluza-Klein reduction from the perspective of the string. We want to study a string moving in the background $\mathbf{R}^{1,24} \times \mathbf{S}^1$. There are two ways in which the compact circle changes the string dynamics.

The first effect of the circle is that the spatial momentum, p , of the string in the circle direction can no longer take any value, but is quantized in integer units

$$p^{25} = \frac{n}{R} \quad n \in \mathbf{Z}$$

The simplest way to see this is simply to require that the string wavefunction, which includes the factor $e^{ip \cdot X}$, is single valued.

The second effect is that we can allow more general boundary conditions for the mode expansion of X . As we move around the string, we no longer need $X(\sigma + 2\pi) = X(\sigma)$, but can relax this to

$$X^{25}(\sigma + 2\pi) = X^{25}(\sigma) + 2\pi m R \quad m \in \mathbf{Z}$$

The integer m tells us how many times the string winds around \mathbf{S}^1 . It is usually simply called the *winding number*.

Let's now follow the familiar path that we described in Section 2 to study the spectrum of the string on the spacetime (8.1). We start by considering only the periodic field X^{25} , highlighting the differences with our previous treatment. The mode expansion of X^{25} is now given by

$$X^{25}(\sigma, \tau) = x^{25} + \frac{\alpha' n}{R} \tau + m R \sigma + \text{oscillator modes}$$

which incorporates both the quantized momentum and the possibility of a winding number. Before splitting $X^{25}(\sigma, \tau)$ into right-moving and left-moving parts, it will be useful to introduce the quantities

$$p_L = \frac{n}{R} + \frac{m R}{\alpha'} \quad , \quad p_R = \frac{n}{R} - \frac{m R}{\alpha'} \tag{8.4}$$

Then we have $X^{25}(\sigma, \tau) = X_L^{25}(\sigma^+) + X_R^{25}(\sigma^-)$, where

$$\begin{aligned} X_L^{25}(\sigma^+) &= \frac{1}{2} x^{25} + \frac{1}{2} \alpha' p_L \sigma^+ + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \tilde{\alpha}_n^{25} e^{-in\sigma^+} , \\ X_R^{25}(\sigma^-) &= \frac{1}{2} x^{25} + \frac{1}{2} \alpha' p_R \sigma^- + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^{25} e^{-in\sigma^-} \end{aligned}$$

This differs from the mode expansion (1.36) only in the terms p_L and p_R . The mode expansion for all the other scalar fields on flat space $\mathbf{R}^{1,24}$ remains unchanged and we don't write them explicitly.

Let's think about what the spectrum of this theory looks like to an observer living in $D = 25$ non-compact directions. Each particle state will be described by a momentum p^μ with $\mu = 0, \dots, 24$. The mass of the particle is

$$M^2 = - \sum_{\mu=0}^{24} p_\mu p^\mu$$

As before, the mass of these particles is fixed in terms of the oscillator modes of the string by the L_0 and \tilde{L}_0 equations. These now read

$$M^2 = p_L^2 + \frac{4}{\alpha'} (\tilde{N} - 1) = p_R^2 + \frac{4}{\alpha'} (N - 1)$$

where N and \tilde{N} are the levels, defined in lightcone quantization by (2.24). (One should take the lightcone coordinate inside $\mathbf{R}^{1,24}$ rather than along the \mathbf{S}^1). The factors of -1 are the necessary normal ordering coefficients that we've seen in several guises in this course.

These equations differ from (2.25) by the presence of the momentum and winding terms around \mathbf{S}^1 on the right-hand side. In particular, level matching no longer tells us that $N = \tilde{N}$, but instead

$$N - \tilde{N} = nm \tag{8.5}$$

Expanding out the mass formula, we have

$$M^2 = \frac{n^2}{R^2} + \frac{m^2 R^2}{\alpha'^2} + \frac{2}{\alpha'} (N + \tilde{N} - 2) \tag{8.6}$$

The new terms in this formula have a simple interpretation. The first term tells us that a string with $n > 0$ units of momentum around the circle gains a contribution to its mass of n/R . This agrees with the result (8.3) that we found from studying the KK reduction of the spacetime theory. The second term is even easier to understand: a string which winds $m > 0$ times around the circle picks up a contribution $2\pi m R T$ to its mass, where $T = 1/2\pi\alpha'$ is the tension of the string.

8.2.1 Massless States

We now restrict attention to the massless states in $\mathbf{R}^{1,24}$. This can be achieved in the mass formula (8.6) by looking at states with zero momentum $n = 0$ and zero winding $m = 0$, obeying the level matching condition $N = \tilde{N} = 1$. The possibilities are

- $\alpha_{-1}^\mu \tilde{\alpha}_{-1}^\nu |0; p\rangle$: Under the $SO(1, 24)$ Lorentz group, these states decompose into a metric $G_{\mu\nu}$, an anti-symmetric tensor $B_{\mu\nu}$ and a scalar Φ .
- $\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} |0; p\rangle$ and $\alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu |0; p\rangle$: These are two vector fields. We can identify the sum of these $(\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} + \alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu) |0; p\rangle$ with the vector field A_μ coming from the metric and the difference $(\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} - \alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu) |0; p\rangle$ with the vector field \tilde{A}_μ coming from the anti-symmetric field.
- $\alpha_{-1}^{25} \tilde{\alpha}_{-1}^{25} |0; p\rangle$: This is another scalar. It is identified with the scalar σ associated to the radius of \mathbf{S}^1 .

We see that the massless spectrum of the string coincides with the massless spectrum associated with the Kaluza-Klein reduction of the previous section.

8.2.2 Charged Fields

One can also check that the KK modes with $n \neq 0$ have charge n under the gauge field A_μ . We can determine the charge of a state under a given $U(1)$ by computing the 3-point function in which two legs correspond to the state of interest, while the third is the appropriate photon. We have two photons, with vertex operators given by,

$$V_\pm(p) \sim \int d^2 z \, \zeta_\mu (\partial X^\mu \bar{\partial} \bar{X}^{25} \pm \partial X^{25} \bar{\partial} \bar{X}^\mu) e^{ip \cdot X}$$

where $+$ corresponds to A_μ and $-$ to \tilde{A}_μ and we haven't been careful about the overall normalization. Meanwhile, any state can be assigned momentum n and winding m by dressing the operator with the factor $e^{ip_L X^{25}(z) + ip_R \bar{X}^{25}(\bar{z})}$. As always, it's simplest to work with the momentum and winding modes of the tachyon, whose vertex operators are of the form

$$V_{m,n}(p) \sim \int d^2 z \, e^{ip \cdot X} e^{ip_L X^{25} + ip_R \bar{X}^{25}}$$

The charge of a state is the coefficient in front of the 3-point coupling of the field and the photon,

$$\langle V_\pm(p_1) V_{m,n}(p_2) V_{-m,-n}(p_3) \rangle \sim \delta^{25} \left(\sum_i p_i \right) \zeta_\mu (p_2^\mu - p_3^\mu) (p_L \pm p_R)$$

The first few factors are merely kinematical. The interesting information is in the last factor. It is telling us that under A_μ , fields have charge $p_L + p_R \sim n/R$. This is in agreement with the Kaluza-Klein analysis that we saw before. However, it's also telling us something new: under \tilde{A}_μ , fields have charge $p_L - p_R \sim mR/\alpha'$. In other words, winding modes are charged under the gauge field that arises from the reduction of $B_{\mu\nu}$. This is not surprising: winding modes correspond to strings wrapping the circle and we saw in Section 7 that strings are electrically charged under $B_{\mu\nu}$.

8.2.3 Enhanced Gauge Symmetry

With a circle in the game, there are other ways to build massless states that don't require us to work at level $N = \tilde{N} = 1$. For example, we can set $N = \tilde{N} = 0$ and look at winding modes $m \neq 0$. The level matching condition (8.5) requires $n = 0$ and the mass of the states is

$$M^2 = \left(\frac{mR}{\alpha'} \right)^2 - \frac{4}{\alpha'}$$

and states can be massless whenever the radius takes special values $R^2 = 4\alpha'/m^2$ with $m \in \mathbf{Z}$. Similarly, we can set the winding to zero $m = 0$ and consider the KK modes of the tachyon which have mass

$$M^2 = \frac{n^2}{R^2} - \frac{4}{\alpha'}$$

which become massless when $R^2 = n^2\alpha'/4$.

However, the richest spectrum of massless states occurs when the radius takes a very special value, namely

$$R = \sqrt{\alpha'}$$

Solutions to the level matching condition (8.5) with $M^2 = 0$ are now given by

- $N = \tilde{N} = 1$ with $m = n = 0$. These give the states described above: a metric, two $U(1)$ gauge fields and two neutral scalars.
- $N = \tilde{N} = 0$ with $n = \pm 2$ and $m = 0$. These are KK modes of the tachyon field. They are scalars in spacetime with charges $(\pm 2, 0)$ under the $U(1) \times U(1)$ gauge symmetry.
- $N = \tilde{N} = 0$ with $n = 0$ and $m = \pm 2$. This is a winding mode of the tachyon field. They are scalars in spacetime with charges $(0, \pm 2)$ under $U(1) \times U(1)$.

- $N = 1$ and $\tilde{N} = 0$ with $n = m = \pm 1$. These are two new spin 1 fields, $\alpha_{-1}^\mu |0; p\rangle$. They carry charge $(\pm 1, \pm 1)$ under the two $U(1) \times U(1)$.
- $N = 1$ and $\tilde{N} = 0$ with $n = -m = \pm 1$. These are a further two spin 1 fields, $\tilde{\alpha}_{-1}^\mu |0; p\rangle$, with charge $(\pm 1, \mp 1)$ under $U(1) \times U(1)$.

How do we interpret these new massless states? Let's firstly look at the spin 1 fields. These are charged under $U(1) \times U(1)$. As we mentioned in Section 7.7, the only way to make sense of charged massless spin 1 fields is in terms of a non-Abelian gauge symmetry. Looking at the charges, we see that at the critical radius $R = \sqrt{\alpha'}$, the theory develops an enhanced gauge symmetry

$$U(1) \times U(1) \rightarrow SU(2) \times SU(2)$$

The massless scalars from the $N = \tilde{N} = 0$ now join with the previous scalars to form adjoint representations of this new symmetry. We move away from the critical radius by changing the vacuum expectation value for σ . This breaks the gauge group back to the Cartan subalgebra by the Higgs mechanism.

From the discussion above, it's clear that this mechanism for generating non-Abelian gauge symmetries relies on the existence of the tachyon. For this reason, this mechanism doesn't work in Type II superstring theories. However, it turns out that it does work in the heterotic string, even though it has no tachyon in its spectrum.

8.3 Why Big Circles are the Same as Small Circles

The formula (8.6) has a rather remarkable property: it is invariant under the exchange

$$R \leftrightarrow \frac{\alpha'}{R} \tag{8.7}$$

if, at the same time, we swap the quantum numbers

$$m \leftrightarrow n \tag{8.8}$$

This means that a string moving on a circle of radius R has the same spectrum as a string moving on a circle of radius α'/R . It achieves this feat by exchanging what it means to wind with that it means to move.

As the radius of the circle becomes large, $R \rightarrow \infty$, the winding modes become very heavy with mass $\sim R/\alpha'$ and are irrelevant for the low-energy dynamics. But the momentum modes become very light, $M \sim 1/R$, and, in the strict limit form a continuum. From the perspective of the energy spectrum, this continuum of energy states is exactly what we mean by the existence of a non-compact direction in space.

In the other limit, $R \rightarrow 0$, the momentum modes become heavy and can be ignored: it takes way too much energy to get anything to move on the \mathbf{S}^1 . In contrast, the winding modes become light and start to form a continuum. The resulting energy spectrum looks as if another dimension of space is opening up!

The equivalence of the string spectrum on circles of radii R and α'/R extends to the full conformal field theory and hence to string interactions. Strings are unable to tell the difference between circles that are very large and circles that are very small. This striking statement has a rubbish name: it is called *T-duality*.

This provides another mechanism in which string theory exhibits a minimum length scale: as you shrink a circle to smaller and smaller sizes, at $R = \sqrt{\alpha'}$, the theory acts as if the circle is growing again, with winding modes playing the role of momentum modes.

The New Direction in Spacetime

So how do we describe this strange new spatial direction that opens up as $R \rightarrow 0$? Under the exchange (8.7) and (8.8), we see that p_L and p_R transform as

$$p_L \rightarrow p_L , \quad p_R \rightarrow -p_R$$

Motivated by this, we define a new scalar field,

$$Y^{25} = X_L^{25}(\sigma^+) - X_R^{25}(\sigma^-)$$

It is simple to check that in the CFT for a free, compact scalar field all OPEs of Y^{25} coincide with the OPEs of X^{25} . This is sufficient to ensure that all interactions defined in the CFT are the same.

We can write the new spatial direction Y directly in terms of the old field X , without first doing the split into left and right-moving pieces. From the definition of Y , one can check that $\partial_\tau X = \partial_\sigma Y$ and $\partial_\sigma X = \partial_\tau Y$. We can write this in a unified way as

$$\partial_\alpha X = \epsilon_{\alpha\beta} \partial^\beta Y \tag{8.9}$$

where $\epsilon_{\alpha\beta}$ is the antisymmetric matrix with $\epsilon_{\tau\sigma} = -\epsilon_{\sigma\tau} = +1$. (The minus sign from $\epsilon_{\sigma\tau}$ in the above equation is canceled by another from the Minkowski worldsheet metric when we lower the index on ∂^β).

The Shift of the Dilaton

The dilaton, or string coupling, also transforms under T-duality. Here we won't derive this in detail, but just give a plausible explanation for why it's the case. The main idea is that a scientist living in a stringy world shouldn't be able to do any experiments that distinguish between a compact circle of radius R and one of radius α'/R . But the first place you would look is simply the low-energy effective action which, working in Einstein frame, contains terms like

$$\frac{2\pi R}{2l_s^{24}g_s^2} \int d^{25}X \sqrt{-\tilde{G}} e^\sigma \mathcal{R} + \dots$$

A scientist cannot tell the difference between R and $\tilde{R} = \alpha'/R$ only if the value of the dilaton is also ambiguous so that the term in front of the action remains invariant: i.e. $R/g_s^2 = \tilde{R}/\tilde{g}_s^2$. This means that, under T-duality, the dilaton must shift so that the coupling constant becomes

$$g_s \rightarrow \tilde{g}_s = \frac{\sqrt{\alpha'} g_s}{R} \quad (8.10)$$

8.3.1 A Path Integral Derivation of T-Duality

There's a simple way to see T-duality of the quantum theory using the path integral. We'll consider just a single periodic scalar field $X \equiv X + 2\pi R$ on the worldsheet. It's useful to change normalization and write $X = R\varphi$, so that the field φ has periodicity 2π . The radius R of the circle now sits in front of the action,

$$S[\varphi] = \frac{R^2}{4\pi\alpha'} \int d^2\sigma \partial_\alpha \varphi \partial^\alpha \varphi \quad (8.11)$$

The Euclidean partition function for this theory is $Z = \int \mathcal{D}\varphi e^{-S[\varphi]}$. We will now play around with this partition function and show that we can rewrite it in terms of new variables that describe the T-dual circle.

The theory (8.11) has a simple shift symmetry $\varphi \rightarrow \varphi + \lambda$. The first step is to make this symmetry local by introducing a gauge field A_α on the worldsheet which transforms as $A_\alpha \rightarrow A_\alpha - \partial_\alpha \lambda$. We then replace the ordinary derivatives with covariant derivatives

$$\partial_\alpha \varphi \rightarrow \mathcal{D}_\alpha \varphi = \partial_\alpha \varphi + A_\alpha$$

This changes our theory. However, we can return to the original theory by adding a new field, θ which couples as

$$S[\varphi, \theta, A] = \frac{R^2}{4\pi\alpha'} \int d^2\sigma \mathcal{D}_\alpha \varphi \mathcal{D}^\alpha \varphi + \frac{i}{2\pi} \int d^2\sigma \theta \epsilon^{\alpha\beta} \partial_\alpha A_\beta \quad (8.12)$$

The new field θ acts as a Lagrange multiplier. Integrating out θ sets $\epsilon^{\alpha\beta}\partial_\alpha A_\beta = 0$. If the worldsheet is topologically \mathbf{R}^2 , then this condition ensures that A_α is pure gauge which, in turn, means that we can pick a gauge such that $A_\alpha = 0$. The quantum theory described by (8.12) is then equivalent to that given by (8.11).

Of course, if the worldsheet is topologically \mathbf{R}^2 then we're missing the interesting physics associated to strings winding around φ . On a non-trivial worldsheet, the condition $\epsilon^{\alpha\beta}\partial_\alpha A_\beta = 0$ does not mean that A_α is pure gauge. Instead, the gauge field can have non-trivial holonomy around the cycles of the worldsheet. One can show that these holonomies are gauge trivial if θ has periodicity 2π . In this case, the partition function defined by (8.12),

$$Z = \frac{1}{\text{Vol}} \int \mathcal{D}\varphi \mathcal{D}\theta \mathcal{D}A e^{-S[\varphi, \theta, A]}$$

is equivalent to the partition function constructed from (8.11) for worldsheets of any topology.

At this stage, we make use of a clever and ubiquitous trick: we reverse the order of integration. We start by integrating out φ which we can do by simply fixing the gauge symmetry so that $\varphi = 0$. The path integral then becomes

$$Z = \int \mathcal{D}\theta \mathcal{D}A \exp \left(-\frac{R^2}{4\pi\alpha'} \int d^2\sigma A_\alpha A^\alpha + \frac{i}{2\pi} \int d^2\sigma \epsilon^{\alpha\beta} (\partial_\alpha \theta) A_\beta \right)$$

where we have also taken the opportunity to integrate the last term by parts. We can now complete the procedure and integrate out A_α . We get

$$Z = \int \mathcal{D}\theta \exp \left(-\frac{\tilde{R}^2}{4\pi\alpha'} \int d^2\sigma \partial_\alpha \theta \partial^\alpha \theta \right)$$

with $\tilde{R} = \alpha'/R$ the radius of the T-dual circle. In the final integration, we threw away the overall factor in the path integral, which is proportional to $\sqrt{\alpha'}/R$. A more careful treatment shows that this gives rise to the appropriate shift in the dilaton (8.10).

8.3.2 T-Duality for Open Strings

What happens to open strings and D-branes under T-duality? Suppose firstly that we compactify a circle in direction X transverse to the brane. This means that X has Dirichlet boundary conditions

$$X = \text{const} \Rightarrow \partial_\tau X^{25} = 0 \quad \text{at } \sigma = 0, \pi$$

But what happens in the T-dual direction Y ? From the definition (8.9) we learn that the new direction has Neumann boundary conditions,

$$\partial_\sigma Y = 0 \quad \text{at } \sigma = 0, \pi$$

We see that T-duality exchanges Neumann and Dirichlet boundary conditions. If we dualize a circle transverse to a Dp -brane, then it turns into a $D(p+1)$ -brane.

The same argument also works in reverse. We can start with a Dp -brane wrapped around the circle direction X , so that the string has Neumann boundary conditions. After T-duality, (8.9) changes these to Dirichlet boundary conditions and the Dp -brane turns into a $D(p-1)$ -brane, localized at some point on the circle Y .

In fact, this was how D-branes were originally discovered: by following the fate of open strings under T-duality.

8.3.3 T-Duality for Superstrings

To finish, let's nod one final time towards the superstring. It turns out that the ten-dimensional superstring theories are not invariant under T-duality. Instead, they map into each other. More precisely, Type IIA and IIB transform into each other under T-duality. This means that Type IIA string theory on a circle of radius R is equivalent to Type IIB string theory on a circle of radius α'/R . This dovetails with the transformation of D-branes, since type IIA has Dp -branes with p even, while IIB has p odd. Similarly, the two heterotic strings transform into each other under T-duality.

8.3.4 Mirror Symmetry

The essence of T-duality is that strings get confused. Their extended nature means that they're unable to tell the difference between big circles and small circles. We can ask whether this confusion extends to more complicated manifolds. The answer is yes. The fact that strings can see different manifolds as the same is known as *mirror symmetry*.

Mirror symmetry is cleanest to state in the context of the Type II superstring, although similar behaviour also holds for the heterotic strings. The simplest example is when the worldsheet of the string is governed by a superconformal non-linear sigma-model with target space given by some Calabi-Yau manifold \mathbf{X} . The claim of mirror symmetry is that this CFT is identical to the CFT describing the string moving on a different Calabi-Yau manifold \mathbf{Y} . The topology of \mathbf{X} and \mathbf{Y} is not the same. Their Hodge diamonds are the mirror of each other; hence the name. The subject of mirror symmetry is an active area of research in geometry and provides a good example of the impact of string theory on mathematics.

8.4 Epilogue

We are now at the end of this introductory course on string theory. We began by trying to make sense of the quantum theory of a relativistic string moving in flat space. It is, admittedly, an odd place to start. But from then on we had no choices to make. The relativistic string leads us ineluctably to conformal field theory, to higher dimensions of spacetime, to Einstein’s theory of gravity at low-energies, to good UV behaviour at high-energies and to Yang-Mills theories living on branes. There are few stories in theoretical physics where such meagre input gives rise to such a rich structure.

This journey continues. There is one further ingredient that it is necessary to add: supersymmetry. Even this is in some sense not a choice, but is necessary to remove the troublesome tachyon that plagued these lectures. From there we may again blindly follow where the string leads, through anomalies (and the lack thereof) in ten dimensions, to dualities and M-theory in eleven dimensions, to mirror symmetry and moduli stabilization and black hole entropy counting and holography and the miraculous AdS/CFT correspondence.

However, the journey is far from complete. There is much about string theory that remains to be understood. This is true both of the mathematical structure of the theory and of its relationship to the world that we observe. The problems that we alluded to in Section 6.4.5 are real. Non-perturbative completions of string theory are only known in spacetimes which are asymptotically anti-de Sitter, but cosmological observations suggest that our home is not among these. In attempts to make contact with the standard models of particle physics and cosmology, we typically return to the old idea of Kaluza-Klein compactifications. Is this the right approach? Or are we missing some important and subtle conceptual ingredient? Or is the existence of this remarkable mathematical structure called string theory merely a red-herring that has nothing to do with the real world?

In the years immediately after its birth, no one knew that string theory was a theory of strings. It seems very possible that we’re currently in a similar situation. When the theory is better understood, it may have little to do with strings. We are certainly still some way from answering the simple question: what is string theory really?

White holes and eternal black holes

Stephen D. H. Hsu*

Institute of Theoretical Science, University of Oregon, Eugene, OR 97403

(Dated: November 2011)

We investigate isolated white holes surrounded by vacuum, which correspond to the time reversal of eternal black holes that do not evaporate. We show that isolated white holes produce quasi-thermal Hawking radiation. The time reversal of this radiation, incident on a black hole precursor, constitutes a special preparation that will cause the black hole to become eternal.

What is a white hole?

White holes have received far less attention from researchers than black holes (for a review, see, e.g., [1]). This is understandable, given that conditions in our universe readily lead to black hole formation, whereas white hole creation has neither been observed nor is expected to have occurred in the history of the universe.

However, white holes are themselves fundamental objects and worthy of further study. White holes are time-reversed black holes, and therefore characterized by the same quantum numbers: mass, angular momentum, charge. While a classical black hole spacetime has a singularity in the future, a white hole has one in the past. If quantum gravitational effects can resolve black hole singularities, then white holes need not result from singular initial conditions. (In any case the initial white hole singularity is not directly visible to observers.)

Standard quantum mechanical reasoning suggests that any initial state which evolves into a black hole also has some nonzero probability to evolve into a white hole. Note we are referring to a quantum gravitational process, and are making the assumption that even in quantum gravity tunneling between two states with the same quantum numbers has non-zero (although possibly very small) probability. For example, it is known that the collision of two sufficiently energetic particles can create a black hole [2]. Because the quantum numbers are the same we would expect that the same energetic particles have a small but non-zero probability of producing a white hole with the same quantum numbers [3]. Since large black holes are long-lived, there are some white hole states (corresponding to time slices late in the black hole's existence) that persist for a long time before exploding. Thus, long-lived white holes are a consequence of quantum mechanics and the properties of black holes.

A class of highly entropic objects whose full spacetime evolution is that of a white hole which explodes outwards, is stopped by gravitational self-attraction, and recollapses to form a black hole, are described in [4].

Hawking's arguments and thermal equilibrium

In his 1976 paper *Black holes and thermodynamics* [5], Hawking analyzed the properties of white holes by considering a box in thermal equilibrium, whose temperature and volume are adjusted so that the most probable configuration is a black hole surrounded by a gas of particles whose temperature is equal to that of a black hole. The black hole emits Hawking radiation but absorbs, on average, as much energy from the gas as it emits. Applying time reversal, the configuration describes a white hole emitting and absorbing radiation. Since there is no arrow of time for a system in thermal equilibrium, Hawking argued that black and white holes must be indistinguishable. More precisely, the properties of white and black holes *in equilibrium with their surroundings* are identical. However, the same cannot be said for black and white holes in isolation (i.e., surrounded by empty vacuum)—we shall see that their properties are radically different.

In elementary particle physics we are accustomed to the idea that time reversal maps particles to their anti-particles. However, in the case of black and white holes the subsequent evolution of the time-reversed object depends on more than just quantum numbers such as M, J, Q . For a hot black hole in a cold environment, there is a statistical arrow of time.

Isolated white holes

Consider a white hole (figure 1) in a spacetime with the property that past null infinity is in the empty vacuum state of ordinary flat space. This implies that space far from the hole is empty and that there is no incoming radiation from the past. This white hole spacetime is the time reversal of a black hole spacetime with no Hawking radiation propagating to future infinity (figure 2). One motivation for considering such objects is that they are localized, as opposed to an entire spacelike slice of a box in thermal equilibrium. Do such white holes exist? What are their properties? (For simplicity, we assume all quantum numbers of the hole, other than its mass, are zero.)

In this discussion we will refer to the diagrams in figures 1 and 2, which depict a black hole spacetime (figure 2) and its time reversal (figure 1). We will refer to past and future null infinity of the black hole spacetime as $\mathcal{J}_\pm^{\text{bh}}$, where the subscript + indicates future, and - indicates past. In the white hole diagram the role of past and future are reversed: $\mathcal{J}_+^{\text{wh}} = \mathcal{J}_-^{\text{bh}}$.

Assuming that the white hole is isolated implies the

*Electronic address: hsu@uoregon.edu

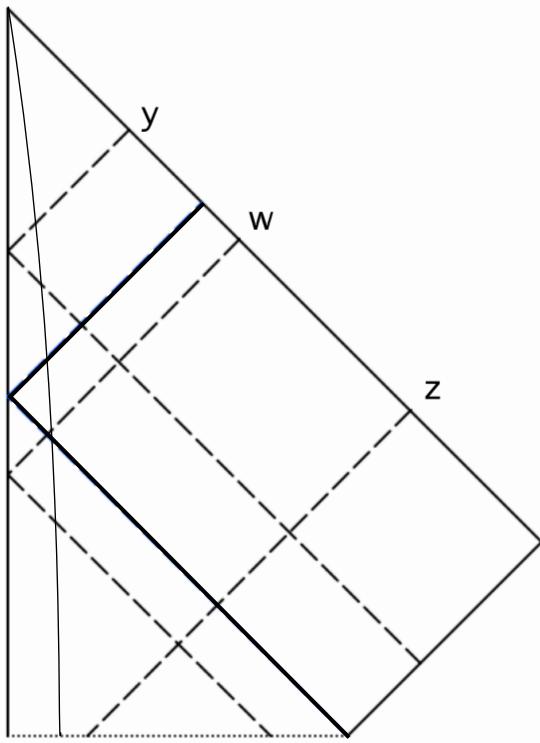


FIG. 1: A white hole spacetime. We impose the condition that past null infinity $\mathcal{J}_-^{\text{wh}}$ is in the vacuum state – there is no incoming radiation from the far past. The dotted black line is the initial singularity, and the thick solid line is the path of a null ray on the anti-horizon. The curved line indicates matter which explodes out of the hole. The dashed black lines refer to modes discussed in the text.

empty vacuum on $\mathcal{J}_-^{\text{wh}}$: there is no incoming radiation from the past.

This condition is equivalent, on the black hole spacetime, to no Hawking radiation propagating to future null infinity $\mathcal{J}_+^{\text{bh}}$. This sounds strange, but can be accomplished by proper choice of initial state from which the black hole is formed. That is, a special arrangement of incoming modes from $\mathcal{J}_-^{\text{bh}}$ is required; see below for details. In the white hole spacetime these modes would be seen exiting the white hole after it explodes from behind its anti-horizon.

In our discussion we *assume* that the black hole spacetime (figure 2) describes a progenitor (e.g., a star) which collapses to form the hole. Because ordinary stars and other progenitors in nearly-flat space obey an entropy bound: $S < A^{3/4}$, where A is their surface area in Planck units, such objects have much lower entropy than a black hole with no constraint on how it was formed [4] [6]. Indeed, a generic black hole, formed in a maximally entropic process (e.g., by allowing an initially small black hole to slowly accrete matter) has entropy of order A , but such objects do not satisfy the isolation condition imposed above. Thus, the objects under study are very exotic (improbable): not only are they white holes, but

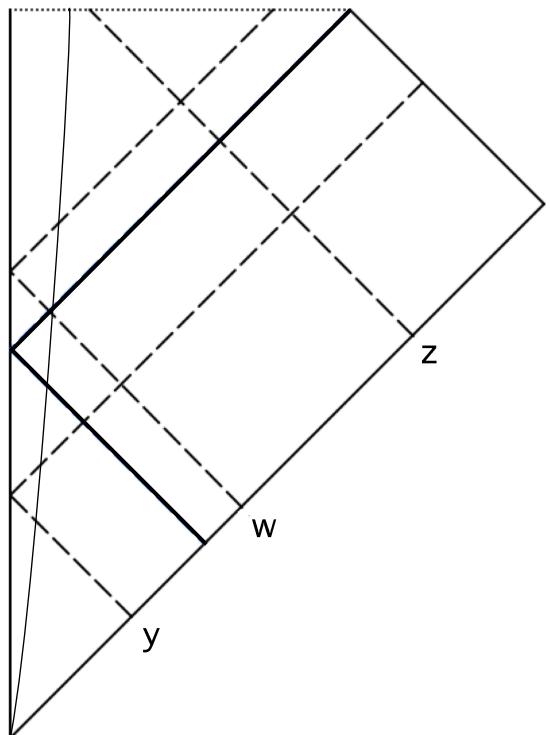


FIG. 2: A black hole spacetime which is the time reversal of figure 1. We impose the condition that future null infinity $\mathcal{J}_+^{\text{bh}}$ is in the vacuum state – there is no outgoing Hawking radiation. The dotted black line is the final singularity, and the thick solid line is the path of a null ray which coincides with the horizon at late times. The curved line indicates matter which collapses to form the hole. The dashed black lines refer to modes discussed in the text.

the condition of isolation further reduces the entropy substantially. Our analysis is mainly of theoretical, rather than practical astrophysical, interest.

In our analysis we only treat the case of a Schwarzschild black hole. Holes with angular momentum or electric charge have a more complex inner structure, including a Cauchy horizon. Interactions between outgoing and backscattered ingoing radiation near this Cauchy horizon lead to a curvature singularity known as mass inflation [7]. The resulting inner structure seems to involve quantum gravitational effects and is still not completely understood.

Following Hawking [8], we define an orthogonal set of modes for a scalar field on the black hole spacetime.

$$\phi = \int d\omega (f_\omega a_\omega + \bar{f}_\omega a_\omega^\dagger) , \quad (1)$$

where the $\{f_\omega\}$ are a complete, orthonormal family of complex solutions of the wave equation. The notation used here is identical to that in [8], except that (see below) we define the destruction operators of the w, y, z modes to be a^w, a^y, a^z rather than g, h, j as Hawking did. In all other respects, we adhere to his definitions, which

we now briefly review.

The modes are defined on an extended Schwarzschild geometry, obtained by analytic continuation, which includes both past and future horizons, \mathcal{H}^\pm , but does not describe the collapse which forms the black hole. Let u and v be the retarded and advanced time coordinates for the Schwarzschild metric. Let U (Kruskal coordinate) be the affine parameter on the past horizon \mathcal{H}^- :

$$u = -4M \ln(-U) , \quad (2)$$

where M is the mass of the black hole. The $\{f_\omega^{(1)}\}$ modes are solutions on the extended spacetime with zero Cauchy data on the past horizon and time dependence of the form $\exp(i\omega v)$ on $\mathcal{J}_+^{\text{bh}}$. The $\{f_\omega^{(2)}\}$ are solutions with zero Cauchy data on $\mathcal{J}_-^{\text{bh}}$ and dependence $\exp(i\omega U)$ on the past horizon. The analytic continuation of u yields two coordinates

$$u_\pm = -4M(\ln U \mp i\pi) \quad (U > 0) , \quad (3)$$

with $u_+ = u_-$ for $U < 0$. These are used to replace the $f^{(2)}$ modes by two orthogonal families of solutions $f^{(3)}$ and $f^{(4)}$. These have zero Cauchy data on $\mathcal{J}_-^{\text{bh}}$ and dependence on the past horizon of the form $\exp(i\omega u_+)$ and $\exp(i\omega u_-)$, respectively.

The physical interpretation of these modes, after continuation back to the collapse spacetime, is as follows. The $f^{(1)}$ modes enter the black hole after a horizon has formed. The $f^{(2)}$ modes (equivalently, the $f^{(3),(4)}$ modes) enter the black hole region before the $f^{(1)}$ modes, at earlier advanced time; in the time-reversed spacetime they would exit the white hole before it emerges from behind its anti-horizon (see figure 1). The quantum states associated with these modes are observable to a detector at $\mathcal{J}_+^{\text{bh}}$, or equivalently at $\mathcal{J}_+^{\text{wh}}$. We define the destruction operators for the f modes to be a^1, a^3, a^4 .

It is useful to define additional bases of modes (see figures): $\{w_\omega\}$, $\{y_\omega\}$ and $\{z_\omega\}$, which are linear combinations of the $\{f_\omega^{(i)}\}$, and are observable by a detector at $\mathcal{J}_+^{\text{bh}}$. The $\{w_\omega\}$ modes have zero Cauchy data on $\mathcal{J}_-^{\text{bh}}$ and on the past horizon for $U < 0$. For $U > 0$ on the past horizon their dependence is of the form $\exp(-i\omega u_+)$. The $\{y_\omega\}$ modes have zero Cauchy data on $\mathcal{J}_-^{\text{bh}}$ and on the past horizon for $U > 0$. For $U < 0$ on the past horizon their dependence is of the form $\exp(i\omega u_+)$. The $\{z_\omega\}$ modes are identical to the $f^{(1)}$ modes already defined. The destruction operators for these new modes are a^w, a^y, a^z .

The physical interpretation of these modes, after continuation back to the collapse spacetime, is as follows. The y modes enter the spatial region where the black hole will be formed (i.e., the precursor), but emerge before a horizon appears. The transmitted y modes (which are not reflected by the gravitational potential back into the hole) are observable at future null infinity of the black hole spacetime, $\mathcal{J}_+^{\text{bh}}$. The w modes propagate in from $\mathcal{J}_-^{\text{bh}}$, enter the black hole region of space (i.e., the precursor) before a horizon is formed, but are trapped and

encounter the future singularity. In the white hole spacetime (see figure 1), w modes emerge first from the anti-horizon, followed by the y modes, which appear after matter begins to explode from behind the anti-horizon.

The Hawking radiation modes p_ω , which are observable by a detector at $\mathcal{J}_+^{\text{bh}}$, are a complete set of orthonormal solutions which contain only positive frequencies at $\mathcal{J}_+^{\text{bh}}$ and are purely outgoing (zero Cauchy data on the horizon of the collapse spacetime). They can be written in terms of the y and z modes [8]:

$$p_\omega = t_\omega y_\omega + r_\omega z_\omega , \quad (4)$$

and the destruction operator for this mode is

$$a_\omega^p = \bar{t}_\omega a_\omega^y + \bar{r}_\omega a_\omega^z . \quad (5)$$

Equation (4) can be understood as follows from figure 2, depicting the black hole spacetime. The modes which reach future infinity are a superposition of transmitted y modes and reflected z modes, where t and r are the transmission and reflection amplitudes for waves incident on the black hole.

The condition that $\mathcal{J}_+^{\text{bh}}$ is in the vacuum state (no Hawking radiation; an *eternal* black hole) is

$$a_\omega^p |0_+^{\text{bh}}\rangle = 0 . \quad (6)$$

This condition is not typically imposed on the future state of the black hole spacetime. Instead, one usually requires that the precursor state (i.e., a collapsing star) is surrounded by vacuum, which is a condition on the past rather than on the future. However, the time reversal symmetry of quantum field theory and of general relativity imply that there must exist initial conditions that lead to the future condition (6). In the white hole case it is natural to impose the vacuum condition on the past, and we explore what its consequences are for the future of the hole.

A sufficient, but not necessary, condition for satisfying (6) is to require

$$a_\omega^y |0_+^{\text{bh}}\rangle = a_\omega^z |0_+^{\text{bh}}\rangle = 0 . \quad (7)$$

In his original discussion of the future vacuum on the black hole spacetime [8], Hawking imposes (7) as well as the additional condition

$$a_\omega^w |0_+^{\text{bh}}\rangle = 0 , \quad (8)$$

requiring that the future vacuum be empty of unnecessary w modes. In our discussion of white holes this condition need not apply since we do not wish to constrain the initial state of the white hole other than to require its isolation.

It is straightforward to calculate the particle number content, mode by mode, for the state defined above. We are specifically interested in the $f^{(i)}$ modes, which are detectable as particles incident on the black hole and its precursor by an observer at $\mathcal{J}_-^{\text{bh}}$ (past infinity of the

black hole spacetime). Equivalently, these modes are detectable as outgoing particles by an observer far outside the white hole. The condition $a_\omega^z |0_{+}^{\text{bh}}\rangle = 0$, imposed in (7), is identical to the condition $a^1 |0_{+}^{\text{bh}}\rangle = 0$, which implies that there are no $f^{(1)}$ or z modes emitted by the white hole (or absorbed by the eternal black hole). These modes are emitted by the white hole long before it explodes from behind its anti-horizon, or equivalently are absorbed by the black hole long after its horizon forms (see figures). The remaining $f^{(3,4)}$ modes are linear combinations of the w and y modes, which are emitted by the white hole just before and after it explodes. The y modes, in particular, appear to be emitted from the ejecta of the hole.

We obtain

$$\langle 0_{+}^{\text{bh}} | a_\omega^{3\dagger} a_\omega^3 + a_\omega^{4\dagger} a_\omega^4 | 0_{+}^{\text{bh}} \rangle = \frac{2x}{1-x} + \frac{1+x}{1-x} \langle 0_{+}^{\text{bh}} | a_\omega^{w\dagger} a_\omega^w | 0_{+}^{\text{bh}} \rangle, \quad (9)$$

where $x = \exp(-\beta\omega)$ and β the Hawking temperature of the black hole. In the simple case with $a_\omega^w |0_{+}^{\text{bh}}\rangle = 0$, the particle occupation numbers of each of the $f^{(3)}$ and $f^{(4)}$ modes are simply those of the blackbody distribution ($i = 3$ or 4):

$$\langle 0_{+}^{\text{bh}} | a_\omega^{i\dagger} a_\omega^i | 0_{+}^{\text{bh}} \rangle = \frac{x}{1-x} = \frac{1}{\exp(\beta\omega) - 1} \quad . \quad (10)$$

Physically, this means that one can construct an *eternal* (i.e., non-radiating) black hole in the minimal state satisfying (7) and (8) by exposing its precursor (and, briefly, the black hole itself) to a special quasi-thermal radiation state. It also implies that an isolated white hole in the state satisfying (8) will radiate quasi-thermally just before and after it explodes from behind its anti-horizon. Note that although the occupation numbers we have calculated are thermal, the state is actually a pure state if the initial white hole state was pure. Unlike in the case of Hawking radiation, we are not forced to trace over any causally disconnected region, and we do not necessarily obtain a mixed state description.

In our analysis so far we have treated the background spacetime as fixed and have neglected backreaction effects. In the original Hawking analysis, one first obtains the thermal spectrum of black hole radiation, and then invokes energy conservation and backreaction to argue that the hole steadily loses mass through radiation, eventually (perhaps) evaporating completely. Since the rate of energy loss is small it is assumed that the semiclassical analysis pertains until the final Planckian stage of evaporation. In our case we can make the same argument regarding the white hole: our calculations initially assume a fixed spacetime, but lead to thermal behavior of the hole just before and after it explodes. Conservation of energy implies that the white hole and the ejecta somehow compensate for the emitted radiation so that the total energy that reaches infinity is the initial ADM mass of the hole. How this happens is not entirely clear, although one can simply regard it as a constraint

on possible final states resulting from an isolated white hole. We note that the mode bases used in this analysis only depend on the asymptotic structure of the black or white hole spacetime. The details of how the black hole is formed, or how the white hole explodes, do not affect the results; indeed, the analysis can be formulated on the extended Schwarzschild spacetime which is the analytic continuation of the realistic geometry which contains a collapsing/exploding body.

A necessary and sufficient condition for isolation of the white hole (as opposed to (7), which was sufficient, but a special case) is

$$\bar{t}_\omega a_\omega^y |0_{+}^{\text{bh}}\rangle = -\bar{r}_\omega a_\omega^z |0_{+}^{\text{bh}}\rangle . \quad (11)$$

As mentioned previously, the condition (11) can be understood (see figure 2) as the requirement that reflected z modes interfere perfectly with transmitted y modes so that no Hawking radiation reaches future infinity of the black hole spacetime. This, more general, condition allows for Hawking-like radiation from the white hole in the form of z modes, which leave the white hole long before its explosion and reach future infinity $\mathcal{J}_{+}^{\text{wh}}$.

In the general case, we obtain the following expression for $f^{(3,4)}$ mode occupation numbers:

$$\langle 0_{+}^{\text{bh}} | \sum_{i=3,4} a_\omega^{i\dagger} a_\omega^i | 0_{+}^{\text{bh}} \rangle = \frac{1}{1-x} \left[2x + \langle 0_{+}^{\text{bh}} | (1+x) \times (a^{y\dagger} a^y + a^{w\dagger} a^w) - 2\sqrt{x}(a^w a^y + a^{w\dagger} a^{y\dagger}) | 0_{+}^{\text{bh}} \rangle \right] . \quad (12)$$

For $|0_{+}^{\text{bh}}\rangle$ which are particle number eigenstates of y and w , one can use the Cauchy-Schwarz inequality

$$|\langle a^w a^y \rangle|^2 \leq \langle a^{w\dagger} a^w a^{y\dagger} a^y \rangle ,$$

and identities $(1+x) \geq 2\sqrt{x}$ and $N_y + N_w \geq 2\sqrt{N_w N_y}$, to see that the expectation value of the mode number is, for every frequency, at least as large as in the simplest case where the conditions (7) and (8) are satisfied.

For a white hole to be indistinguishable from an ordinary black hole it must emit Hawking-like radiation from the beginning, with thermal occupation numbers for $\langle a^z a^z \rangle$. Condition (11) then requires non-zero occupation numbers for $\langle a^{y\dagger} a^y \rangle$, leading to more energy radiated in $f^{(3,4)}$ modes at late times. A significant amount of energy in this form must emerge *after* the white hole explodes, which limits how much can be radiated before it explodes. It is hard to see how an isolated white hole can behave so as to be indistinguishable from an ordinary black hole of equal mass. This only seems possible if we remove the condition of isolation, allowing the white hole to both emit and absorb energy, as would be the case for the thermal box considered originally by Hawking [5].

Conclusions

We summarize our main results below. These results have not, to our knowledge appeared previously in the literature.

1. Isolated white holes behave very differently from isolated black holes. This is due to the lack of time reversal symmetry in the surrounding environment: the statistical arrow of time implies that isolated black holes evaporate into their cold surroundings, whereas isolated white holes are, by definition, not bathed in incident radiation. Complete time reversal symmetry is only present in thermal equilibrium, the case originally analyzed by Hawking.

2. Isolated white holes with initial state given by the simple conditions (7) and (8) emit quasi-thermal radiation just before and after exploding from behind their anti-horizon. Modifying the initial state, while retaining the condition of isolation, likely implies even more radiation at late stages. There do not seem to be isolated white holes which are indistinguishable from isolated black holes of the same mass.

3. As a byproduct of our investigation, we note the existence of eternal – non-evaporating – black holes, formed from special quantum initial states. We do not know whether such holes are stable against perturbations. That is, if one prepares a black hole in this “eternal” state, but the hole subsequently interacts with a small probe (whose existence was not anticipated in the original preparation), does this cause only a small leakage of Hawking radiation, or does the hole revert to ordinary evaporation? Another interesting question is the relative entropy of eternal and ordinary black holes.

Acknowledgments — The author thanks Roberto Casadio, Ted Jacobson and David Reeb for comments or discussions. This work is supported by the Department of Energy under grant DE-FG02-96ER40969.

-
- [1] V. P. Frolov and I. D. Novikov, Black Hole Physics: Basic Concepts and New Developments, Kluwer Academic Publishers (1998).
- [2] D. M. Eardley and S. B. Giddings, Phys. Rev. D **66**, 044011 (2002) [[gr-qc/0201034](#)]; S. D. H. Hsu, Phys. Lett. B **555**, 92 (2003) [[hep-ph/0203154](#)].
- [3] To be more specific, take a spacelike slice (e.g., horizontal line) near the top of figure 2 (black hole Penrose diagram), but below the singularity. The time reversal of this state has the same values (and time derivatives) of metric and matter fields as a corresponding spacelike slice near the bottom of figure 1 (white hole diagram). If colliding particles can tunnel to the first (black hole) state, they can also tunnel to the corresponding white hole state (i.e., time reversal does not alter any quantum numbers). Note that because this is a quantum process, there is no singularity in the past of the white hole created. The spacetime with two colliding particles is simply patched onto the initial condition described by the spacelike slice from figure 1.
- [4] S. D. H. Hsu and D. Reeb, Phys. Lett. B **658**, 244 (2008) [[arXiv:0706.3239 \[hep-th\]](#)]; Class. Quant. Grav. **25**, 235007 (2008) [[arXiv:0803.4212 \[hep-th\]](#)]; Mod. Phys. Lett. A **24**, 1875 (2009) [[arXiv:0908.1265 \[gr-qc\]](#)].
- [5] S. W. Hawking, Phys. Rev. D **13**, 191 (1976).
- [6] P. Frampton, S. D. H. Hsu, D. Reeb and T. W. Kephart, Class. Quant. Grav. **26**, 145005 (2009) [[arXiv:0801.1847 \[hep-th\]](#)].
- [7] E. Poisson and W. Israel, Phys. Rev. Lett. **63**, 1663 (1989); Phys. Lett. B **233**, 74 (1989).
- [8] S. W. Hawking, Phys. Rev. D **14**, 2460 (1976).

FÜRUE WHI

RELATIVITY

THE SPECIAL AND GENERAL THEORY

ALBERT EINSTEIN

DIGITAL REPRINT

Elegant Ebooks

COPYRIGHT INFORMATION

Book: *Relativity: The Special and General Theory*

Author: Albert Einstein, 1879–1955

First published: 1920

The original book is in the public domain in the United States. However, since Einstein died in 1955, it may still be under copyright in many other countries, for example, those that use the life of the author + 60 years or life + 70 years for the duration of copyright. Readers outside the United States should check their own countries' copyright laws to be certain they can legally download this ebook. The [Online Books Page](#) has an [FAQ](#) which gives a summary of copyright durations for many other countries, as well as links to more official sources.

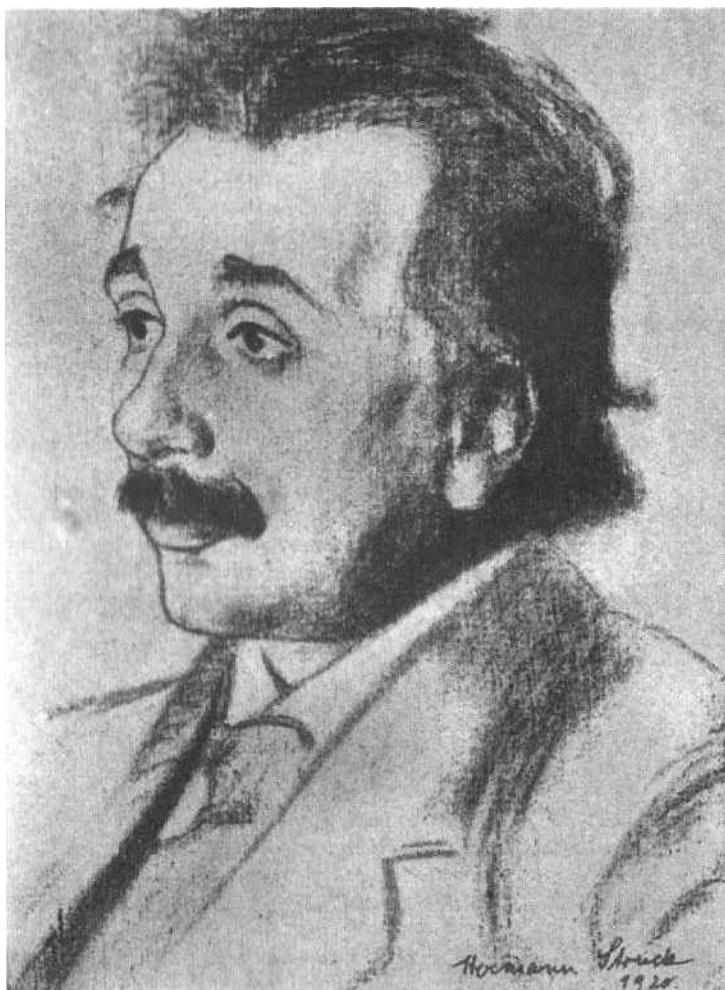
This PDF ebook was
created by José Menéndez.

NOTE ON THE TEXT

The text used in this ebook is from the first English translation, published in 1920, of *Über die spezielle und die allgemeine Relativitätstheorie*. A few misprints in the original text have been corrected, but to preserve all of the book, the original misprints are included in footnotes enclosed in square brackets and signed “J.M.”

In addition, the line breaks and pagination of the original book have been reproduced, and the PDF’s page numbers are linked to [page images](#) of the book from the [Google Books Library Project](#) for easy comparison. Please note: Google’s images may not be available to people outside the U.S. and may be unavailable to U.S. residents at times.

Retaining the original pagination made it possible to include a fully functional (clickable) copy of the original [index](#).



A. Einstein

FÉUE WHI

RELATIVITY

THE SPECIAL AND GENERAL THEORY

BY

ALBERT EINSTEIN, Ph.D.

PROFESSOR OF PHYSICS IN THE UNIVERSITY OF BERLIN

TRANSLATED BY

ROBERT W. LAWSON, M.Sc.

UNIVERSITY OF SHEFFIELD



NEW YORK
HENRY HOLT AND COMPANY
1920

601

FÉUE WHI

**COPYRIGHT, 1920
BY
HENRY HOLT AND COMPANY**

PREFACE

THE present book is intended, as far as possible, to give an exact insight into the theory of Relativity to those readers who, from a general scientific and philosophical point of view, are interested in the theory, but who are not conversant with the mathematical apparatus¹ of theoretical physics. The work presumes a standard of education corresponding to that of a university matriculation examination, and, despite the shortness of the book, a fair amount of patience and force of will on the part of the reader. The author has spared himself no pains in his endeavour to present the main ideas in the simplest and most intelligible form, and on the

¹ The mathematical fundaments of the special theory of relativity are to be found in the original papers of H. A. Lorentz, A. Einstein, H. Minkowski,* published under the title *Das Relativitätsprinzip* (The Principle of Relativity) in B. G. Teubner's collection of monographs *Fortschritte der mathematischen Wissenschaften* (Advances in the Mathematical Sciences), also in M. Laue's exhaustive book *Das Relativitätsprinzip* — published by Friedr. Vieweg & Son, Braunschweig. The general theory of relativity, together with the necessary parts of the theory of invariants, is dealt with in the author's book *Die Grundlagen der allgemeinen Relativitätstheorie* (The Foundations of the General Theory of Relativity) — Joh. Ambr. Barth, 1916; this book assumes some familiarity with the special theory of relativity.

RELATIVITY

whole, in the sequence and connection in which they actually originated. In the interest of clearness, it appeared to me inevitable that I should repeat myself frequently, without paying the slightest attention to the elegance of the presentation. I adhered scrupulously to the precept of that brilliant theoretical physicist, L. Boltzmann, according to whom matters of elegance ought to be left to the tailor and to the cobbler. I make no pretence of having withheld from the reader difficulties which are inherent to the subject. On the other hand, I have purposely treated the empirical physical foundations of the theory in a “step-motherly” fashion, so that readers unfamiliar with physics may not feel like the wanderer who was unable to see the forest for trees. May the book bring some one a few happy hours of suggestive thought!

A. EINSTEIN

December, 1916

NOTE TO THE THIRD EDITION

IN the present year (1918) an excellent and detailed manual on the general theory of relativity, written by H. Weyl, was published by the firm Julius Springer (Berlin). This book, entitled *Raum — Zeit — Materie* (Space — Time — Matter), may be warmly recommended to mathematicians and physicists.

BIOGRAPHICAL NOTE

ALBERT EINSTEIN is the son of German-Jewish parents. He was born in 1879 in the town of Ulm, Würtemberg, Germany. His schooldays were spent in Munich, where he attended the *Gymnasium* until his sixteenth year. After leaving school at Munich, he accompanied his parents to Milan, whence he proceeded to Switzerland six months later to continue his studies.

From 1896 to 1900 Albert Einstein studied mathematics and physics at the Technical High School in Zurich, as he intended becoming a secondary school (*Gymnasium*) teacher. For some time afterwards he was a private tutor, and having meanwhile become naturalised, he obtained a post as engineer in the Swiss Patent Office in 1902, which position he occupied till 1909. The main ideas involved in the most important of Einstein's theories date back to this period. Amongst these may be mentioned: *The Special Theory of Relativity*, *Inertia of Energy*, *Theory of the Brownian Movement*, and the *Quantum-Law of the Emission and Absorption of Light* (1905). These were followed some years later by the

RELATIVITY

Theory of the Specific Heat of Solid Bodies, and the fundamental idea of the *General Theory of Relativity*.

During the interval 1909 to 1911 he occupied the post of Professor *Extraordinarius* at the University of Zurich, afterwards being appointed to the University of Prague, Bohemia, where he remained as Professor *Ordinarius* until 1912. In the latter year Professor Einstein accepted a similar chair at the *Polytechnikum*, Zurich, and continued his activities there until 1914, when he received a call to the Prussian Academy of Science, Berlin, as successor to Van't Hoff. Professor Einstein is able to devote himself freely to his studies at the Berlin Academy, and it was here that he succeeded in completing his work on the *General Theory of Relativity* (1915–17). Professor Einstein also lectures on various special branches of physics at the University of Berlin, and, in addition, he is Director of the Institute* for Physical Research of the *Kaiser Wilhelm Gesellschaft*.

Professor Einstein has been twice married. His first wife, whom he married at Berne in 1903, was a fellow-student from Serbia. There were two sons of this marriage, both of whom are living in Zurich, the elder being sixteen years of age. Recently Professor Einstein married a widowed cousin, with whom he is now living in Berlin.

R. W. L.

[* Institutte — J.M.]

TRANSLATOR'S NOTE

IN presenting this translation to the English-reading public, it is hardly necessary for me to enlarge on the Author's prefatory remarks, except to draw attention to those additions to the book which do not appear in the original.

At my request, Professor Einstein kindly supplied me with a portrait of himself, by one of Germany's most celebrated artists. Appendix III, on "The Experimental Confirmation of the General Theory of Relativity," has been written specially for this translation. Apart from these valuable additions to the book, I have included a biographical note on the Author, and, at the end of the book, an **Index** and a **list** of English references to the subject. This list, which is more suggestive than exhaustive, is intended as a guide to those readers who wish to pursue the subject farther.

I desire to tender my best thanks to my colleagues Professor S. R. Milner, D.Sc., and Mr. W. E. Curtis, A.R.C.Sc., F.R.A.S., also to my friend Dr. Arthur Holmes, A.R.C.Sc., F.G.S.,

RELATIVITY

of the Imperial College, for their kindness in reading through the manuscript, for helpful criticism, and for numerous suggestions. I owe an expression of thanks also to Messrs. Methuen for their ready counsel and advice, and for the care they have bestowed on the work during the course of its publication.

ROBERT W. LAWSON

THE PHYSICS LABORATORY
THE UNIVERSITY OF SHEFFIELD
June 12, 1920

CONTENTS

PART I

THE SPECIAL THEORY OF RELATIVITY

	PAGE
I. Physical Meaning of Geometrical Propositions	1
II. The System of Co-ordinates	5
III. Space and Time in Classical Mechanics	9
IV. The Galileian System of Co-ordinates	12
V. The Principle of Relativity (in the Restricted Sense)	14
VI. The Theorem of the Addition of Velocities employed in Classical Mechanics	19
VII. The Apparent Incompatibility of the Law of Propagation of Light with the Principle of Relativity	21
VIII. On the Idea of Time in Physics	25
IX. The Relativity of Simultaneity	30
X. On the Relativity of the Conception of Distance	34
XI. The Lorentz Transformation	36
XII. The Behaviour of Measuring-Rods and Clocks in Motion	42

RELATIVITY

	PAGE
XIII. Theorem of the Addition of Velocities. The Experiment of Fizeau	45
XIV. The Heuristic Value of the Theory of Relativity	50
XV. General Results of the Theory	52
XVI. Experience and the Special Theory of Relativity*	58
XVII. Minkowski's Four-dimensional Space . .	65

PART II

THE GENERAL THEORY OF RELATIVITY

XVIII. Special and General Principle of Relativity	69
XIX. The Gravitational Field	74
XX. The Equality of Inertial and Gravitational Mass as an Argument for the General Postulate of Relativity	78
XXI. In what Respects are the Foundations of Classical Mechanics and of the Special Theory of Relativity unsatisfactory? .	84
XXII. A Few Inferences from the General Theory [†] of Relativity	87
XXIII. Behaviour of Clocks and Measuring-Rods on a Rotating Body of Reference . . .	93
XXIV. Euclidean and Non-Euclidean Continuum	98
XXV. Gaussian Co-ordinates	103
XXVI. The Space-time Continuum of the Special Theory of Relativity considered as a Euclidean Continuum	108

[* Relativity — J.M.]

[[†] “Theory” was changed to “Principle” in later editions. — J.M.]

CONTENTS

xiii

PAGE

XXVII. The Space-time Continuum of the General Theory of Relativity is not a Euclidean Continuum	111
XXVIII. Exact Formulation of the General Principle of Relativity	115
XXIX. The Solution of the Problem of Gravitation on the Basis of the General Principle of Relativity	119

PART III

CONSIDERATIONS ON THE UNIVERSE AS A WHOLE

XXX. Cosmological Difficulties of Newton's Theory	125
XXXI. The Possibility of a "Finite" and yet "Unbounded" Universe	128
XXXII. The Structure of Space according to the General Theory of Relativity	135

APPENDICES

I. Simple Derivation of the Lorentz Transformation	139
II. Minkowski's Four-dimensional Space ("World") [Supplementary to Section XVII.]	146
III. The Experimental Confirmation of the General Theory of Relativity	148
(a) Motion of the Perihelion of Mercury	150
(b) Deflection of Light by a Gravitational Field	152
(c) Displacement of Spectral Lines towards the Red	155
BIBLIOGRAPHY	161
INDEX	165

611

FÉUE WHI

FILEUE WHI

RELATIVITY

PART I

THE SPECIAL THEORY OF RELATIVITY

I

PHYSICAL MEANING OF GEOMETRICAL PROPOSITIONS

IN your schooldays most of you who read this book made acquaintance with the noble building of Euclid's geometry, and you remember — perhaps with more respect than love — the magnificent structure, on the lofty staircase of which you were chased about for uncounted hours by conscientious teachers. By reason of your past experience, you would certainly regard every one with disdain who should pronounce even the most out-of-the-way proposition of this science to be untrue. But perhaps this feeling of proud certainty would leave you immediately if some one were to ask you: "What, then, do you mean by the assertion that these propositions are true?" Let us proceed to give this question a little consideration.

Geometry sets out from certain conceptions such as "plane," "point," and "straight line," with

2 SPECIAL THEORY OF RELATIVITY

which we are able to associate more or less definite ideas, and from certain simple propositions (axioms) which, in virtue of these ideas, we are inclined to accept as "true." Then, on the basis of a logical process, the justification of which we feel ourselves compelled to admit, all remaining propositions are shown to follow from those axioms, *i.e.* they are proven. A proposition is then correct ("true") when it has been derived in the recognised manner from the axioms. The question of the "truth" of the individual geometrical propositions is thus reduced to one of the "truth" of the axioms. Now it has long been known that the last question is not only unanswerable by the methods of geometry, but that it is in itself entirely without meaning. We cannot ask whether it is true that only one straight line goes through two points. We can only say that Euclidean geometry deals with things called "straight lines," to each of which is ascribed the property of being uniquely determined by two points situated on it. The concept "true" does not tally with the assertions of pure geometry, because by the word "true" we are eventually in the habit of designating always the correspondence with a "real" object; geometry, however, is not concerned with the relation of the ideas involved in it to objects of experience, but only with the logical connection of these ideas among themselves.

It is not difficult to understand why, in spite of this, we feel constrained to call the propositions of geometry "true." Geometrical ideas correspond to more or less exact objects in nature, and these last are undoubtedly the exclusive cause of the genesis of those ideas. Geometry ought to refrain from such a course, in order to give to its structure the largest possible logical unity. The practice, for example, of seeing in a "distance" two marked positions on a practically rigid body is something which is lodged deeply in our habit of thought. We are accustomed further to regard three points as being situated on a straight line, if their apparent positions can be made to coincide for observation with one eye, under suitable choice of our place of observation.

If, in pursuance of our habit of thought, we now supplement the propositions of Euclidean geometry by the single proposition that two points on a practically rigid body always correspond to the same distance (line-interval), independently of any changes in position to which we may subject the body, the propositions of Euclidean geometry then resolve themselves into propositions on the possible relative position of practically rigid bodies.¹

¹ It follows that a natural object is associated also with a straight line. Three points A , B and C on a rigid body thus lie in a straight line when, the points A and C being given, B is chosen such that the sum of the distances AB and BC is as short as possible. This incomplete suggestion will suffice for our present purpose.

4 SPECIAL THEORY OF RELATIVITY

Geometry which has been supplemented in this way is then to be treated as a branch of physics. We can now legitimately ask as to the "truth" of geometrical propositions interpreted in this way, since we are justified in asking whether these propositions are satisfied for those real things we have associated with the geometrical ideas. In less exact terms we can express this by saying that by the "truth" of a geometrical proposition in this sense we understand its validity for a construction with ruler and compasses.

Of course the conviction of the "truth" of geometrical propositions in this sense is founded exclusively on rather incomplete experience. For the present we shall assume the "truth" of the geometrical propositions, then at a later stage (in the general theory of relativity) we shall see that this "truth" is limited, and we shall consider the extent of its limitation.

II

THE SYSTEM OF CO-ORDINATES

ON the basis of the physical interpretation of distance which has been indicated, we are also in a position to establish the distance between two points on a rigid body by means of measurements. For this purpose we require a “distance” (rod S) which is to be used once and for all, and which we employ as a standard measure. If, now, A and B are two points on a rigid body, we can construct the line joining them according to the rules of geometry; then, starting from A , we can mark off the distance S time after time until we reach B . The number of these operations required is the numerical measure of the distance AB . This is the basis of all measurement of length.¹

Every description of the scene of an event or of the position of an object in space is based on the specification of the point on a rigid body (body of reference) with which that event or object coin-

¹ Here we have assumed that there is nothing left over, *i.e.* that the measurement gives a whole number. This difficulty is got over by the use of divided measuring-rods, the introduction of which does not demand any fundamentally new method.

6 SPECIAL THEORY OF RELATIVITY

cides. This applies not only to scientific description, but also to everyday life. If I analyse the place specification "Trafalgar Square, London,"¹ I arrive at the following result. The earth is the rigid body to which the specification of place refers; "Trafalgar Square, London" is a well-defined point, to which a name has been assigned, and with which the event coincides in space.²

This primitive method of place specification deals only with places on the surface of rigid bodies, and is dependent on the existence of points on this surface which are distinguishable from each other. But we can free ourselves from both of these limitations without altering the nature of our specification of position. If, for instance, a cloud is hovering over Trafalgar Square, then we can determine its position relative to the surface of the earth by erecting a pole perpendicularly on the Square, so that it reaches the cloud. The length of the pole measured with the standard measuring-rod, combined with the specification of the position of the foot of the pole, supplies us with a complete place specification. On the basis

¹ I have chosen this as being more familiar to the English reader than the "Potsdamer Platz, Berlin," which is referred to in the original. (R. W. L.)

² It is not necessary here to investigate further the significance of the expression "coincidence in space." This conception is sufficiently obvious to ensure that differences of opinion are scarcely likely to arise as to its applicability in practice.

THE SYSTEM OF CO-ORDINATES 7

of this illustration, we are able to see the manner in which a refinement of the conception of position has been developed.

(a) We imagine the rigid body, to which the place specification is referred, supplemented in such a manner that the object whose position we require is reached by the completed rigid body.

(b) In locating the position of the object, we make use of a number (here the length of the pole measured with the measuring-rod) instead of designated points of reference.

(c) We speak of the height of the cloud even when the pole which reaches the cloud has not been erected. By means of optical observations of the cloud from different positions on the ground, and taking into account the properties of the propagation of light, we determine the length of the pole we should have required in order to reach the cloud.

From this consideration we see that it will be advantageous if, in the description of position, it should be possible by means of numerical measures to make ourselves independent of the existence of marked positions (possessing names) on the rigid body of reference. In the physics of measurement this is attained by the application of the Cartesian system of co-ordinates.

This consists of three plane surfaces perpendicular to each other and rigidly attached to a rigid

8 SPECIAL THEORY OF RELATIVITY

body. Referred to a system of co-ordinates, the scene of any event will be determined (for the main part) by the specification of the lengths of the three perpendiculars or co-ordinates (x , y , z) which can be dropped from the scene of the event to those three plane surfaces. The lengths of these three perpendiculars can be determined by a series of manipulations with rigid measuring-rods performed according to the rules and methods laid down by Euclidean geometry.

In practice, the rigid surfaces which constitute the system of co-ordinates are generally not available; furthermore, the magnitudes of the co-ordinates are not actually determined by constructions with rigid rods, but by indirect means. If the results of physics and astronomy are to maintain their clearness, the physical meaning of specifications of position must always be sought in accordance with the above considerations.¹

We thus obtain the following result: Every description of events in space involves the use of a rigid body to which such events have to be referred. The resulting relationship takes for granted that the laws of Euclidean geometry hold for "distances," the "distance" being represented physically by means of the convention of two marks on a rigid body.

¹ A refinement and modification of these views does not become necessary until we come to deal with the general theory of relativity, treated in the second part of this book.

III

SPACE AND TIME IN CLASSICAL MECHANICS

“**T**HE purpose of mechanics is to describe how bodies change their position in space with time.” I should load my conscience with grave sins against the sacred spirit of lucidity were I to formulate the aims of mechanics in this way, without serious reflection and detailed explanations. Let us proceed to disclose these sins.

It is not clear what is to be understood here by “position” and “space.” I stand at the window of a railway carriage which is travelling uniformly, and drop a stone on the embankment, without throwing it. Then, disregarding the influence of the air resistance, I see the stone descend in a straight line. A pedestrian who observes the misdeed from the footpath notices that the stone falls to earth in a parabolic curve. I now ask: Do the “positions” traversed by the stone lie “in reality” on a straight line or on a parabola? Moreover, what is meant here by motion “in space”? From the considerations of the previous section the answer is self-evident. In the first place, we entirely shun the vague word “space,”

10 SPECIAL THEORY OF RELATIVITY

of which, we must honestly acknowledge, we cannot form the slightest conception, and we replace it by "motion relative to a practically rigid body of reference." The positions relative to the body of reference (railway carriage or embankment) have already been defined in detail in the preceding section. If instead of "body of reference" we insert "system of co-ordinates," which is a useful idea for mathematical description, we are in a position to say: The stone traverses a straight line relative to a system of co-ordinates rigidly attached to the carriage, but relative to a system of co-ordinates rigidly attached to the ground (embankment) it describes a parabola. With the aid of this example it is clearly seen that there is no such thing as an independently existing trajectory (lit. "path-curve"¹), but only a trajectory relative to a particular body of reference.

In order to have a *complete* description of the motion, we must specify how the body alters its position *with time*; *i.e.* for every point on the trajectory it must be stated at what time the body is situated there. These data must be supplemented by such a definition of time that, in virtue of this definition, these time-values can be regarded essentially as magnitudes (results of measurements) capable of observation. If we take our stand on the ground of classical me-

¹ That is, a curve along which the body moves.

chanics, we can satisfy this requirement for our illustration in the following manner. We imagine two clocks of identical construction; the man at the railway-carriage window is holding one of them, and the man on the footpath the other. Each of the observers determines the position on his own reference-body occupied by the stone at each tick of the clock he is holding in his hand. In this connection we have not taken account of the inaccuracy involved by the finiteness of the velocity of propagation of light. With this and with a second difficulty prevailing here we shall have to deal in detail later.

IV

**THE GALILEIAN SYSTEM OF
CO-ORDINATES**

AS is well known, the fundamental law of the mechanics of Galilei-Newton, which is known as the *law of inertia*, can be stated thus: A body removed sufficiently far from other bodies continues in a state of rest or of uniform motion in a straight line. This law not only says something about the motion of the bodies, but it also indicates the reference-bodies or systems of co-ordinates, permissible in mechanics, which can be used in mechanical description. The visible fixed stars are bodies for which the law of inertia certainly holds to a high degree of approximation. Now if we use a system of co-ordinates which is rigidly attached to the earth, then, relative to this system, every fixed star describes a circle of immense radius in the course of an astronomical day, a result which is opposed to the statement of the law of inertia. So that if we adhere to this law we must refer these motions only to systems of co-ordinates relative to which the fixed stars do not move in a circle. A system of co-ordinates of

which the state of motion is such that the law of inertia holds relative to it is called a “Galileian system of co-ordinates.” The laws of the mechanics of Galilei-Newton can be regarded as valid only for a Galileian system of co-ordinates.

V

THE PRINCIPLE OF RELATIVITY (IN THE RESTRICTED SENSE)

IN order to attain the greatest possible clearness, let us return to our example of the railway carriage supposed to be travelling uniformly. We call its motion a uniform translation ("uniform" because it is of constant velocity and direction, "translation" because although the carriage changes its position relative to the embankment yet it does not rotate in so doing). Let us imagine a raven flying through the air in such a manner that its motion, as observed from the embankment, is uniform and in a straight line. If we were to observe the flying raven from the moving railway carriage, we should find that the motion of the raven would be one of different velocity and direction, but that it would still be uniform and in a straight line. Expressed in an abstract manner we may say: If a mass m is moving uniformly in a straight line with respect to a co-ordinate system K , then it will also be moving uniformly and in a straight line relative to a second co-ordinate system K' , provided that

the latter is executing a uniform translatory motion with respect to K . In accordance with the discussion contained in the preceding section, it follows that:

If K is a Galileian co-ordinate system, then every other co-ordinate system K' is a Galileian one, when, in relation to K , it is in a condition of uniform motion of translation. Relative to K' the mechanical laws of Galilei-Newton hold good exactly as they do with respect to K .

We advance a step farther in our generalisation when we express the tenet thus: If, relative to K , K' is a uniformly moving co-ordinate system devoid of rotation, then natural phenomena run their course with respect to K' according to exactly the same general laws as with respect to K . This statement is called the *principle of relativity* (in the restricted sense).

As long as one was convinced that all natural phenomena were capable of representation with the help of classical mechanics, there was no need to doubt the validity of this principle of relativity. But in view of the more recent development of electrodynamics and optics it became more and more evident that classical mechanics affords an insufficient foundation for the physical description of all natural phenomena. At this juncture the question of the validity of the principle of relativity became ripe for discussion, and it did not appear

16 SPECIAL THEORY OF RELATIVITY

impossible that the answer to this question might be in the negative.

Nevertheless, there are two general facts which at the outset speak very much in favour of the validity of the principle of relativity. Even though classical mechanics does not supply us with a sufficiently broad basis for the theoretical presentation of all physical phenomena, still we must grant it a considerable measure of "truth," since it supplies us with the actual motions of the heavenly bodies with a delicacy of detail little short of wonderful. The principle of relativity must therefore apply with great accuracy in the domain of *mechanics*. But that a principle of such broad generality should hold with such exactness in one domain of phenomena, and yet should be invalid for another, is *a priori* not very probable.

We now proceed to the second argument, to which, moreover, we shall return later. If the principle of relativity (in the restricted sense) does not hold, then the Galileian co-ordinate systems K , K' , K'' , etc., which are moving uniformly relative to each other, will not be *equivalent* for the description of natural phenomena. In this case we should be constrained to believe that natural laws are capable of being formulated in a particularly simple manner, and of course only on condition that, from amongst all possible Galileian

co-ordinate systems, we should have chosen *one* (K_0) of a particular state of motion as our body of reference. We should then be justified (because of its merits for the description of natural phenomena) in calling this system "absolutely at rest," and all other Galileian systems K "in motion." If, for instance, our embankment were the system K_0 , then our railway carriage would be a system K , relative to which less simple laws would hold than with respect to K_0 . This diminished simplicity would be due to the fact that the carriage K would be in motion (*i.e.* "really") with respect to K_0 . In the general laws of nature which have been formulated with reference to K , the magnitude and direction of the velocity of the carriage would necessarily play a part. We should expect, for instance, that the note emitted by an organ-pipe placed with its axis parallel to the direction of travel would be different from that emitted if the axis of the pipe were placed perpendicular to this direction. Now in virtue of its motion in an orbit round the sun, our earth is comparable with a railway carriage travelling with a velocity of about 30 kilometres per second. If the principle of relativity were not valid we should therefore expect that the direction of motion of the earth at any moment would enter into the laws of nature, and also that physical systems in their behaviour would be dependent on the orientation in space

18 SPECIAL THEORY OF RELATIVITY

with respect to the earth. For owing to the alteration in direction of the velocity of rotation* of the earth in the course of a year, the earth cannot be at rest relative to the hypothetical system K_0 throughout the whole year. However, the most careful observations have never revealed such anisotropic properties in terrestrial physical space, *i.e.* a physical non-equivalence of different directions. This is a very powerful argument in favour of the principle of relativity.

[* The word “rotation” was correctly changed to “revolution” in later editions. — J.M.]

VI

**THE THEOREM OF THE ADDITION OF
VELOCITIES EMPLOYED IN CLASSI-
CAL MECHANICS**

LET us suppose our old friend the railway carriage to be travelling along the rails with a constant velocity v , and that a man traverses the length of the carriage in the direction of travel with a velocity w . How quickly, or, in other words, with what velocity W does the man advance relative to the embankment during the process? The only possible answer seems to result from the following consideration: If the man were to stand still for a second, he would advance relative to the embankment through a distance v equal numerically to the velocity of the carriage. As a consequence of his walking, however, he traverses an additional distance w relative to the carriage, and hence also relative to the embankment, in this second, the distance w being numerically equal to the velocity with which he is walking. Thus in total he covers the distance $W = v + w$ relative to the embankment in the second considered. We shall see later that this result, which expresses the theorem of the addi-

20 SPECIAL THEORY OF RELATIVITY

tion of velocities employed in classical mechanics, cannot be maintained; in other words, the law that we have just written down does not hold in reality. For the time being, however, we shall assume its correctness.

VII

**THE APPARENT INCOMPATIBILITY OF THE
LAW OF PROPAGATION OF LIGHT WITH
THE PRINCIPLE OF RELATIVITY**

THREE is hardly a simpler law in physics than that according to which light is propagated in empty space. Every child at school knows, or believes he knows, that this propagation takes place in straight lines with a velocity $c = 300,000$ km./sec. At all events we know with great exactness that this velocity is the same for all colours, because if this were not the case, the minimum of emission would not be observed simultaneously for different colours during the eclipse of a fixed star by its dark neighbour. By means of similar considerations based on observations of double stars, the Dutch astronomer De Sitter was also able to show that the velocity of propagation of light cannot depend on the velocity of motion of the body emitting the light. The assumption that this velocity of propagation is dependent on the direction "in space" is in itself improbable.

In short, let us assume that the simple law of the constancy of the velocity of light c (in vacuum)

22 SPECIAL THEORY OF RELATIVITY

is justifiably believed by the child at school. Who would imagine that this simple law has plunged the conscientiously thoughtful physicist into the greatest intellectual difficulties? Let us consider how these difficulties arise.

Of course we must refer the process of the propagation of light (and indeed every other process) to a rigid reference-body (co-ordinate system). As such a system let us again choose our embankment. We shall imagine the air above it to have been removed. If a ray of light be sent along the embankment, we see from the above that the tip of the ray will be transmitted with the velocity c relative to the embankment. Now let us suppose that our railway carriage is again travelling along the railway lines with the velocity v , and that its direction is the same as that of the ray of light, but its velocity of course much less. Let us inquire about the velocity of propagation of the ray of light relative to the carriage. It is obvious that we can here apply the consideration of the previous section, since the ray of light plays the part of the man walking along relatively to the carriage. The velocity W of the man relative to the embankment is here replaced by the velocity of light relative to the embankment. w is the required velocity of light with respect to the carriage, and we have

$$w = c - v.$$

The velocity of propagation of a ray of light relative to the carriage thus comes out smaller than c .

But this result comes into conflict with the principle of relativity set forth in Section V. For, like every other general law of nature, the law of the transmission of light *in vacuo* must, according to the principle of relativity, be the same for the railway carriage as reference-body as when the rails are the body of reference. But, from our above consideration, this would appear to be impossible. If every ray of light is propagated relative to the embankment with the velocity c , then for this reason it would appear that another law of propagation of light must necessarily hold with respect to the carriage — a result contradictory to the principle of relativity.

In view of this dilemma there appears to be nothing else for it than to abandon either the principle of relativity or the simple law of the propagation of light *in vacuo*. Those of you who have carefully followed the preceding discussion are almost sure to expect that we should retain the principle of relativity, which appeals so convincingly to the intellect because it is so natural and simple. The law of the propagation of light *in vacuo* would then have to be replaced by a more complicated law conformable to the principle of relativity. The development of theoretical

24 SPECIAL THEORY OF RELATIVITY

physics shows, however, that we cannot pursue this course. The epoch-making theoretical investigations of H. A. Lorentz on the electrodynamical and optical phenomena connected with moving bodies show that experience in this domain leads conclusively to a theory of electromagnetic phenomena, of which the law of the constancy of the velocity of light *in vacuo* is a necessary consequence. Prominent theoretical physicists were therefore more inclined to reject the principle of relativity, in spite of the fact that no empirical data had been found which were contradictory to this principle.

At this juncture the theory of relativity entered the arena. As a result of an analysis of the physical conceptions of time and space, it became evident that *in reality there is not the least incompatibility between the principle of relativity and the law of propagation of light*, and that by systematically holding fast to both these laws a logically rigid theory could be arrived at. This theory has been called the *special theory of relativity* to distinguish it from the extended theory, with which we shall deal later. In the following pages we shall present the fundamental ideas of the special theory of relativity.

VIII

ON THE IDEA OF TIME IN PHYSICS

LIGHTNING has struck the rails on our railway embankment at two places *A* and *B* far distant from each other. I make the additional assertion that these two lightning flashes occurred simultaneously. If now I ask you whether there is sense in this statement, you will answer my question with a decided "Yes." But if I now approach you with the request to explain to me the sense of the statement more precisely, you find after some consideration that the answer to this question is not so easy as it appears at first sight.

After some time perhaps the following answer would occur to you: "The significance of the statement is clear in itself and needs no further explanation; of course it would require some consideration if I were to be commissioned to determine by observations whether in the actual case the two events took place simultaneously or not." I cannot be satisfied with this answer for the following reason. Supposing that as a result of ingenious considerations an able meteorologist were to dis-

26 SPECIAL THEORY OF RELATIVITY

cover that the lightning must always strike the places A and B simultaneously, then we should be faced with the task of testing whether or not this theoretical result is in accordance with the reality. We encounter the same difficulty with all physical statements in which the conception "simultaneous" plays a part. The concept does not exist for the physicist until he has the possibility of discovering whether or not it is fulfilled in an actual case. We thus require a definition of simultaneity such that this definition supplies us with the method by means of which, in the present case, he can decide by experiment whether or not both the lightning strokes occurred simultaneously. As long as this requirement is not satisfied, I allow myself to be deceived as a physicist (and of course the same applies if I am not a physicist), when I imagine that I am able to attach a meaning to the statement of simultaneity. (I would ask the reader not to proceed farther until he is fully convinced on this point.)

After thinking the matter over for some time you then offer the following suggestion with which to test simultaneity. By measuring along the rails, the connecting line AB should be measured up and an observer placed at the mid-point M of the distance AB . This observer should be supplied with an arrangement (e.g. two mirrors inclined at 90°) which allows him visually to ob-

serve both places A and B at the same time. If the observer perceives the two flashes of lightning at the same time, then they are simultaneous.

I am very pleased with this suggestion, but for all that I cannot regard the matter as quite settled, because I feel constrained to raise the following objection: "Your definition would certainly be right, if I only knew that the light by means of which the observer at M perceives the lightning flashes travels along the length $A \rightarrow M$ with the same velocity as along the length $B \rightarrow M$. But an examination of this supposition would only be possible if we already had at our disposal the means of measuring time. It would thus appear as though we were moving here in a logical circle."

After further consideration you cast a somewhat disdainful glance at me — and rightly so — and you declare: "I maintain my previous definition nevertheless, because in reality it assumes absolutely nothing about light. There is only *one* demand to be made of the definition of simultaneity, namely, that in every real case it must supply us with an empirical decision as to whether or not the conception that has to be defined is fulfilled. That my definition satisfies this demand is indisputable. That light requires the same time to traverse the path $A \rightarrow M$ as for the path $B \rightarrow M$ is in reality neither a *supposition* nor a *hypothesis* about the physical nature of light,

28 SPECIAL THEORY OF RELATIVITY

but a *stipulation* which I can make of my own freewill in order to arrive at a definition of simultaneity."

It is clear that this definition can be used to give an exact meaning not only to *two* events, but to as many events as we care to choose, and independently of the positions of the scenes of the events with respect to the body of reference¹ (here the railway embankment). We are thus led also to a definition of "time" in physics. For this purpose we suppose that clocks of identical construction are placed at the points *A*, *B* and *C* of the railway line (co-ordinate system), and that they are set in such a manner that the positions of their pointers are simultaneously (in the above sense) the same. Under these conditions we understand by the "time" of an event the reading (position of the hands) of that one of these clocks which is in the immediate vicinity (in space) of the event. In this manner a time-value is associated with every event which is essentially capable of observation.

This stipulation contains a further physical

¹ We suppose further that, when three events *A*, *B* and *C* take place in different places in such a manner that, if *A* is simultaneous with *B*, and *B* is simultaneous with *C* (simultaneous in the sense of the above definition), then the criterion for the simultaneity of the pair of events *A*, *C* is also satisfied. This assumption is a physical hypothesis about the law of propagation of light; it must certainly be fulfilled if we are to maintain the law of the constancy of the velocity of light *in vacuo*.

hypothesis, the validity of which will hardly be doubted without empirical evidence to the contrary. It has been assumed that all these clocks go *at the same rate* if they are of identical construction. Stated more exactly: When two clocks arranged at rest in different places of a reference-body are set in such a manner that a *particular* position of the pointers of the one clock is *simultaneous* (in the above sense) with the *same* position of the pointers of the other clock, then identical “settings” are always simultaneous (in the sense of the above definition).

IX

THE RELATIVITY OF SIMULTANEITY

UP to now our considerations have been referred to a particular body of reference, which we have styled a "railway embankment." We suppose a very long train travelling along the rails with the constant velocity v and in the direction indicated in Fig. 1. People travelling in this train will with advantage use the train as a rigid reference-body (co-ordinate system); they regard all events in reference to

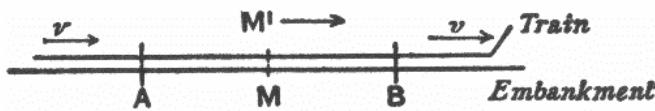


FIG. I.

the train. Then every event which takes place along the line also takes place at a particular point of the train. Also the definition of simultaneity can be given relative to the train in exactly the same way as with respect to the embankment. As a natural consequence, however, the following question arises:

Are two events (*e.g.* the two strokes of lightning A and B) which are simultaneous *with reference to*

the railway embankment also simultaneous relatively to the train? We shall show directly that the answer must be in the negative.

When we say that the lightning strokes A and B are simultaneous with respect to the embankment, we mean: the rays of light emitted at the places A and B , where the lightning occurs, meet each other at the mid-point M of the length $A \rightarrow B$ of the embankment. But the events A and B also correspond to positions A and B on the train. Let M' be the mid-point of the distance $A \rightarrow B$ on the travelling train. Just when the flashes¹ of lightning occur, this point M' naturally coincides with the point M , but it moves towards the right in the diagram with the velocity v of the train. If an observer sitting in the position M' in the train did not possess this velocity, then he would remain permanently at M , and the light rays emitted by the flashes of lightning A and B would reach him simultaneously, *i.e.* they would meet just where he is situated. Now in reality (considered with reference to the railway embankment) he is hastening towards the beam of light coming from B , whilst he is riding on ahead of the beam of light coming from A . Hence the observer will see the beam of light emitted from B earlier than he will see that emitted from A . Observers who take the railway train as their reference-body

¹ As judged from the embankment.

32 SPECIAL THEORY OF RELATIVITY

must therefore come to the conclusion that the lightning flash *B* took place earlier than the lightning flash *A*. We thus arrive at the important result:

Events which are simultaneous with reference to the embankment are not simultaneous with respect to the train, and *vice versa* (relativity of simultaneity). Every reference-body (co-ordinate system) has its own particular time; unless we are told the reference-body to which the statement of time refers, there is no meaning in a statement of the time of an event.

Now before the advent of the theory of relativity it had always tacitly been assumed in physics that the statement of time had an absolute significance, *i.e.* that it is independent of the state of motion of the body of reference. But we have just seen that this assumption is incompatible with the most natural definition of simultaneity; if we discard this assumption, then the conflict between the law of the propagation of light *in vacuo* and the principle of relativity (developed in Section VII) disappears.

We were led to that conflict by the considerations of Section VI, which are now no longer tenable. In that section we concluded that the man in the carriage, who traverses the distance w *per second* relative to the carriage, traverses the same distance also with respect to the embank-

ment *in each second* of time. But, according to the foregoing considerations, the time required by a particular occurrence with respect to the carriage must not be considered equal to the duration of the same occurrence as judged from the embankment (as reference-body). Hence it cannot be contended that the man in walking travels the distance w relative to the railway line in a time which is equal to one second as judged from the embankment.

Moreover, the considerations of Section VI are based on yet a second assumption, which, in the light of a strict consideration, appears to be arbitrary, although it was always tacitly made even before the introduction of the theory of relativity.

X

**ON THE RELATIVITY OF THE CONCEPTION
OF DISTANCE**

LET us consider two particular points on the train¹ travelling along the embankment with the velocity v , and inquire as to their distance apart. We already know that it is necessary to have a body of reference for the measurement of a distance, with respect to which body the distance can be measured up. It is the simplest plan to use the train itself as the reference-body (co-ordinate system). An observer in the train measures the interval by marking off his measuring-rod in a straight line (*e.g.* along the floor of the carriage) as many times as is necessary to take him from the one marked point to the other. Then the number which tells us how often the rod has to be laid down is the required distance.

It is a different matter when the distance has to be judged from the railway line. Here the following method suggests itself. If we call A' and B' the two points on the train whose distance apart is required, then both of these points are

¹ *e.g.* the middle of the first and of the hundredth carriage.

moving with the velocity v along the embankment. In the first place we require to determine the points A and B of the embankment which are just being passed by the two points A' and B' at a particular time t — judged from the embankment. These points A and B of the embankment can be determined by applying the definition of time given in Section VIII. The distance between these points A and B is then measured by repeated application of the measuring-rod along the embankment.

A priori it is by no means certain that this last measurement will supply us with the same result as the first. Thus the length of the train as measured from the embankment may be different from that obtained by measuring in the train itself. This circumstance leads us to a second objection which must be raised against the apparently obvious consideration of Section VI. Namely, if the man in the carriage covers the distance w in a unit of time — *measured from the train*, — then this distance — *as measured from the embankment* — is not necessarily also equal to w .

XI

THE LORENTZ TRANSFORMATION

THE results of the last three sections show that the apparent incompatibility of the law of propagation of light with the principle of relativity (Section VII) has been derived by means of a consideration which borrowed two unjustifiable hypotheses from classical mechanics; these are as follows:

- (1) The time-interval (time) between two events is independent of the condition of motion of the body of reference.
- (2) The space-interval (distance) between two points of a rigid body is independent of the condition of motion of the body of reference.

If we drop these hypotheses, then the dilemma of Section VII disappears, because the theorem of the addition of velocities derived in Section VI becomes invalid. The possibility presents itself that the law of the propagation of light *in vacuo* may be compatible with the principle of relativity, and the question arises: How have we to modify the considerations of Section VI in order to remove

THE LORENTZ TRANSFORMATION 37

the apparent disagreement between these two fundamental results of experience? This question leads to a general one. In the discussion of Section VI we have to do with places and times relative both to the train and to the embankment. How are we to find the place and time of an event in relation to the train, when we know the place and time of the event with respect to the railway embankment? Is there a thinkable answer to this question of such a nature that the law of transmission of light *in vacuo* does not contradict the principle of relativity? In other words: Can we conceive of a relation between place and time of the individual events relative to both reference-bodies, such that every ray of light possesses the velocity of transmission c relative to the embankment and relative to the train? This question leads to a quite definite positive answer, and to a perfectly definite transformation law for the space-time magnitudes of an event when changing over from one body of reference to another.

Before we deal with this, we shall introduce the following incidental consideration. Up to the present we have only considered events taking place along the embankment, which had mathematically to assume the function of a straight line. In the manner indicated in Section II we can imagine this reference-body supplemented laterally and in a vertical direction by means of a

38 SPECIAL THEORY OF RELATIVITY

framework of rods, so that an event which takes place anywhere can be localised with reference to this framework. Similarly, we can imagine the train travelling with the velocity v to be continued across the whole of space, so that every event, no matter how far off it may be, could also be localised with respect to the second framework. Without committing any fundamental error, we can disregard the fact that in reality these frameworks would continually interfere with each other, owing to the impenetrability of solid bodies. In every such framework we imagine three surfaces perpendicular to each other marked out, and designated as "co-ordinate planes" ("co-ordinate system"). A co-ordinate system K then corresponds to the embankment, and a co-ordinate system K' to the train. An event, wherever it may have taken place, would be fixed in space with respect to K by the three perpendicularly x, y, z on the co-ordinate planes, and with regard to time by a time-value t . Relative to K' , *the same event* would be fixed in respect of space and time by corresponding values x', y', z', t' , which of course are not identical with x, y, z, t . It has already been set forth in detail how these magnitudes are to be regarded as results of physical measurements.

Obviously our problem can be exactly formulated in the following manner. What are the

THE LORENTZ TRANSFORMATION 39

values x' , y' , z' , t' of an event with respect to K' , when the magnitudes x , y , z , t , of the same event with respect to K are given? The relations must be so chosen that the law of the transmission of light *in vacuo* is satisfied for one and the same ray of light (and of course for every ray) with respect to K and K' . For the relative orientation in space of the co-ordinate systems indicated in the diagram (Fig. 2), this problem is solved by means of the equations:

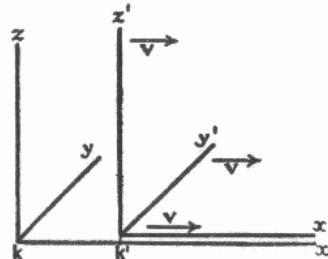


FIG. 2.

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$y' = y$$

$$z' = z$$

$$t' = \frac{t - \frac{v}{c^2} \cdot x}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

This system of equations is known as the "Lorentz transformation."¹

If in place of the law of transmission of light we had taken as our basis the tacit assumptions of the older mechanics as to the absolute character

¹ A simple derivation of the Lorentz transformation is given in Appendix I.

40 SPECIAL THEORY OF RELATIVITY

of times and lengths, then instead of the above we should have obtained the following equations:

$$\begin{aligned}x' &= x - vt \\y' &= y \\z' &= z \\t' &= t.\end{aligned}$$

This system of equations is often termed the “Galilei transformation.” The Galilei transformation can be obtained from the Lorentz transformation by substituting an infinitely large value for the velocity of light c in the latter transformation.

Aided by the following illustration, we can readily see that, in accordance with the Lorentz transformation, the law of the transmission of light *in vacuo* is satisfied both for the reference-body K and for the reference-body K' . A light-signal is sent along the positive x -axis, and this light-stimulus advances in accordance with the equation

$$x = ct,$$

i.e. with the velocity c . According to the equations of the Lorentz transformation, this simple relation between x and t involves a relation between x' and t' . In point of fact, if we substitute for x the value ct in the first and fourth equations of the Lorentz transformation, we obtain:

$$x' = \frac{(c-v)t}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$t' = \frac{\left(1 - \frac{v}{c}\right)t}{\sqrt{1 - \frac{v^2}{c^2}}},$$

from which, by division, the expression

$$x' = ct'$$

immediately follows. If referred to the system K' , the propagation of light takes place according to this equation. We thus see that the velocity of transmission relative to the reference-body K' is also equal to c . The same result is obtained for rays of light advancing in any other direction whatsoever. Of course this is not surprising, since the equations of the Lorentz transformation were derived conformably to this point of view.

XII

THE BEHAVIOUR OF MEASURING-RODS AND CLOCKS IN MOTION

I PLACE a metre-rod in the x' -axis of K' in such a manner that one end (the beginning) coincides with the point $x' = 0$, whilst the other end (the end of the rod) coincides with the point $x' = 1$. What is the length of the metre-rod relatively to the system K ? In order to learn this, we need only ask where the beginning of the rod and the end of the rod lie with respect to K at a particular time t of the system K . By means of the first equation of the Lorentz transformation the values of these two points at the time $t = 0$ can be shown to be

$$x_{(\text{beginning of rod})} = 0 \cdot \sqrt{1 - \frac{v^2}{c^2}}$$

$$x_{(\text{end of rod})} = 1 \cdot \sqrt{1 - \frac{v^2}{c^2}},$$

the distance between the points being $\sqrt{1 - \frac{v^2}{c^2}}$.

But the metre-rod is moving with the velocity v relative to K . It therefore follows that the length of a rigid metre-rod moving in the direction of its length with a velocity v is $\sqrt{1 - v^2/c^2}$ of a metre. The rigid rod is thus shorter when in motion than

when at rest, and the more quickly it is moving, the shorter is the rod. For the velocity $v=c$ we should have $\sqrt{1-v^2/c^2} = 0$, and for still greater velocities the square-root becomes imaginary. From this we conclude that in the theory of relativity the velocity c plays the part of a limiting velocity, which can neither be reached nor exceeded by any real body.

Of course this feature of the velocity c as a limiting velocity also clearly follows from the equations of the Lorentz transformation, for these become meaningless if we choose values of v greater than c .

If, on the contrary, we had considered a metre-rod at rest in the x -axis with respect to K , then we should have found that the length of the rod as judged from K' would have been $\sqrt{1-v^2/c^2}$; this is quite in accordance with the principle of relativity which forms the basis of our considerations.

A priori it is quite clear that we must be able to learn something about the physical behaviour of measuring-rods and clocks from the equations of transformation, for the magnitudes x, y, z, t , are nothing more nor less than the results of measurements obtainable by means of measuring-rods and clocks. If we had based our considerations on the Galilei transformation we should not have obtained a contraction of the rod as a consequence of its motion.

44 SPECIAL THEORY OF RELATIVITY

Let us now consider a seconds-clock which is permanently situated at the origin ($x' = 0$) of K' . $t' = 0$ and $t' = 1$ are two successive ticks of this clock. The first and fourth equations of the Lorentz transformation give for these two ticks:

$$t = 0$$

and

$$t = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

As judged from K , the clock is moving with the velocity v ; as judged from this reference-body, the time which elapses between two strokes of the clock is not one second, but $\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$ seconds, *i.e.*

a somewhat larger time. As a consequence of its motion the clock goes more slowly than when at rest. Here also the velocity c plays the part of an unattainable limiting velocity.

XIII

**THEOREM OF THE ADDITION OF VELOCITIES.
THE EXPERIMENT OF FIZEAU**

NOW in practice we can move clocks and measuring-rods only with velocities that are small compared with the velocity of light; hence we shall hardly be able to compare the results of the previous section directly with the reality. But, on the other hand, these results must strike you as being very singular, and for that reason I shall now draw another conclusion from the theory, one which can easily be derived from the foregoing considerations, and which has been most elegantly confirmed by experiment.

In Section VI we derived the theorem of the addition of velocities in one direction in the form which also results from the hypotheses of classical mechanics. This theorem can also be deduced readily from the Galilei transformation (Section XI). In place of the man walking inside the carriage, we introduce a point moving relatively to the co-ordinate system K' in accordance with the equation

$$x' = wt'.$$

By means of the first and fourth equations of the

46 SPECIAL THEORY OF RELATIVITY

Galilei transformation we can express x' and t' in terms of x and t , and we then obtain

$$x = (v + w)t.$$

This equation expresses nothing else than the law of motion of the point with reference to the system K (of the man with reference to the embankment). We denote this velocity by the symbol W , and we then obtain, as in Section VI,

$$W = v + w \dots \dots \dots \quad (\text{A}).$$

But we can carry out this consideration just as well on the basis of the theory of relativity. In the equation

$$x' = wt'$$

we must then express x' and t' in terms of x and t , making use of the first and fourth equations of the *Lorentz transformation*. Instead of the equation (A) we then obtain the equation

$$W = \frac{v + w}{1 + \frac{vw}{c^2}} \dots \dots \dots \quad (\text{B}),$$

which corresponds to the theorem of addition for velocities in one direction according to the theory of relativity. The question now arises as to which of these two theorems is the better in accord with experience. On this point we are enlightened by a most important experiment which the brilliant physicist Fizeau performed more than half a century ago, and which has been repeated since

then by some of the best experimental physicists, so that there can be no doubt about its result. The experiment is concerned with the following question. Light travels in a motionless liquid with a particular velocity w . How quickly does it travel in the direction of the arrow in the tube T (see the accompanying diagram, Fig. 3) when the liquid above mentioned is flowing through the tube with a velocity v ?

In accordance with the principle of relativity we shall certainly have to take for granted that the propagation of light always takes place with the same velocity w *with respect to the liquid*, whether the latter is in motion with reference to other bodies or not. The velocity of light relative to the liquid and the velocity of the latter relative to the tube are thus known, and we require the velocity of light relative to the tube.

It is clear that we have the problem of Section VI again before us. The tube plays the part of

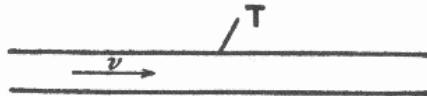


FIG. 3

the railway embankment or of the co-ordinate system K , the liquid plays the part of the carriage or of the co-ordinate system K' , and finally, the light plays the part of the man walking along the carriage, or of the moving point in the present

48 SPECIAL THEORY OF RELATIVITY

section. If we denote the velocity of the light relative to the tube by W , then this is given by the equation (A) or (B), according as the Galilei transformation or the Lorentz transformation corresponds to the facts. Experiment¹ decides in favour of equation (B) derived from the theory of relativity, and the agreement is, indeed, very exact. According to recent and most excellent measurements by Zeeman, the influence of the velocity of flow v on the propagation of light is represented by formula (B) to within one per cent.

Nevertheless we must now draw attention to the fact that a theory of this phenomenon was given by H. A. Lorentz long before the statement of the theory of relativity. This theory was of a purely electrodynamical nature, and was obtained by the use of particular hypotheses as to the electromagnetic structure of matter. This circumstance, however, does not in the least diminish the conclusiveness of the experiment as a crucial test in favour of the theory of relativity, for the

¹ Fizeau found $W = w + v \left(1 - \frac{1}{n^2}\right)$, where $n = \frac{c}{w}$ is the index of refraction of the liquid. On the other hand, owing to the smallness of $\frac{vw}{c^2}$ as compared with 1, we can replace (B) in the first place by $W = (w + v) \left(1 - \frac{vw}{c^2}\right)$, or to the same order of approximation by $w + v \left(1 - \frac{1}{n^2}\right)$, which agrees with Fizeau's result.

electrodynamics of Maxwell-Lorentz, on which the original theory was based, in no way opposes the theory of relativity. Rather has the latter been developed from electrodynamics as an astoundingly simple combination and generalisation of the hypotheses, formerly independent of each other, on which electrodynamics was built.

XIV

**THE HEURISTIC VALUE OF THE THEORY OF
RELATIVITY**

OUR train of thought in the foregoing pages can be epitomised in the following manner.

Experience has led to the conviction that, on the one hand, the principle of relativity holds true, and that on the other hand the velocity of transmission of light *in vacuo* has to be considered equal to a constant c . By uniting these two postulates we obtained the law of transformation for the rectangular co-ordinates x , y , z and the time t of the events which constitute the processes of nature. In this connection we did not obtain the Galilei transformation, but, differing from classical mechanics, the *Lorentz transformation*.

The law of transmission of light, the acceptance of which is justified by our actual knowledge, played an important part in this process of thought. Once in possession of the Lorentz transformation, however, we can combine this with the principle of relativity, and sum up the theory thus:

Every general law of nature must be so constituted that it is transformed into a law of exactly the same form when, instead of the space-

HEURISTIC VALUE OF RELATIVITY 51

time variables x, y, z, t of the original co-ordinate system K , we introduce new space-time variables x', y', z', t' of a co-ordinate system K' . In this connection the relation between the ordinary and the accented magnitudes is given by the Lorentz transformation. Or, in brief: General laws of nature are co-variant with respect to Lorentz transformations.

This is a definite mathematical condition that the theory of relativity demands of a natural law, and in virtue of this, the theory becomes a valuable heuristic aid in the search for general laws of nature. If a general law of nature were to be found which did not satisfy this condition, then at least one of the two fundamental assumptions of the theory would have been disproved. Let us now examine what general results the latter theory has hitherto evinced.

XV

GENERAL RESULTS OF THE THEORY

IT is clear from our previous considerations that the (special) theory of relativity has grown out of electrodynamics and optics. In these fields it has not appreciably altered the predictions of theory, but it has considerably simplified the theoretical structure, *i.e.* the derivation of laws, and — what is incomparably more important — it has considerably reduced the number of independent hypotheses forming the basis of theory. The special theory of relativity has rendered the Maxwell-Lorentz theory so plausible, that the latter would have been generally accepted by physicists even if experiment had decided less unequivocally in its favour.

Classical mechanics required to be modified before it could come into line with the demands of the special theory of relativity. For the main part, however, this modification affects only the laws for rapid motions, in which the velocities of matter v are not very small as compared with the velocity of light. We have experience of such rapid motions only in the case of electrons and

GENERAL RESULTS OF THEORY 53

ions; for other motions the variations from the laws of classical mechanics are too small to make themselves evident in practice. We shall not consider the motion of stars until we come to speak of the general theory of relativity. In accordance with the theory of relativity the kinetic energy of a material point of mass m is no longer given by the well-known expression

$$m \frac{v^2}{2},$$

but by the expression

$$\frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}.$$
*

This expression approaches infinity as the velocity v approaches the velocity of light c . The velocity must therefore always remain less than c , however great may be the energies used to produce the acceleration. If we develop the expression for the kinetic energy in the form of a series, we obtain

$$mc^2 + m \frac{v^2}{2} + \frac{3}{8} m \frac{v^4}{c^2} + \dots$$

When $\frac{v^2}{c^2}$ is small compared with unity, the third of these terms is always small in comparison with the second, which last is alone considered in classical mechanics. The first term mc^2 does not contain the velocity, and requires no consideration if we

$$[{}^* \sqrt{\frac{mc^2}{1 - \frac{v^2}{c^2}}} — J.M.]$$

54 SPECIAL THEORY OF RELATIVITY

are only dealing with the question as to how the energy of a point-mass depends on the velocity. We shall speak of its essential significance later.

The most important result of a general character to which the special theory of relativity has led is concerned with the conception of mass. Before the advent of relativity, physics recognised two conservation laws of fundamental importance, namely, the law of the conservation of energy and the law of the conservation of mass; these two fundamental laws appeared to be quite independent of each other. By means of the theory of relativity they have been united into one law. We shall now briefly consider how this unification came about, and what meaning is to be attached to it.

The principle of relativity requires that the law of the conservation of energy should hold not only with reference to a co-ordinate system K , but also with respect to every co-ordinate system K' which is in a state of uniform motion of translation relative to K , or, briefly, relative to every "Galileian" system of co-ordinates. In contrast to classical mechanics, the Lorentz transformation is the deciding factor in the transition from one such system to another.

By means of comparatively simple considerations we are led to draw the following conclusion from these premises, in conjunction with the

fundamental equations of the electrodynamics of Maxwell: A body moving with the velocity v , which absorbs¹ an amount of energy E_0 in the form of radiation without suffering an alteration in velocity in the process, has, as a consequence, its energy increased by an amount

$$\frac{E_0}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

In consideration of the expression given above for the kinetic energy of the body, the required energy of the body comes out to be

$$\frac{\left(m + \frac{E_0}{c^2}\right)c^2}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Thus the body has the same energy as a body of mass $\left(m + \frac{E_0}{c^2}\right)$ moving with the velocity v .

Hence we can say: If a body takes up an amount of energy E_0 , then its inertial mass increases by an amount $\frac{E_0}{c^2}$; the inertial mass of a body is not a constant, but varies according to the change in the energy of the body. The inertial mass of a system of bodies can even be regarded as a measure

¹ E_0 is the energy taken up, as judged from a co-ordinate system moving with the body.

$$[\frac{E_0}{\sqrt{1 - \frac{v^2}{c^2}}} — J.M.]$$

56 SPECIAL THEORY OF RELATIVITY

of its energy. The law of the conservation of the mass of a system becomes identical with the law of the conservation of energy, and is only valid provided that the system neither takes up nor sends out energy. Writing the expression for the energy in the form

$$\frac{mc^2 + E_0}{\sqrt{1 - \frac{v^2}{c^2}}},$$

we see that the term mc^2 , which has hitherto attracted our attention, is nothing else than the energy possessed by the body¹ before it absorbed the energy E_0 .

A direct comparison of this relation with experiment is not possible at the present time, owing to the fact that the changes in energy E_0 to which we can subject a system are not large enough to make themselves perceptible as a change in the inertial mass of the system. $\frac{E_0}{c^2}$ is too small in comparison with the mass m , which was present before the alteration of the energy. It is owing to this circumstance that classical mechanics was able to establish successfully the conservation of mass as a law of independent validity.

Let me add a final remark of a fundamental nature. The success of the Faraday-Maxwell

¹ As judged from a co-ordinate system moving with the body.

GENERAL RESULTS OF THEORY 57

interpretation of electromagnetic action at a distance resulted in physicists becoming convinced that there are no such things as instantaneous actions at a distance (not involving an intermediary medium) of the type of Newton's law of gravitation. According to the theory of relativity, action at a distance with the velocity of light always takes the place of instantaneous action at a distance or of action at a distance with an infinite velocity of transmission. This is connected with the fact that the velocity c plays a fundamental rôle in this theory. In Part II we shall see in what way this result becomes modified in the general theory of relativity.

XVI**EXPERIENCE AND THE SPECIAL THEORY
OF RELATIVITY**

TO what extent is the special theory of relativity supported by experience? This question is not easily answered for the reason already mentioned in connection with the fundamental experiment of Fizeau. The special theory of relativity has crystallised out from the Maxwell-Lorentz theory of electromagnetic phenomena. Thus all facts of experience which support the electromagnetic theory also support the theory of relativity. As being of particular importance, I mention here the fact that the theory of relativity enables us to predict the effects produced on the light reaching us from the fixed stars. These results are obtained in an exceedingly simple manner, and the effects indicated, which are due to the relative motion of the earth with reference to those fixed stars, are found to be in accord with experience. We refer to the yearly movement of the apparent position of the fixed stars resulting from the motion of the earth round the sun (aberration), and to the influence of the radial

components of the relative motions of the fixed stars with respect to the earth on the colour of the light reaching us from them. The latter effect manifests itself in a slight displacement of the spectral lines of the light transmitted to us from a fixed star, as compared with the position of the same spectral lines when they are produced by a terrestrial source of light (Doppler principle). The experimental arguments in favour of the Maxwell-Lorentz theory, which are at the same time arguments in favour of the theory of relativity, are too numerous to be set forth here. In reality they limit the theoretical possibilities to such an extent, that no other theory than that of Maxwell and Lorentz has been able to hold its own when tested by experience.

But there are two classes of experimental facts hitherto obtained which can be represented in the Maxwell-Lorentz theory only by the introduction of an auxiliary hypothesis, which in itself — *i.e.* without making use of the theory of relativity — appears extraneous.

It is known that cathode rays and the so-called β -rays emitted by radioactive substances consist of negatively electrified particles (electrons) of very small inertia and large velocity. By examining the deflection of these rays under the influence of electric and magnetic fields, we can study the law of motion of these particles very exactly.

60 SPECIAL THEORY OF RELATIVITY

In the theoretical treatment of these electrons, we are faced with the difficulty that electrodynamic theory of itself is unable to give an account of their nature. For since electrical masses of one sign repel each other, the negative electrical masses constituting the electron would necessarily be scattered under the influence of their mutual repulsions, unless there are forces of another kind operating between them, the nature of which has hitherto remained obscure to us.¹ If we now assume that the relative distances between the electrical masses constituting the electron remain unchanged during the motion of the electron (rigid connection in the sense of classical mechanics), we arrive at a law of motion of the electron which does not agree with experience. Guided by purely formal points of view, H. A. Lorentz was the first to introduce the hypothesis that the particles constituting the electron experience a contraction in the direction of motion in consequence of that motion, the amount of this contraction being proportional to the expression

$\sqrt{1 - \frac{v^2}{c^2}}$. * This hypothesis, which is not justifiable by any electrodynamical facts, supplies us then with that particular law of motion which has been confirmed with great precision in recent years.

¹ The general theory of relativity renders it likely that the electrical masses of an electron are held together by gravitational forces.

$$[{}^* \sqrt{1 - \frac{v^2}{c^2}} — J.M.]$$

The theory of relativity leads to the same law of motion, without requiring any special hypothesis whatsoever as to the structure and the behaviour of the electron. We arrived at a similar conclusion in Section XIII in connection with the experiment of Fizeau, the result of which is foretold by the theory of relativity without the necessity of drawing on hypotheses as to the physical nature of the liquid.

The second class of facts to which we have alluded has reference to the question whether or not the motion of the earth in space can be made perceptible in terrestrial experiments. We have already remarked in Section V that all attempts of this nature led to a negative result. Before the theory of relativity was put forward, it was difficult to become reconciled to this negative result, for reasons now to be discussed. The inherited prejudices about time and space did not allow any doubt to arise as to the prime importance of the Galilei transformation for changing over from one body of reference to another. Now assuming that the Maxwell-Lorentz equations hold for a reference-body K , we then find that they do not hold for a reference-body K' moving uniformly with respect to K , if we assume that the relations of the Galileian transformation exist between the co-ordinates of K and K' . It thus appears that of all Galileian co-ordinate

62 SPECIAL THEORY OF RELATIVITY

systems one (K) corresponding to a particular state of motion is physically unique. This result was interpreted physically by regarding K as at rest with respect to a hypothetical æther of space. On the other hand, all co-ordinate systems K' moving relatively to K were to be regarded as in motion with respect to the æther. To this motion of K' against the æther ("æther-drift" relative to K') were assigned the more complicated laws which were supposed to hold relative to K' . Strictly speaking, such an æther-drift ought also to be assumed relative to the earth, and for a long time the efforts of physicists were devoted to attempts to detect the existence of an æther-drift at the earth's surface.

In one of the most notable of these attempts Michelson devised a method which appears as though it must be decisive. Imagine two mirrors so arranged on a rigid body that the reflecting surfaces face each other. A ray of light requires a perfectly definite time T to pass from one mirror to the other and back again, if the whole system be at rest with respect to the æther. It is found by calculation, however, that a slightly different time T' is required for this process, if the body, together with the mirrors, be moving relatively to the æther. And yet another point: it is shown by calculation that for a given velocity v with reference to the æther, this time T' is different

when the body is moving perpendicularly to the planes of the mirrors from that resulting when the motion is parallel to these planes. Although the estimated difference between these two times is exceedingly small, Michelson and Morley performed an experiment involving interference in which this difference should have been clearly detectable. But the experiment gave a negative result — a fact very perplexing to physicists. Lorentz and FitzGerald rescued the theory from this difficulty by assuming that the motion of the body relative to the æther produces a contraction of the body in the direction of motion, the amount of contraction being just sufficient to compensate for the difference in time mentioned above. Comparison with the discussion in Section XII shows that from the standpoint also of the theory of relativity this solution of the difficulty was the right one. But on the basis of the theory of relativity the method of interpretation is incomparably more satisfactory. According to this theory there is no such thing as a “specially favoured” (unique) co-ordinate system to occasion the introduction of the æther-idea, and hence there can be no æther-drift, nor any experiment with which to demonstrate it. Here the contraction of moving bodies follows from the two fundamental principles of the theory without the introduction of particular hypotheses; and as the

64 SPECIAL THEORY OF RELATIVITY

prime factor involved in this contraction we find, not the motion in itself, to which we cannot attach any meaning, but the motion with respect to the body of reference chosen in the particular case in point. Thus for a co-ordinate system moving with the earth the mirror system of Michelson and Morley is not shortened, but it *is* shortened for a co-ordinate system which is at rest relatively to the sun.

XVII

MINKOWSKI'S FOUR-DIMENSIONAL SPACE

THE non-mathematician is seized by a mysterious shuddering when he hears of “four-dimensional” things, by a feeling not unlike that awakened by thoughts of the occult. And yet there is no more common-place statement than that the world in which we live is a four-dimensional space-time continuum.

Space is a three-dimensional continuum. By this we mean that it is possible to describe the position of a point (at rest) by means of three numbers (co-ordinates) x, y, z , and that there is an indefinite number of points in the neighbourhood of this one, the position of which can be described by co-ordinates such as x_1, y_1, z_1 , which may be as near as we choose to the respective values of the co-ordinates x, y, z of the first point. In virtue of the latter property we speak of a “continuum,” and owing to the fact that there are three co-ordinates we speak of it as being “three-dimensional.”

Similarly, the world of physical phenomena which was briefly called “world” by Minkowski

66 SPECIAL THEORY OF RELATIVITY

is naturally four-dimensional in the space-time sense. For it is composed of individual events, each of which is described by four numbers, namely, three space co-ordinates x, y, z and a time co-ordinate, the time-value t . The “world” is in this sense also a continuum; for to every event there are as many “neighbouring” events (realised or at least thinkable) as we care to choose, the co-ordinates x_1, y_1, z_1, t_1 of which differ by an indefinitely small amount from those of the event x, y, z, t originally considered. That we have not been accustomed to regard the world in this sense as a four-dimensional continuum is due to the fact that in physics, before the advent of the theory of relativity, time played a different and more independent rôle, as compared with the space co-ordinates. It is for this reason that we have been in the habit of treating time as an independent continuum. As a matter of fact, according to classical mechanics, time is absolute, *i.e.* it is independent of the position and the condition of motion of the system of co-ordinates. We see this expressed in the last equation of the Galileian transformation ($t' = t$).

The four-dimensional mode of consideration of the “world” is natural on the theory of relativity, since according to this theory time is robbed of its independence. This is shown by the fourth equation of the Lorentz transformation:

$$t' = \frac{t - \frac{v}{c^2} x}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Moreover, according to this equation the time difference $\Delta t'$ of two events with respect to K' does not in general vanish, even when the time difference Δt of the same events with reference to K vanishes. Pure "space-distance" of two events with respect to K results in "time-distance" of the same events with respect to K' . But the discovery of Minkowski, which was of importance for the formal development of the theory of relativity, does not lie here. It is to be found rather in the fact of his recognition that the four-dimensional space-time continuum of the theory of relativity, in its most essential formal properties, shows a pronounced relationship to the three-dimensional continuum of Euclidean geometrical space.¹ In order to give due prominence to this relationship, however, we must replace the usual time co-ordinate t by an imaginary magnitude $\sqrt{-1} \cdot ct$ proportional to it. Under these conditions, the natural laws satisfying the demands of the (special) theory of relativity assume mathematical forms, in which the time co-ordinate plays exactly the same rôle as the three space co-ordinates. Formally, these four co-ordinates

¹ Cf. the somewhat more detailed discussion in Appendix II.

68 SPECIAL THEORY OF RELATIVITY

correspond exactly to the three space co-ordinates in Euclidean geometry. It must be clear even to the non-mathematician that, as a consequence of this purely formal addition to our knowledge, the theory perforce gained clearness in no mean measure.

These inadequate remarks can give the reader only a vague notion of the important idea contributed by Minkowski. Without it the general theory of relativity, of which the fundamental ideas are developed in the following pages, would perhaps have got no farther than its long clothes. Minkowski's work is doubtless difficult of access to anyone inexperienced in mathematics, but since it is not necessary to have a very exact grasp of this work in order to understand the fundamental ideas of either the special or the general theory of relativity, I shall at present leave it here, and shall revert to it only towards the end of Part II.

PART II

THE GENERAL THEORY OF RELATIVITY

XVIII

SPECIAL AND GENERAL PRINCIPLE OF RELATIVITY

THE basal principle, which was the pivot of all our previous considerations, was the *special* principle of relativity, *i.e.* the principle of the physical relativity of all *uniform* motion. Let us once more analyse its meaning carefully.

It was at all times clear that, from the point of view of the idea it conveys to us, every motion must only be considered as a relative motion. Returning to the illustration we have frequently used of the embankment and the railway carriage, we can express the fact of the motion here taking place in the following two forms, both of which are equally justifiable:

- (a) The carriage is in motion relative to the embankment.
- (b) The embankment is in motion relative to the carriage.

In (a) the embankment, in (b) the carriage, serves as the body of reference in our statement

70 GENERAL THEORY OF RELATIVITY

of the motion taking place. If it is simply a question of detecting or of describing the motion involved, it is in principle immaterial to what reference-body we refer the motion. As already mentioned, this is self-evident, but it must not be confused with the much more comprehensive statement called “the principle of relativity,” which we have taken as the basis of our investigations.

The principle we have made use of not only maintains that we may equally well choose the carriage or the embankment as our reference-body for the description of any event (for this, too, is self-evident). Our principle rather asserts what follows: If we formulate the general laws of nature as they are obtained from experience, by making use of

- (a) the embankment as reference-body,
- (b) the railway carriage as reference-body,

then these general laws of nature (*e.g.* the laws of mechanics or the law of the propagation of light *in vacuo*) have exactly the same form in both cases. This can also be expressed as follows: For the *physical* description of natural processes, neither of the reference-bodies K , K' is unique (lit. “specially marked out”) as compared with the other. Unlike the first, this latter statement need not of necessity hold *a priori*; it is not contained in the conceptions of “motion” and “reference-

SPECIAL AND GENERAL PRINCIPLE 71

body" and derivable from them; only *experience* can decide as to its correctness or incorrectness.

Up to the present, however, we have by no means maintained the equivalence of *all* bodies of reference K in connection with the formulation of natural laws. Our course was more on the following lines. In the first place, we started out from the assumption that there exists a reference-body K , whose condition of motion is such that the Galileian law holds with respect to it: A particle left to itself and sufficiently far removed from all other particles moves uniformly in a straight line. With reference to K (Galileian reference-body) the laws of nature were to be as simple as possible. But in addition to K , all bodies of reference K' should be given preference in this sense, and they should be exactly equivalent to K for the formulation of natural laws, provided that they are in a state of *uniform rectilinear and non-rotary motion* with respect to K ; all these bodies of reference are to be regarded as Galileian reference-bodies. The validity of the principle of relativity was assumed only for these reference-bodies, but not for others (*e.g.* those possessing motion of a different kind). In this sense we speak of the *special* principle of relativity, or special theory of relativity.

In contrast to this we wish to understand by the "general principle of relativity" the following

72 GENERAL THEORY OF RELATIVITY

statement: All bodies of reference K , K' , etc., are equivalent for the description of natural phenomena (formulation of the general laws of nature), whatever may be their state of motion. But before proceeding farther, it ought to be pointed out that this formulation must be replaced later by a more abstract one, for reasons which will become evident at a later stage.

Since the introduction of the special principle of relativity has been justified, every intellect which strives after generalisation must feel the temptation to venture the step towards the general principle of relativity. But a simple and apparently quite reliable consideration seems to suggest that, for the present at any rate, there is little hope of success in such an attempt. Let us imagine ourselves transferred to our old friend the railway carriage, which is travelling at a uniform rate. As long as it is moving uniformly, the occupant of the carriage is not sensible of its motion, and it is for this reason that he can unreluctantly interpret the facts of the case as indicating that the carriage is at rest, but the embankment in motion. Moreover, according to the special principle of relativity, this interpretation is quite justified also from a physical point of view.

If the motion of the carriage is now changed into a non-uniform motion, as for instance by a

SPECIAL AND GENERAL PRINCIPLE 73

powerful application of the brakes, then the occupant of the carriage experiences a correspondingly powerful jerk forwards. The retarded motion is manifested in the mechanical behaviour of bodies relative to the person in the railway carriage. The mechanical behaviour is different from that of the case previously considered, and for this reason it would appear to be impossible that the same mechanical laws hold relatively to the non-uniformly moving carriage, as hold with reference to the carriage when at rest or in uniform motion. At all events it is clear that the Galileian law does not hold with respect to the non-uniformly moving carriage. Because of this, we feel compelled at the present juncture to grant a kind of absolute physical reality to non-uniform motion, in opposition to the general principle of relativity. But in what follows we shall soon see that this conclusion cannot be maintained.

XIX**THE GRAVITATIONAL FIELD**

“**I**F we pick up a stone and then let it go, why does it fall to the ground?” The usual answer to this question is: “Because it is attracted by the earth.” Modern physics formulates the answer rather differently for the following reason. As a result of the more careful study of electromagnetic phenomena, we have come to regard action at a distance as a process impossible without the intervention of some intermediary medium. If, for instance, a magnet attracts a piece of iron, we cannot be content to regard this as meaning that the magnet acts directly on the iron through the intermediate empty space, but we are constrained to imagine — after the manner of Faraday — that the magnet always calls into being something physically real in the space around it, that something being what we call a “magnetic field.” In its turn this magnetic field operates on the piece of iron, so that the latter strives to move towards the magnet. We shall not discuss here the justification for this incidental conception, which is indeed a somewhat arbi-

trary one. We shall only mention that with its aid electromagnetic phenomena can be theoretically represented much more satisfactorily than without it, and this applies particularly to the transmission of electromagnetic waves. The effects of gravitation also are regarded in an analogous manner.

The action of the earth on the stone takes place indirectly. The earth produces in its surroundings a gravitational field, which acts on the stone and produces its motion of fall. As we know from experience, the intensity of the action on a body diminishes according to a quite definite law, as we proceed farther and farther away from the earth. From our point of view this means: The law governing the properties of the gravitational field in space must be a perfectly definite one, in order correctly to represent the diminution of gravitational action with the distance from operative bodies. It is something like this: The body (*e.g.* the earth) produces a field in its immediate neighbourhood directly; the intensity and direction of the field at points farther removed from the body are thence determined by the law which governs the properties in space of the gravitational fields themselves.

In contrast to electric and magnetic fields, the gravitational field exhibits a most remarkable property, which is of fundamental importance

76 GENERAL THEORY OF RELATIVITY

for what follows. Bodies which are moving under the sole influence of a gravitational field receive an acceleration, *which does not in the least depend either on the material or on the physical state of the body*. For instance, a piece of lead and a piece of wood fall in exactly the same manner in a gravitational field (*in vacuo*), when they start off from rest or with the same initial velocity. This law, which holds most accurately, can be expressed in a different form in the light of the following consideration.

According to Newton's law of motion, we have

$$\text{(Force)} = \text{(inertial mass)} \times \text{(acceleration)},$$

where the "inertial mass" is a characteristic constant of the accelerated body. If now gravitation is the cause of the acceleration, we then have

$$\text{(Force)} = \text{(gravitational mass)} \times \text{(intensity of the gravitational field)},$$

where the "gravitational mass" is likewise a characteristic constant for the body. From these two relations follows:

$$\text{(acceleration)} = \frac{\text{(gravitational mass)}}{\text{(inertial mass)}} \times \text{(intensity of the gravitational field)}.$$

If now, as we find from experience, the acceleration is to be independent of the nature and the condition of the body and always the same for a

given gravitational field, then the ratio of the gravitational to the inertial mass must likewise be the same for all bodies. By a suitable choice of units we can thus make this ratio equal to unity. We then have the following law: The *gravitational* mass of a body is equal to its *inertial* mass.

It is true that this important law had hitherto been recorded in mechanics, but it had not been *interpreted*. A satisfactory interpretation can be obtained only if we recognise the following fact: *The same* quality of a body manifests itself according to circumstances as “inertia” or as “weight” (lit. “heaviness”). In the following section we shall show to what extent this is actually the case, and how this question is connected with the general postulate of relativity.

XX

THE EQUALITY OF INERTIAL AND GRAVITATIONAL MASS AS AN ARGUMENT FOR THE GENERAL POSTULATE OF RELATIVITY

WE imagine a large portion of empty space, so far removed from stars and other appreciable masses that we have before us approximately the conditions required by the fundamental law of Galilei. It is then possible to choose a Galileian reference-body for this part of space (world), relative to which points at rest remain at rest and points in motion continue permanently in uniform rectilinear motion. As reference-body let us imagine a spacious chest resembling a room with an observer inside who is equipped with apparatus. Gravitation naturally does not exist for this observer. He must fasten himself with strings to the floor, otherwise the slightest impact against the floor will cause him to rise slowly towards the ceiling of the room.

To the middle of the lid of the chest is fixed externally a hook with rope attached, and now a "being" (what kind of a being is immaterial to

us) begins pulling at this with a constant force. The chest together with the observer then begin to move "upwards" with a uniformly accelerated motion. In course of time their velocity will reach unheard-of values — provided that we are viewing all this from another reference-body which is not being pulled with a rope.

But how does the man in the chest regard the process? The acceleration of the chest will be transmitted to him by the reaction of the floor of the chest. He must therefore take up this pressure by means of his legs if he does not wish to be laid out full length on the floor. He is then standing in the chest in exactly the same way as anyone stands in a room of a house on our earth. If he release a body which he previously had in his hand, the acceleration of the chest will no longer be transmitted to this body, and for this reason the body will approach the floor of the chest with an accelerated relative motion. The observer will further convince himself *that the acceleration of the body towards the floor of the chest is always of the same magnitude, whatever kind of body he may happen to use for the experiment.*

Relying on his knowledge of the gravitational field (as it was discussed in the preceding section), the man in the chest will thus come to the conclusion that he and the chest are in a gravitational field which is constant with regard to time. Of

80 GENERAL THEORY OF RELATIVITY

course he will be puzzled for a moment as to why the chest does not fall in this gravitational field. Just then, however, he discovers the hook in the middle of the lid of the chest and the rope which is attached to it, and he consequently comes to the conclusion that the chest is suspended at rest in the gravitational field.

Ought we to smile at the man and say that he errs in his conclusion? I do not believe we ought if we wish to remain consistent; we must rather admit that his mode of grasping the situation violates neither reason nor known mechanical laws. Even though it is being accelerated with respect to the "Galileian space" first considered, we can nevertheless regard the chest as being at rest. We have thus good grounds for extending the principle of relativity to include bodies of reference which are accelerated with respect to each other, and as a result we have gained a powerful argument for a generalised postulate of relativity.

We must note carefully that the possibility of this mode of interpretation rests on the fundamental property of the gravitational field of giving all bodies the same acceleration, or, what comes to the same thing, on the law of the equality of inertial and gravitational mass. If this natural law did not exist, the man in the accelerated chest would not be able to interpret the behaviour of

INERTIAL AND GRAVITATIONAL MASS 81

the bodies around him on the supposition of a gravitational field, and he would not be justified on the grounds of experience in supposing his reference-body to be "at rest."

Suppose that the man in the chest fixes a rope to the inner side of the lid, and that he attaches a body to the free end of the rope. The result of this will be to stretch the rope so that it will hang "vertically" downwards. If we ask for an opinion of the cause of tension in the rope, the man in the chest will say: "The suspended body experiences a downward force in the gravitational field, and this is neutralised by the tension of the rope; what determines the magnitude of the tension of the rope is the *gravitational mass* of the suspended body." On the other hand, an observer who is poised freely in space will interpret the condition of things thus: "The rope must perforce take part in the accelerated motion of the chest, and it transmits this motion to the body attached to it. The tension of the rope is just large enough to effect the acceleration of the body. That which determines the magnitude of the tension of the rope is the *inertial mass* of the body." Guided by this example, we see that our extension of the principle of relativity implies the *necessity* of the law of the equality of inertial and gravitational mass. Thus we have obtained a physical interpretation of this law.

82 GENERAL THEORY OF RELATIVITY

From our consideration of the accelerated chest we see that a general theory of relativity must yield important results on the laws of gravitation. In point of fact, the systematic pursuit of the general idea of relativity has supplied the laws satisfied by the gravitational field. Before proceeding farther, however, I must warn the reader against a misconception suggested by these considerations. A gravitational field exists for the man in the chest, despite the fact that there was no such field for the co-ordinate system first chosen. Now we might easily suppose that the existence of a gravitational field is always only an *apparent* one. We might also think that, regardless of the kind of gravitational field which may be present, we could always choose another reference-body such that *no* gravitational field exists with reference to it. This is by no means true for all gravitational fields, but only for those of quite special form. It is, for instance, impossible to choose a body of reference such that, as judged from it, the gravitational field of the earth (in its entirety) vanishes.

We can now appreciate why that argument is not convincing, which we brought forward against the general principle of relativity at the end of Section XVIII. It is certainly true that the observer in the railway carriage experiences a jerk forwards as a result of the application of the

INERTIAL AND GRAVITATIONAL MASS 83

brake, and that he recognises in this the non-uniformity of motion (retardation) of the carriage. But he is compelled by nobody to refer this jerk to a "real" acceleration (retardation) of the carriage. He might also interpret his experience thus: "My body of reference (the carriage) remains permanently at rest. With reference to it, however, there exists (during the period of application of the brakes) a gravitational field which is directed forwards and which is variable with respect to time. Under the influence of this field, the embankment together with the earth moves non-uniformly in such a manner that their original velocity in the backwards direction is continuously reduced."

XXI**IN WHAT RESPECTS ARE THE FOUNDATIONS
OF CLASSICAL MECHANICS AND OF THE
SPECIAL THEORY OF RELATIVITY UN-
SATISFACTORY?**

WE have already stated several times that classical mechanics starts out from the following law: Material particles sufficiently far removed from other material particles continue to move uniformly in a straight line or continue in a state of rest. We have also repeatedly emphasised that this fundamental law can only be valid for bodies of reference K which possess certain unique states of motion, and which are in uniform translational motion relative to each other. Relative to other reference-bodies K the law is not valid. Both in classical mechanics and in the special theory of relativity we therefore differentiate between reference-bodies K relative to which the recognised “laws of nature” can be said to hold, and reference-bodies K relative to which these laws do not hold.

But no person whose mode of thought is logical can rest satisfied with this condition of things. He asks: “How does it come that certain refer-

ence-bodies (or their states of motion) are given priority over other reference-bodies (or their states of motion)? *What is the reason for this preference?* In order to show clearly what I mean by this question, I shall make use of a comparison.

I am standing in front of a gas range. Standing alongside of each other on the range are two pans so much alike that one may be mistaken for the other. Both are half full of water. I notice that steam is being emitted continuously from the one pan, but not from the other. I am surprised at this, even if I have never seen either a gas range or a pan before. But if I now notice a luminous something of bluish colour under the first pan but not under the other, I cease to be astonished, even if I have never before seen a gas flame. For I can only say that this bluish something will cause the emission of the steam, or at least *possibly* it may do so. If, however, I notice the bluish something in neither case, and if I observe that the one continuously emits steam whilst the other does not, then I shall remain astonished and dissatisfied until I have discovered some circumstance to which I can attribute the different behaviour of the two pans.

Analogously, I seek in vain for a real something in classical mechanics (or in the special theory of relativity) to which I can attribute the different behaviour of bodies considered with respect to

86 GENERAL THEORY OF RELATIVITY

the reference-systems K and K' .¹ Newton saw this objection and attempted to invalidate it, but without success. But E. Mach recognised it most clearly of all, and because of this objection he claimed that mechanics must be placed on a new basis. It can only be got rid of by means of a physics which is conformable to the general principle of relativity, since the equations of such a theory hold for every body of reference, whatever may be its state of motion.

¹ The objection is of importance more especially when the state of motion of the reference-body is of such a nature that it does not require any external agency for its maintenance, *e.g.* in the case when the reference-body is rotating uniformly.

XXII

A FEW INFERENCES FROM THE GENERAL THEORY^{*} OF RELATIVITY

THE considerations of Section XX show that the general theory^{*} of relativity puts us in a position to derive properties of the gravitational field in a purely theoretical manner. Let us suppose, for instance, that we know the space-time “course” for any natural process whatsoever, as regards the manner in which it takes place in the Galileian domain relative to a Galileian body of reference K . By means of purely theoretical operations (*i.e.* simply by calculation) we are then able to find how this known natural process appears, as seen from a reference-body K' which is accelerated relatively to K . But since a gravitational field exists with respect to this new body of reference K' , our consideration also teaches us how the gravitational field influences the process studied.

For example, we learn that a body which is in a state of uniform rectilinear motion with respect to K (in accordance with the law of Galilei) is executing an accelerated and in general

[^{*} The word “theory” was changed to “principle” in both places in later editions. — J.M.]

88 GENERAL THEORY OF RELATIVITY

curvilinear motion with respect to the accelerated reference-body K' (chest). This acceleration or curvature corresponds to the influence on the moving body of the gravitational field prevailing relatively to K' . It is known that a gravitational field influences the movement of bodies in this way, so that our consideration supplies us with nothing essentially new.

However, we obtain a new result of fundamental importance when we carry out the analogous consideration for a ray of light. With respect to the Galilean reference-body K , such a ray of light is transmitted rectilinearly with the velocity c . It can easily be shown that the path of the same ray of light is no longer a straight line when we consider it with reference to the accelerated chest (reference-body K'). From this we conclude, *that, in general, rays of light are propagated curvilinearly in gravitational fields*. In two respects this result is of great importance.

In the first place, it can be compared with the reality. Although a detailed examination of the question shows that the curvature of light rays required by the general theory of relativity is only exceedingly small for the gravitational fields at our disposal in practice, its estimated magnitude for light rays passing the sun at grazing incidence is nevertheless 1.7 seconds of arc. This ought to manifest itself in the following way.

As seen from the earth, certain fixed stars appear to be in the neighbourhood of the sun, and are thus capable of observation during a total eclipse of the sun. At such times, these stars ought to appear to be displaced outwards from the sun by an amount indicated above, as compared with their apparent position in the sky when the sun is situated at another part of the heavens. The examination of the correctness or otherwise of this deduction is a problem of the greatest importance, the early solution of which is to be expected of astronomers.¹

In the second place our result shows that, according to the general theory of relativity, the law of the constancy of the velocity of light *in vacuo*, which constitutes one of the two fundamental assumptions in the special theory of relativity and to which we have already frequently referred, cannot claim any unlimited validity. A curvature of rays of light can only take place when the velocity of propagation of light varies with position. Now we might think that as a consequence of this, the special theory of relativity and with it the whole theory of relativity would be laid in the dust. But in reality this is not the

¹ By means of the star photographs of two expeditions equipped by a Joint Committee of the Royal and Royal Astronomical Societies, the existence of the deflection of light demanded by theory was confirmed during the solar eclipse of 29th May, 1919. (Cf. Appendix III.)

90 GENERAL THEORY OF RELATIVITY

case. We can only conclude that the special theory of relativity cannot claim an unlimited domain of validity; its results hold only so long as we are able to disregard the influences of gravitational fields on the phenomena (*e.g.* of light).

Since it has often been contended by opponents of the theory of relativity that the special theory of relativity is overthrown by the general theory of relativity, it is perhaps advisable to make the facts of the case clearer by means of an appropriate comparison. Before the development of electrodynamics the laws of electrostatics and the laws of electricity were regarded indiscriminately. At the present time we know that electric fields can be derived correctly from electrostatic considerations only for the case, which is never strictly realised, in which the electrical masses are quite at rest relatively to each other, and to the co-ordinate system. Should we be justified in saying that for this reason electrostatics is overthrown by the field-equations of Maxwell in electrodynamics? Not in the least. Electrostatics is contained in electrodynamics as a limiting case; the laws of the latter lead directly to those of the former for the case in which the fields are invariable with regard to time. No fairer destiny could be allotted to any physical theory, than that it should of itself point out the

way to the introduction of a more comprehensive theory, in which it lives on as a limiting case.

In the example of the transmission of light just dealt with, we have seen that the general theory of relativity enables us to derive theoretically the influence of a gravitational field on the course of natural processes, the laws of which are already known when a gravitational field is absent. But the most attractive problem, to the solution of which the general theory of relativity supplies the key, concerns the investigation of the laws satisfied by the gravitational field itself. Let us consider this for a moment.

We are acquainted with space-time domains which behave (approximately) in a “Galileian” fashion under suitable choice of reference-body, *i.e.* domains in which gravitational fields are absent. If we now refer such a domain to a reference-body K' possessing any kind of motion, then relative to K' there exists a gravitational field which is variable with respect to space and time.¹ The character of this field will of course depend on the motion chosen for K' . According to the general theory of relativity, the general law of the gravitational field must be satisfied for all gravitational fields obtainable in this way. Even though by no means all gravitational fields

¹ This follows from a generalisation of the discussion in Section XX.

92 GENERAL THEORY OF RELATIVITY

can be produced in this way, yet we may entertain the hope that the general law of gravitation will be derivable from such gravitational fields of a special kind. This hope has been realised in the most beautiful manner. But between the clear vision of this goal and its actual realisation it was necessary to surmount a serious difficulty, and as this lies deep at the root of things, I dare not withhold it from the reader. We require to extend our ideas of the space-time continuum still farther.

XXIII

**BEHAVIOUR OF CLOCKS AND MEASURING-
RODS ON A ROTATING BODY
OF REFERENCE**

HITHERTO I have purposely refrained from speaking about the physical interpretation of space- and time-data in the case of the general theory of relativity. As a consequence, I am guilty of a certain slovenliness of treatment, which, as we know from the special theory of relativity, is far from being unimportant and pardonable. It is now high time that we remedy this defect; but I would mention at the outset, that this matter lays no small claims on the patience and on the power of abstraction of the reader.

We start off again from quite special cases, which we have frequently used before. Let us consider a space-time domain in which no gravitational field exists relative to a reference-body K whose state of motion has been suitably chosen. K is then a Galileian reference-body as regards the domain considered, and the results of the special theory of relativity hold relative to K . Let us suppose the same domain referred to a

94 GENERAL THEORY OF RELATIVITY

second body of reference K' , which is rotating uniformly with respect to K . In order to fix our ideas, we shall imagine K' to be in the form of a plane circular disc, which rotates uniformly in its own plane about its centre. An observer who is sitting eccentrically on the disc K' is sensible of a force which acts outwards in a radial direction, and which would be interpreted as an effect of inertia (centrifugal force) by an observer who was at rest with respect to the original reference-body K . But the observer on the disc may regard his disc as a reference-body which is "at rest"; on the basis of the general principle of relativity he is justified in doing this. The force acting on himself, and in fact on all other bodies which are at rest relative to the disc, he regards as the effect of a gravitational field. Nevertheless, the space-distribution of this gravitational field is of a kind that would not be possible on Newton's theory of gravitation.¹ But since the observer believes in the general theory of relativity, this does not disturb him; he is quite in the right when he believes that a general law of gravitation can be formulated — a law which not only explains the motion of the stars correctly, but also the field of force experienced by himself.

¹ The field disappears at the centre of the disc and increases proportionally to the distance from the centre as we proceed outwards.

The observer performs experiments on his circular disc with clocks and measuring-rods. In doing so, it is his intention to arrive at exact definitions for the signification of time- and space-data with reference to the circular disc K' , these definitions being based on his observations. What will be his experience in this enterprise?

To start with, he places one of two identically constructed clocks at the centre of the circular disc, and the other on the edge of the disc, so that they are at rest relative to it. We now ask ourselves whether both clocks go at the same rate from the standpoint of the non-rotating Galileian reference-body K . As judged from this body, the clock at the centre of the disc has no velocity, whereas the clock at the edge of the disc is in motion relative to K in consequence of the rotation. According to a result obtained in Section XII, it follows that the latter clock goes at a rate permanently slower than that of the clock at the centre of the circular disc, *i.e.* as observed from K . It is obvious that the same effect would be noted by an observer whom we will imagine sitting alongside his clock at the centre of the circular disc. Thus on our circular disc, or, to make the case more general, in every gravitational field, a clock will go more quickly or less quickly, according to the position in which the clock is situated (at rest). For this reason it is not

96 GENERAL THEORY OF RELATIVITY

possible to obtain a reasonable definition of time with the aid of clocks which are arranged at rest with respect to the body of reference. A similar difficulty presents itself when we attempt to apply our earlier definition of simultaneity in such a case, but I do not wish to go any farther into this question.

Moreover, at this stage the definition of the space co-ordinates also presents unsurmountable difficulties. If the observer applies his standard measuring-rod (a rod which is short as compared with the radius of the disc) tangentially to the edge of the disc, then, as judged from the Galileian system, the length of this rod will be less than 1, since, according to Section XII, moving bodies suffer a shortening in the direction of the motion. On the other hand, the measuring-rod will not experience a shortening in length, as judged from K , if it is applied to the disc in the direction of the radius. If, then, the observer first measures the circumference of the disc with his measuring-rod and then the diameter of the disc, on dividing the one by the other, he will not obtain as quotient the familiar number $\pi = 3.14 \dots$, but a larger number,¹ whereas of course, for a disc which is at rest with respect to K , this operation

¹ Throughout this consideration we have to use the Galileian (non-rotating) system K as reference-body, since we may only assume the validity of the results of the special theory of relativity relative to K (relative to K' a gravitational field prevails).

BEHAVIOUR OF CLOCKS AND RODS 97

would yield π exactly. This proves that the propositions of Euclidean geometry cannot hold exactly on the rotating disc, nor in general in a gravitational field, at least if we attribute the length 1 to the rod in all positions and in every orientation. Hence the idea of a straight line also loses its meaning. We are therefore not in a position to define exactly the co-ordinates x, y, z relative to the disc by means of the method used in discussing the special theory, and as long as the co-ordinates and times of events have not been defined we cannot assign an exact meaning to the natural laws in which these occur.

Thus all our previous conclusions based on general relativity would appear to be called in question. In reality we must make a subtle detour in order to be able to apply the postulate of general relativity exactly. I shall prepare the reader for this in the following paragraphs.

XXIV

EUCLIDEAN AND NON-EUCLIDEAN
CONTINUUM

THE surface of a marble table is spread out in front of me. I can get from any one point on this table to any other point by passing continuously from one point to a “neighbouring” one, and repeating this process a (large) number of times, or, in other words, by going from point to point without executing “jumps.” I am sure the reader will appreciate with sufficient clearness what I mean here by “neighbouring” and by “jumps” (if he is not too pedantic). We express this property of the surface by describing the latter as a continuum.

Let us now imagine that a large number of little rods of equal length have been made, their lengths being small compared with the dimensions of the marble slab. When I say they are of equal length, I mean that one can be laid on any other without the ends overlapping. We next lay four of these little rods on the marble slab so that they constitute a quadrilateral figure (a square), the diagonals of which are equally long. To ensure the equality of the diagonals, we make use of a

little testing-rod. To this square we add similar ones, each of which has one rod in common with the first. We proceed in like manner with each of these squares until finally the whole marble slab is laid out with squares. The arrangement is such, that each side of a square belongs to two squares and each corner to four squares.

It is a veritable wonder that we can carry out this business without getting into the greatest difficulties. We only need to think of the following. If at any moment three squares meet at a corner, then two sides of the fourth square are already laid, and as a consequence, the arrangement of the remaining two sides of the square is already completely determined. But I am now no longer able to adjust the quadrilateral so that its diagonals may be equal. If they are equal of their own accord, then this is an especial favour of the marble slab and of the little rods about which I can only be thankfully surprised. We must needs experience many such surprises if the construction is to be successful.

If everything has really gone smoothly, then I say that the points of the marble slab constitute a Euclidean continuum with respect to the little rod, which has been used as a “distance” (line-interval). By choosing one corner of a square as “origin,” I can characterise every other corner of a square with reference to this origin by means

100 GENERAL THEORY OF RELATIVITY

of two numbers. I only need state how many rods I must pass over when, starting from the origin, I proceed towards the "right" and then "upwards," in order to arrive at the corner of the square under consideration. These two numbers are then the "Cartesian co-ordinates" of this corner with reference to the "Cartesian co-ordinate system" which is determined by the arrangement of little rods.

By making use of the following modification of this abstract experiment, we recognise that there must also be cases in which the experiment would be unsuccessful. We shall suppose that the rods "expand" by an amount proportional to the increase of temperature. We heat the central part of the marble slab, but not the periphery, in which case two of our little rods can still be brought into coincidence at every position on the table. But our construction of squares must necessarily come into disorder during the heating, because the little rods on the central region of the table expand, whereas those on the outer part do not.

With reference to our little rods — defined as unit lengths — the marble slab is no longer a Euclidean continuum, and we are also no longer in the position of defining Cartesian co-ordinates directly with their aid, since the above construction can no longer be carried out. But since

there are other things which are not influenced in a similar manner to the little rods (or perhaps not at all) by the temperature of the table, it is possible quite naturally to maintain the point of view that the marble slab is a “Euclidean continuum.” This can be done in a satisfactory manner by making a more subtle stipulation about the measurement or the comparison of lengths.

But if rods of every kind (*i.e.* of every material) were to behave *in the same way* as regards the influence of temperature when they are on the variably heated marble slab, and if we had no other means of detecting the effect of temperature than the geometrical behaviour of our rods in experiments analogous to the one described above, then our best plan would be to assign the distance *one* to two points on the slab, provided that the ends of one of our rods could be made to coincide with these two points; for how else should we define the distance without our proceeding being in the highest measure grossly arbitrary? The method of Cartesian co-ordinates must then be discarded, and replaced by another which does not assume the validity of Euclidean geometry for rigid bodies.¹ The reader will notice that

¹ Mathematicians have been confronted with our problem in the following form. If we are given a surface (*e.g.* an ellipsoid) in Euclidean three-dimensional space, then there exists for this surface a two-dimensional geometry, just as much as for a plane surface.

102 GENERAL THEORY OF RELATIVITY

the situation depicted here corresponds to the one brought about by the general postulate of relativity (Section **XXIII**).

Gauss undertook the task of treating this two-dimensional geometry from first principles, without making use of the fact that the surface belongs to a Euclidean continuum of three dimensions. If we imagine constructions to be made with rigid rods *in the surface* (similar to that above with the marble slab), we should find that different laws hold for these from those resulting on the basis of Euclidean plane geometry. The surface is not a Euclidean continuum with respect to the rods, and we cannot define Cartesian co-ordinates *in the surface*. Gauss indicated the principles according to which we can treat the geometrical relationships in the surface, and thus pointed out the way to the method of Riemann of treating multi-dimensional, non-Euclidean *continua*. Thus it is that mathematicians long ago solved the formal problems to which we are led by the general postulate of relativity.

XXV

GAUSSIAN CO-ORDINATES

ACCORDING to Gauss, this combined analytical and geometrical mode of handling the problem can be arrived at in the following way. We imagine a system of arbitrary curves (see Fig. 4) drawn on the surface of the table. These we designate as u -curves, and we indicate each of them by means of a number. The curves $u = 1$, $u = 2$ and $u = 3$ are drawn in the diagram. Between the curves $u = 1$ and $u = 2$ we must imagine an infinitely large number to be drawn, all of which correspond to real numbers lying between 1 and 2. We have then a system of u -curves, and this “infinitely dense” system covers the whole surface of the table. These u -curves must not intersect each other, and through each point of the surface one and only one curve must pass. Thus a perfectly definite value of u belongs to every point on the surface of the marble slab. In like manner we

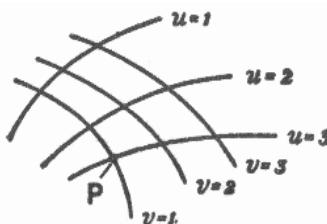


FIG. 4.

104 GENERAL THEORY OF RELATIVITY

imagine a system of v -curves drawn on the surface. These satisfy the same conditions as the u -curves, they are provided with numbers in a corresponding manner, and they may likewise be of arbitrary shape. It follows that a value of u and a value of v belong to every point on the surface of the table. We call these two numbers the co-ordinates of the surface of the table (Gaussian co-ordinates). For example, the point P in the diagram has the Gaussian co-ordinates $u = 3$, $v = 1$. Two neighbouring points P and P' on the surface then correspond to the co-ordinates

$$\begin{array}{ll} P: & u, v \\ P': & u + du, v + dv, \end{array}$$

where du and dv signify very small numbers. In a similar manner we may indicate the distance (line-interval) between P and P' , as measured with a little rod, by means of the very small number ds . Then according to Gauss we have

$$ds^2 = g_{11} du^2 + 2g_{12} du dv + g_{22} dv^2,$$

where g_{11} , g_{12} , g_{22} , are magnitudes which depend in a perfectly definite way on u and v . The magnitudes g_{11} , g_{12} and g_{22} determine the behaviour of the rods relative to the u -curves and v -curves, and thus also relative to the surface of the table. For the case in which the points of the surface considered form a Euclidean continuum with reference to the measuring-rods, but only in this case, it is possible to draw the u -curves and

v -curves and to attach numbers to them, in such a manner, that we simply have:

$$ds^2 = du^2 + dv^2.$$

Under these conditions, the u -curves and v -curves are straight lines in the sense of Euclidean geometry, and they are perpendicular to each other. Here the Gaussian co-ordinates are simply Cartesian ones. It is clear that Gauss co-ordinates are nothing more than an association of two sets of numbers with the points of the surface considered, of such a nature that numerical values differing very slightly from each other are associated with neighbouring points "in space."

So far, these considerations hold for a continuum of two dimensions. But the Gaussian method can be applied also to a continuum of three, four or more dimensions. If, for instance, a continuum of four dimensions be supposed available, we may represent it in the following way. With every point of the continuum we associate arbitrarily four numbers, x_1, x_2, x_3, x_4 , which are known as "co-ordinates." Adjacent points correspond to adjacent values of the co-ordinates. If a distance ds is associated with the adjacent points P and P' , this distance being measurable and well-defined from a physical point of view, then the following formula holds:

$$ds^2 = g_{11}dx_1^2 + 2g_{12}dx_1dx_2 + \dots + g_{44}dx_4^2,$$

106 GENERAL THEORY OF RELATIVITY

where the magnitudes g_{11} , etc., have values which vary with the position in the continuum. Only when the continuum is a Euclidean one is it possible to associate the co-ordinates $x_1 \dots x_4$ with the points of the continuum so that we have simply

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2.$$

In this case relations hold in the four-dimensional continuum which are analogous to those holding in our three-dimensional measurements.

However, the Gauss treatment for ds^2 which we have given above is not always possible. It is only possible when sufficiently small regions of the continuum under consideration may be regarded as Euclidean continua. For example, this obviously holds in the case of the marble slab of the table and local variation of temperature. The temperature is practically constant for a small part of the slab, and thus the geometrical behaviour of the rods is *almost* as it ought to be according to the rules of Euclidean geometry. Hence the imperfections of the construction of squares in the previous section do not show themselves clearly until this construction is extended over a considerable portion of the surface of the table.

We can sum this up as follows: Gauss invented a method for the mathematical treatment of continua in general, in which “size-relations”

(“distances” between neighbouring points) are defined. To every point of a continuum are assigned as many numbers (Gaussian co-ordinates) as the continuum has dimensions. This is done in such a way, that only one meaning can be attached to the assignment, and that numbers (Gaussian co-ordinates) which differ by an indefinitely small amount are assigned to adjacent points. The Gaussian co-ordinate system is a logical generalisation of the Cartesian co-ordinate system. It is also applicable to non-Euclidean continua, but only when, with respect to the defined “size” or “distance,” small parts of the continuum under consideration behave more nearly like a Euclidean system, the smaller the part of the continuum under our notice.

XXVI

THE SPACE-TIME CONTINUUM OF THE SPECIAL THEORY OF RELATIVITY CONSIDERED AS A EUCLIDEAN CONTINUUM

WE are now in a position to formulate more exactly the idea of Minkowski, which was only vaguely indicated in Section XVII. In accordance with the special theory of relativity, certain co-ordinate systems are given preference for the description of the four-dimensional, space-time continuum. We called these "Galileian co-ordinate systems." For these systems, the four co-ordinates x , y , z , t , which determine an event or—in other words—a point of the four-dimensional continuum, are defined physically in a simple manner, as set forth in detail in the first part of this book. For the transition from one Galileian system to another, which is moving uniformly with reference to the first, the equations of the Lorentz transformation are valid. These last form the basis for the derivation of deductions from the special theory of relativity, and in themselves they are nothing more than the expression of the universal

validity of the law of transmission of light for all Galileian systems of reference.

Minkowski found that the Lorentz transformations satisfy the following simple conditions. Let us consider two neighbouring events, the relative position of which in the four-dimensional continuum is given with respect to a Galileian reference-body K by the space co-ordinate differences dx, dy, dz and the time-difference dt . With reference to a second Galileian system we shall suppose that the corresponding differences for these two events are dx', dy', dz', dt' . Then these magnitudes always fulfil the condition.¹

$$dx^2 + dy^2 + dz^2 - c^2 dt^2 = dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2.$$

The validity of the Lorentz transformation follows from this condition. We can express this as follows: The magnitude

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2,$$

which belongs to two adjacent points of the four-dimensional space-time continuum, has the same value for all selected (Galileian) reference-bodies. If we replace $x, y, z, \sqrt{-1} ct$, by x_1, x_2, x_3, x_4 , we also obtain the result that

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2 ^*$$

is independent of the choice of the body of refer-

¹ Cf. Appendices I and II. The relations which are derived there for the co-ordinates themselves are valid also for co-ordinate differences, and thus also for co-ordinate differentials (indefinitely small differences).

[* $ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$ — J.M.]

110 GENERAL THEORY OF RELATIVITY

ence. We call the magnitude ds the “distance” apart of the two events or four-dimensional points.

Thus, if we choose as time-variable the imaginary variable $\sqrt{-1} ct$ instead of the real quantity t , we can regard the space-time continuum — in accordance with the special theory of relativity — as a “Euclidean” four-dimensional continuum, a result which follows from the considerations of the preceding section.

XXVII

**THE SPACE-TIME CONTINUUM OF THE
GENERAL THEORY OF RELATIVITY IS
NOT A EUCLIDEAN CONTINUUM**

IN the first part of this book we were able to make use of space-time co-ordinates which allowed of a simple and direct physical interpretation, and which, according to Section [XXVI](#), can be regarded as four-dimensional Cartesian co-ordinates. This was possible on the basis of the law of the constancy of the velocity of light. But according to Section [XXII](#),* the general theory of relativity cannot retain this law. On the contrary, we arrived at the result that according to this latter theory the velocity of light must always depend on the co-ordinates when a gravitational field is present. In connection with a specific illustration in Section [XXIII](#), we found that the presence of a gravitational field invalidates the definition of the co-ordinates and the time, which led us to our objective in the special theory of relativity.

In view of the results of these considerations we are led to the conviction that, according to

112 GENERAL THEORY OF RELATIVITY

the general principle of relativity, the space-time continuum cannot be regarded as a Euclidean one, but that here we have the general case, corresponding to the marble slab with local variations of temperature, and with which we made acquaintance as an example of a two-dimensional continuum. Just as it was there impossible to construct a Cartesian co-ordinate system from equal rods, so here it is impossible to build up a system (reference-body) from rigid bodies and clocks, which shall be of such a nature that measuring-rods and clocks, arranged rigidly with respect to one another, shall indicate position and time directly. Such was the essence of the difficulty with which we were confronted in Section **XXIII**.

But the considerations of Sections **XXV** and **XXVI** show us the way to surmount this difficulty. We refer the four-dimensional space-time continuum in an arbitrary manner to Gauss co-ordinates. We assign to every point of the continuum (event) four numbers, x_1, x_2, x_3, x_4 (co-ordinates), which have not the least direct physical significance, but only serve the purpose of numbering the points of the continuum in a definite but arbitrary manner. This arrangement does not even need to be of such a kind that we must regard x_1, x_2, x_3 , as “space” co-ordinates and x_4 as a “time” co-ordinate.

The reader may think that such a description of the world would be quite inadequate. What does it mean to assign to an event the particular co-ordinates x_1, x_2, x_3, x_4 , if in themselves these co-ordinates have no significance? More careful consideration shows, however, that this anxiety is unfounded. Let us consider, for instance, a material point with any kind of motion. If this point had only a momentary existence without duration, then it would be described in space-time by a single system of values x_1, x_2, x_3, x_4 . Thus its permanent existence must be characterised by an infinitely large number of such systems of values, the co-ordinate values of which are so close together as to give continuity; corresponding to the material point, we thus have a (uni-dimensional) line in the four-dimensional continuum. In the same way, any such lines in our continuum correspond to many points in motion. The only statements having regard to these points which can claim a physical existence are in reality the statements about their encounters. In our mathematical treatment, such an encounter is expressed in the fact that the two lines which represent the motions of the points in question have a particular system of co-ordinate values, x_1, x_2, x_3, x_4 , in common. After mature consideration the reader will doubtless admit that in reality such encounters con-

114 GENERAL THEORY OF RELATIVITY

stitute the only actual evidence of a time-space nature with which we meet in physical statements.

When we were describing the motion of a material point relative to a body of reference, we stated nothing more than the encounters of this point with particular points of the reference-body. We can also determine the corresponding values of the time by the observation of encounters of the body with clocks, in conjunction with the observation of the encounter of the hands of clocks with particular points on the dials. It is just the same in the case of space-measurements by means of measuring-rods, as a little consideration will show.

The following statements hold generally: Every physical description resolves itself into a number of statements, each of which refers to the space-time coincidence of two events *A* and *B*. In terms of Gaussian co-ordinates, every such statement is expressed by the agreement of their four co-ordinates x_1, x_2, x_3, x_4 . Thus in reality, the description of the time-space continuum by means of Gauss co-ordinates completely replaces the description with the aid of a body of reference, without suffering from the defects of the latter mode of description; it is not tied down to the Euclidean character of the continuum which has to be represented.

XXVIII

EXACT FORMULATION OF THE GENERAL PRINCIPLE OF RELATIVITY

WE are now in a position to replace the provisional formulation of the general principle of relativity given in Section XVIII by an exact formulation. The form there used, "All bodies of reference K , K' , etc., are equivalent for the description of natural phenomena (formulation of the general laws of nature), whatever may be their state of motion," cannot be maintained, because the use of rigid reference-bodies, in the sense of the method followed in the special theory of relativity, is in general not possible in space-time description. The Gauss co-ordinate system has to take the place of the body of reference. The following statement corresponds to the fundamental idea of the general principle of relativity: "*All Gaussian co-ordinate systems are essentially equivalent for the formulation of the general laws of nature.*"

We can state this general principle of relativity in still another form, which renders it yet more clearly intelligible than it is when in the form of

116 GENERAL THEORY OF RELATIVITY

the natural extension of the special principle of relativity. According to the special theory of relativity, the equations which express the general laws of nature pass over into equations of the same form when, by making use of the Lorentz transformation, we replace the space-time variables x, y, z, t , of a (Galileian) reference-body K by the space-time variables x', y', z', t' , of a new reference-body K' . According to the general theory of relativity, on the other hand, by application of *arbitrary substitutions* of the Gauss variables x_1, x_2, x_3, x_4 , the equations must pass over into equations of the same form; for every transformation (not only the Lorentz transformation) corresponds to the transition of one Gauss co-ordinate system into another.

If we desire to adhere to our “old-time” three-dimensional view of things, then we can characterise the development which is being undergone by the fundamental idea of the general theory of relativity as follows: The special theory of relativity has reference to Galileian domains, *i.e.* to those in which no gravitational field exists. In this connection a Galileian reference-body serves as body of reference, *i.e.* a rigid body the state of motion of which is so chosen that the Galileian law of the uniform rectilinear motion of “isolated” material points holds relatively to it.

GENERAL PRINCIPLE OF RELATIVITY 117

Certain considerations suggest that we should refer the same Galileian domains to *non-Galileian* reference-bodies also. A gravitational field of a special kind is then present with respect to these bodies (cf. Sections XX and XXIII).

In gravitational fields there are no such things as rigid bodies with Euclidean properties; thus the fictitious rigid body of reference is of no avail in the general theory of relativity. The motion of clocks is also influenced by gravitational fields, and in such a way that a physical definition of time which is made directly with the aid of clocks has by no means the same degree of plausibility as in the special theory of relativity.

For this reason non-rigid reference-bodies are used which are as a whole not only moving in any way whatsoever, but which also suffer alterations in form *ad lib.* during their motion. Clocks, for which the law of motion is of any kind, however irregular, serve for the definition of time. We have to imagine each of these clocks fixed at a point on the non-rigid reference-body. These clocks satisfy only the one condition, that the “readings” which are observed simultaneously on adjacent clocks (in space) differ from each other by an indefinitely small amount. This non-rigid reference-body, which might appropriately be termed a “reference-mollusk,” is in the main equivalent to a Gaussian four-dimensional co-ordinate sys-

118 GENERAL THEORY OF RELATIVITY

tem chosen arbitrarily. That which gives the "mollusk" a certain comprehensibleness as compared with the Gauss co-ordinate system is the (really unqualified^{*}) formal retention of the separate existence of the space co-ordinates as opposed to the time co-ordinate. Every point on the mollusk is treated as a space-point, and every material point which is at rest relatively to it as at rest, so long as the mollusk is considered as reference-body. The general principle of relativity requires that all these mollusks can be used as reference-bodies with equal right and equal success in the formulation of the general laws of nature; the laws themselves must be quite independent of the choice of mollusk.

The great power possessed by the general principle of relativity lies in the comprehensive limitation which is imposed on the laws of nature in consequence of what we have seen above.

[^{*} The word "unqualified" was correctly changed to "unjustified" in later editions. — J.M.]

XXIX

THE SOLUTION OF THE PROBLEM OF GRAVITATION ON THE BASIS OF THE GENERAL PRINCIPLE OF RELATIVITY

If the reader has followed all our previous considerations, he will have no further difficulty in understanding the methods leading to the solution of the problem of gravitation.

We start off from a consideration of a Galileian domain, *i.e.* a domain in which there is no gravitational field relative to the Galileian reference-body K . The behaviour of measuring-rods and clocks with reference to K is known from the special theory of relativity, likewise the behaviour of “isolated” material points; the latter move uniformly and in straight lines.

Now let us refer this domain to a random Gauss co-ordinate system or to a “mollusk” as reference-body K' . Then with respect to K' there is a gravitational field G (of a particular kind). We learn the behaviour of measuring-rods and clocks and also of freely-moving material points with reference to K' simply by mathematical transformation. We interpret this behaviour as the

120 GENERAL THEORY OF RELATIVITY

behaviour of measuring-rods, clocks and material points under the influence of the gravitational field G . Hereupon we introduce a hypothesis: that the influence of the gravitational field on measuring-rods, clocks and freely-moving material points continues to take place according to the same laws, even in the case when the prevailing gravitational field is *not* derivable from the Galileian special case, simply by means of a transformation of co-ordinates.

The next step is to investigate the space-time behaviour of the gravitational field G , which was derived from the Galileian special case simply by transformation of the co-ordinates. This behaviour is formulated in a law, which is always valid, no matter how the reference-body (mollusk) used in the description may be chosen.

This law is not yet the *general* law of the gravitational field, since the gravitational field under consideration is of a special kind. In order to find out the general law-of-field of gravitation we still require to obtain a generalisation of the law as found above. This can be obtained without caprice, however, by taking into consideration the following demands:

- (a) The required generalisation must likewise satisfy the general postulate of relativity.
- (b) If there is any matter in the domain under consideration, only its inertial mass, and

SOLUTION OF GRAVITATION 121

thus according to Section **XV** only its energy is of importance for its effect in exciting a field.

(c) Gravitational field and matter together must satisfy the law of the conservation of energy (and of impulse).

Finally, the general principle of relativity permits us to determine the influence of the gravitational field on the course of all those processes which take place according to known laws when a gravitational field is absent, *i.e.* which have already been fitted into the frame of the special theory of relativity. In this connection we proceed in principle according to the method which has already been explained for measuring-rods, clocks and freely-moving material points.

The theory of gravitation derived in this way from the general postulate of relativity excels not only in its beauty; nor in removing the defect attaching to classical mechanics which was brought to light in Section **XXI**; nor in interpreting the empirical law of the equality of inertial and gravitational mass; but it has also already explained a result of observation in astronomy, against which classical mechanics is powerless.

If we confine the application of the theory to the case where the gravitational fields can be regarded as being weak, and in which all masses move with respect to the co-ordinate system with

122 GENERAL THEORY OF RELATIVITY

velocities which are small compared with the velocity of light, we then obtain as a first approximation the Newtonian theory. Thus the latter theory is obtained here without any particular assumption, whereas Newton had to introduce the hypothesis that the force of attraction between mutually attracting material points is inversely proportional to the square of the distance between them. If we increase the accuracy of the calculation, deviations from the theory of Newton make their appearance, practically all of which must nevertheless escape the test of observation owing to their smallness.

We must draw attention here to one of these deviations. According to Newton's theory, a planet moves round the sun in an ellipse, which would permanently maintain its position with respect to the fixed stars, if we could disregard the motion of the fixed stars themselves and the action of the other planets under consideration. Thus, if we correct the observed motion of the planets for these two influences, and if Newton's theory be strictly correct, we ought to obtain for the orbit of the planet an ellipse, which is fixed with reference to the fixed stars. This deduction, which can be tested with great accuracy, has been confirmed for all the planets save one, with the precision that is capable of being obtained by the delicacy of observation

attainable at the present time. The sole exception is Mercury, the planet which lies nearest the sun. Since the time of Leverrier, it has been known that the ellipse corresponding to the orbit of Mercury, after it has been corrected for the influences mentioned above, is not stationary with respect to the fixed stars, but that it rotates exceedingly slowly in the plane of the orbit and in the sense of the orbital motion. The value obtained for this rotary movement of the orbital ellipse was 43 seconds of arc per century, an amount ensured to be correct to within a few seconds of arc. This effect can be explained by means of classical mechanics only on the assumption of hypotheses which have little probability, and which were devised solely for this purpose.

On the basis of the general theory of relativity, it is found that the ellipse of every planet round the sun must necessarily rotate in the manner indicated above; that for all the planets, with the exception of Mercury, this rotation is too small to be detected with the delicacy of observation possible at the present time; but that in the case of Mercury it must amount to 43 seconds of arc per century, a result which is strictly in agreement with observation.

Apart from this one, it has hitherto been possible to make only two deductions from the theory

124 GENERAL THEORY OF RELATIVITY

which admit of being tested by observation, to wit, the curvature of light rays by the gravitational field of the sun,¹ and a displacement of the spectral lines of light reaching us from large stars, as compared with the corresponding lines for light produced in an analogous manner terrestrially (*i.e.* by the same kind of molecule^{*}). I do not doubt that these deductions from the theory will be confirmed also.

¹ Observed by Eddington and others in 1919. (Cf. Appendix III.)

[^{*} The word “molecule” was correctly changed to “atom” in later editions. Cf. Appendix III, pg. 157. — J.M.]

PART III

CONSIDERATIONS ON THE UNIVERSE AS A WHOLE

XXX

COSMOLOGICAL DIFFICULTIES OF NEWTON'S THEORY

APART from the difficulty discussed in Section [XXI](#), there is a second fundamental difficulty attending classical celestial mechanics, which, to the best of my knowledge, was first discussed in detail by the astronomer Seeliger. If we ponder over the question as to how the universe, considered as a whole, is to be regarded, the first answer that suggests itself to us is surely this: As regards space (and time) the universe is infinite. There are stars everywhere, so that the density of matter, although very variable in detail, is nevertheless on the average everywhere the same. In other words: However far we might travel through space, we should find everywhere an attenuated swarm of fixed stars of approximately the same kind and density.

126 CONSIDERATIONS ON THE UNIVERSE

This view is not in harmony with the theory of Newton. The latter theory rather requires that the universe should have a kind of centre in which the density of the stars is a maximum, and that as we proceed outwards from this centre the group-density of the stars should diminish, until finally, at great distances, it is succeeded by an infinite region of emptiness. The stellar universe ought to be a finite island in the infinite ocean of space.¹

This conception is in itself not very satisfactory. It is still less satisfactory because it leads to the result that the light emitted by the stars and also individual stars of the stellar system are perpetually passing out into infinite space, never to return, and without ever again coming into interaction with other objects of nature. Such a finite material universe would be destined to become gradually but systematically impoverished.

¹ *Proof.* — According to the theory of Newton, the number of “lines of force” which come from infinity and terminate in a mass m is proportional to the mass m . If, on the average, the mass-density ρ_0 is constant throughout the universe, then a sphere of volume V will enclose the average mass ρ_0V . Thus the number of lines of force passing through the surface F of the sphere into its interior is proportional to ρ_0V . For unit area of the surface of the sphere the number of lines of force which enters the sphere is thus proportional to $\rho_0 \frac{V}{F}$ * or ρ_0R . Hence the intensity of the field at the surface would ultimately become infinite with increasing radius R of the sphere, which is impossible.

[* $\rho_0 \frac{V}{F}$ — J.M.]

In order to escape this dilemma, Seeliger suggested a modification of Newton's law, in which he assumes that for great distances the force of attraction between two masses diminishes more rapidly than would result from the inverse square law. In this way it is possible for the mean density of matter to be constant everywhere, even to infinity, without infinitely large gravitational fields being produced. We thus free ourselves from the distasteful conception that the material universe ought to possess something of the nature of a centre. Of course we purchase our emancipation from the fundamental difficulties mentioned, at the cost of a modification and complication of Newton's law which has neither empirical nor theoretical foundation. We can imagine innumerable laws which would serve the same purpose, without our being able to state a reason why one of them is to be preferred to the others; for any one of these laws would be founded just as little on more general theoretical principles as is the law of Newton.

XXXI

**THE POSSIBILITY OF A “FINITE” AND YET
“UNBOUNDED” UNIVERSE**

BUT speculations on the structure of the universe also move in quite another direction. The development of non-Euclidean geometry led to the recognition of the fact, that we can cast doubt on the *infiniteness* of our space without coming into conflict with the laws of thought or with experience (Riemann, Helmholtz). These questions have already been treated in detail and with unsurpassable lucidity by Helmholtz and Poincaré, whereas I can only touch on them briefly here.

In the first place, we imagine an existence in two-dimensional space. Flat beings with flat implements, and in particular flat rigid measuring-rods, are free to move in a *plane*. For them nothing exists outside of this plane: that which they observe to happen to themselves and to their flat “things” is the all-inclusive reality of their plane. In particular, the constructions of plane Euclidean geometry can be carried out by means of the rods, *e.g.* the lattice construction, con-

sidered in Section XXIV. In contrast to ours, the universe of these beings is two-dimensional; but, like ours, it extends to infinity. In their universe there is room for an infinite number of identical squares made up of rods, *i.e.* its volume (surface) is infinite. If these beings say their universe is “plane,” there is sense in the statement, because they mean that they can perform the constructions of plane Euclidean geometry with their rods. In this connection the individual rods always represent the same distance, independently of their position.

Let us consider now a second two-dimensional existence, but this time on a spherical surface instead of on a plane. The flat beings with their measuring-rods and other objects fit exactly on this surface and they are unable to leave it. Their whole universe of observation extends exclusively over the surface of the sphere. Are these beings able to regard the geometry of their universe as being plane geometry and their rods withal as the realisation of “distance”? They cannot do this. For if they attempt to realise a straight line, they will obtain a curve, which we “three-dimensional beings” designate as a great circle, *i.e.* a self-contained line of definite finite length, which can be measured up by means of a measuring-rod. Similarly, this universe has a finite area, that can be compared with the area of a

130 CONSIDERATIONS ON THE UNIVERSE

square constructed with rods. The great charm resulting from this consideration lies in the recognition of the fact that *the universe of these beings is finite and yet has no limits.*

But the spherical-surface beings do not need to go on a world-tour in order to perceive that they are not living in a Euclidean universe. They can convince themselves of this on every part of their “world,” provided they do not use too small a piece of it. Starting from a point, they draw “straight lines” (arcs of circles as judged in three-dimensional space) of equal length in all directions. They will call the line joining the free ends of these lines a “circle.” For a plane surface, the ratio of the circumference of a circle to its diameter, both lengths being measured with the same rod, is, according to Euclidean geometry of the plane, equal to a constant value π , which is independent of the diameter of the circle. On their spherical surface our flat beings would find for this ratio the value

$$\pi \frac{\sin\left(\frac{r}{R}\right)}{\left(\frac{r}{R}\right)},$$

i.e. a smaller value than π , the difference being the more considerable, the greater is the radius of the circle in comparison with the radius R of the “world-sphere.” By means of this relation

the spherical beings can determine the radius of their universe (“world”), even when only a relatively small part of their world-sphere is available for their measurements. But if this part is very small indeed, they will no longer be able to demonstrate that they are on a spherical “world” and not on a Euclidean plane, for a small part of a spherical surface differs only slightly from a piece of a plane of the same size.

Thus if the spherical-surface beings are living on a planet of which the solar system occupies only a negligibly small part of the spherical universe, they have no means of determining whether they are living in a finite or in an infinite universe, because the “piece of universe” to which they have access is in both cases practically plane, or Euclidean. It follows directly from this discussion, that for our sphere-beings the circumference of a circle first increases with the radius until the “circumference of the universe” is reached, and that it thenceforward gradually decreases to zero for still further increasing values of the radius. During this process the area of the circle continues to increase more and more, until finally it becomes equal to the total area of the whole “world-sphere.”

Perhaps the reader will wonder why we have placed our “beings” on a sphere rather than on another closed surface. But this choice has its

132 CONSIDERATIONS ON THE UNIVERSE

justification in the fact that, of all closed surfaces, the sphere is unique in possessing the property that all points on it are equivalent. I admit that the ratio of the circumference c of a circle to its radius r depends on r , but for a given value of r it is the same for all points of the “world-sphere”; in other words, the “world-sphere” is a “surface of constant curvature.”

To this two-dimensional sphere-universe there is a three-dimensional analogy, namely, the three-dimensional spherical space which was discovered by Riemann. Its points are likewise all equivalent. It possesses a finite volume, which is determined by its “radius” ($2\pi^2R^3$). Is it possible to imagine a spherical space? To imagine a space means nothing else than that we imagine an epitome of our “space” experience, *i.e.* of experience that we can have in the movement of “rigid” bodies. In this sense we *can* imagine a spherical space.

Suppose we draw lines or stretch strings in all directions from a point, and mark off from each of these the distance r with a measuring-rod. All the free end-points of these lengths lie on a spherical surface. We can specially measure up the area (F) of this surface by means of a square made up of measuring-rods. If the universe is Euclidean, then $F = 4\pi r^2$; if it is spherical, then F is always less than $4\pi r^2$. With increasing values

of r , F increases from zero up to a maximum value which is determined by the “world-radius,” but for still further increasing values of r , the area gradually diminishes to zero. At first, the straight lines which radiate from the starting point diverge farther and farther from one another, but later they approach each other, and finally they run together again at a “counter-point” to the starting point. Under such conditions they have traversed the whole spherical space. It is easily seen that the three-dimensional spherical space is quite analogous to the two-dimensional spherical surface. It is finite (*i.e.* of finite volume), and has no bounds.

It may be mentioned that there is yet another kind of curved space: “elliptical space.” It can be regarded as a curved space in which the two “counter-points” are identical (indistinguishable from each other). An elliptical universe can thus be considered to some extent as a curved universe possessing central symmetry.

It follows from what has been said, that closed spaces without limits are conceivable. From amongst these, the spherical space (and the elliptical) excels in its simplicity, since all points on it are equivalent. As a result of this discussion, a most interesting question arises for astronomers and physicists, and that is whether the universe in which we live is infinite, or whether it is finite

134 CONSIDERATIONS ON THE UNIVERSE

in the manner of the spherical universe. Our experience is far from being sufficient to enable us to answer this question. But the general theory of relativity permits of our answering it with a moderate degree of certainty, and in this connection the difficulty mentioned in Section **XXX** finds its solution.

XXXII

**THE STRUCTURE OF SPACE ACCORDING TO
THE GENERAL THEORY OF RELATIVITY**

ACCORDING to the general theory of relativity, the geometrical properties of space are not independent, but they are determined by matter. Thus we can draw conclusions about the geometrical structure of the universe only if we base our considerations on the state of the matter as being something that is known. We know from experience that, for a suitably chosen co-ordinate system, the velocities of the stars are small as compared with the velocity of transmission of light. We can thus as a rough approximation arrive at a conclusion as to the nature of the universe as a whole, if we treat the matter as being at rest.

We already know from our previous discussion that the behaviour of measuring-rods and clocks is influenced by gravitational fields, *i.e.* by the distribution of matter. This in itself is sufficient to exclude the possibility of the exact validity of Euclidean geometry in our universe. But it is conceivable that our universe differs only slightly

136 CONSIDERATIONS ON THE UNIVERSE

from a Euclidean one, and this notion seems all the more probable, since calculations show that the metrics of surrounding space is influenced only to an exceedingly small extent by masses even of the magnitude of our sun. We might imagine that, as regards geometry, our universe behaves analogously to a surface which is irregularly curved in its individual parts, but which nowhere departs appreciably from a plane: something like the rippled surface of a lake. Such a universe might fittingly be called a quasi-Euclidean universe. As regards its space it would be infinite. But calculation shows that in a quasi-Euclidean universe the average density of matter would necessarily be *nil*. Thus such a universe could not be inhabited by matter everywhere; it would present to us that unsatisfactory picture which we portrayed in Section **XXX**.

If we are to have in the universe an average density of matter which differs from zero, however small may be that difference, then the universe cannot be quasi-Euclidean. On the contrary, the results of calculation indicate that if matter be distributed uniformly, the universe would necessarily be spherical (or elliptical). Since in reality the detailed distribution of matter is not uniform, the real universe will deviate in individual parts from the spherical, *i.e.* the universe will be quasi-spherical. But it will be

THE STRUCTURE OF SPACE 137

necessarily finite. In fact, the theory supplies us with a simple connection¹ between the space-expansive of the universe and the average density of matter in it.

¹ For the "radius" R of the universe we obtain the equation

$$R^2 = \frac{2}{\kappa\rho}$$

The use of the C.G.S. system in this equation gives $\frac{2}{\kappa} = 1.08 \cdot 10^{27}$; ρ is the average density of the matter.

749

FÉUE WHI

APPENDIX I

SIMPLE DERIVATION OF THE LORENTZ TRANSFORMATION [SUPPLEMENTARY TO SEC- TION XI]

FOR the relative orientation of the co-ordinate systems indicated in Fig. 2, the x -axes of both systems permanently coincide. In the present case we can divide the problem into parts by considering first only events which are localised on the x -axis. Any such event is represented with respect to the coordinate system K by the abscissa x and the time t , and with respect to the system K' by the abscissa x' and the time t' . We require to find x' and t' when x and t are given.

A light-signal, which is proceeding along the positive axis of x , is transmitted according to the equation

$$x = ct$$

or

$$x - ct = 0 \dots \dots \dots \dots \dots \quad (1).$$

Since the same light-signal has to be transmitted relative to K' with the velocity c , the propagation

APPENDIX I

relative to the system K' will be represented by the analogous formula

$$x' - ct' = 0 \quad \dots \dots \dots \quad (2).$$

Those space-time points (events) which satisfy (1) must also satisfy (2). Obviously this will be the case when the relation

$$(x' - ct') = \lambda(x - ct) \quad \dots \dots \dots \quad (3)$$

is fulfilled in general, where λ indicates a constant; for, according to (3), the disappearance of $(x - ct)$ involves the disappearance of $(x' - ct')$.

If we apply quite similar considerations to light rays which are being transmitted along the negative x -axis, we obtain the condition

$$(x' + ct') = \mu(x + ct) \quad \dots \dots \dots \quad (4).$$

By adding (or subtracting) equations (3) and (4), and introducing for convenience the constants a and b in place of the constants λ and μ where

$$a = \frac{\lambda + \mu}{2}$$

and

$$b = \frac{\lambda - \mu}{2},$$

we obtain the equations

$$\left. \begin{aligned} x' &= ax - bct \\ ct' &= act - bx \end{aligned} \right\} \quad \dots \dots \dots \quad (5).$$

We should thus have the solution of our problem, if the constants a and b were known. These result from the following discussion.

For the origin of K' we have permanently $x' = 0$, and hence according to the first of the equations (5)

$$x = \frac{bc}{a}t.$$

If we call v the velocity with which the origin of K' is moving relative to K , we then have

$$v = \frac{bc}{a} \quad \dots \dots \dots \dots \quad (6).$$

The same value v can be obtained from equations * (5), if we calculate the velocity of another point of K' relative to K , or the velocity (directed towards the negative x -axis) of a point of K with respect to K' . In short, we can designate v as the relative velocity of the two systems.

Furthermore, the principle of relativity teaches us that, as judged from K , the length of a unit measuring-rod which is at rest with reference to K' must be exactly the same as the length, as judged from K' , of a unit measuring-rod which is at rest relative to K . In order to see how the points of the x' -axis appear as viewed from K , we only require to take a "snapshot" of K' from K ; this means that we have to insert a particular value of t (time of K), e.g. [†] $t = 0$. For this value of t we then obtain from the first of the equations (5)

$$x' = ax.$$

Two points of the x' -axis which are separated by the distance $\Delta x' = 1^{\ddagger}$ when measured in the K' system are thus separated in our instantaneous photograph by the distance

$$\Delta x = \frac{1}{a} \quad \dots \dots \dots \dots \quad (7).$$

[* tion — J.M.]

[† e.g. — J.M.]

[‡ $x' = 1$ — J.M.]

APPENDIX I

But if the snapshot be taken from K' ($t' = 0$), and if we eliminate t from the equations (5), taking into account the expression (6), we obtain

$$x' = a \left(1 - \frac{v^2}{c^2} \right) x.$$

From this we conclude that two points on the x -axis and separated by the distance 1 (relative to K) will be represented on our snapshot by the distance

$$\Delta x' = a \left(1 - \frac{v^2}{c^2} \right) \dots \quad (7a).$$

But from what has been said, the two snapshots must be identical; hence Δx in (7) must be equal to $\Delta x'$ in (7a), so that we obtain

$$a^2 = \frac{1}{1 - \frac{v^2}{c^2}} \quad \dots \quad (7b).$$

The equations (6) and (7b) determine the constants a and b . By inserting the values of these constants in (5), we obtain the first and the fourth of the equations given in Section XI.

$$\left. \begin{aligned} x' &= \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \\ t' &= \frac{t - \frac{v}{c^2} x}{\sqrt{1 - \frac{v^2}{c^2}}} \end{aligned} \right\} \quad \dots \quad (8).$$

THE LORENTZ TRANSFORMATION 143

Thus we have obtained the Lorentz transformation for events on the x -axis. It satisfies the condition

$$x'^2 - c^2 t'^2 = x^2 - c^2 t^2 \quad \dots \quad (8a).$$

The extension of this result, to include events which take place outside the x -axis, is obtained by retaining equations (8) and supplementing them by the relations

$$\left. \begin{array}{l} y' = y \\ z' = z \end{array} \right\} \quad \dots \quad (9).$$

In this way we satisfy the postulate of the constancy of the velocity of light *in vacuo* for rays of light of arbitrary direction, both for the system K and for the system K' . This may be shown in the following manner.

We suppose a light-signal sent out from the origin of K at the time $t = 0$. It will be propagated according to the equation

$$r = \sqrt{x^2 + y^2 + z^2} = ct,$$

or, if we square this equation, according to the equation

$$x^2 + y^2 + z^2 - c^2 t^2 = 0 \quad \dots \quad (10).$$

It is required by the law of propagation of light, in conjunction with the postulate of relativity, that the transmission of the signal in question should take place—as judged from K' —in accordance with the corresponding formula

$$r' = ct'$$

or,

$$x'^2 + y'^2 + z'^2 - c^2 t'^2 = 0 \quad \dots \quad (10a).$$

APPENDIX I

In order that equation (10a) may be a consequence of equation (10), we must have

$$x'^2 + y'^2 + z'^2 - c^2 t'^2 = \sigma(x^2 + y^2 + z^2 - c^2 t^2) \quad (11).$$

Since equation (8a) must hold for points on the x -axis, we thus have $\sigma = 1$. It is easily seen that the Lorentz transformation really satisfies equation (11) for $\sigma = 1$; for (11) is a consequence of (8a) and (9), and hence also of (8) and (9). We have thus derived the Lorentz transformation.

The Lorentz transformation represented by (8) and (9) still requires to be generalised. Obviously it is immaterial whether the axes of K' be chosen so that they are spatially parallel to those of K . It is also not essential that the velocity of translation of K' with respect to K should be in the direction of the x -axis. A simple consideration shows that we are able to construct the Lorentz transformation in this general sense from two kinds of transformations, viz. from Lorentz transformations in the special sense and from purely spatial transformations, which corresponds to the replacement of the rectangular co-ordinate system by a new system with its axes pointing in other directions.

Mathematically, we can characterise the generalised Lorentz transformation thus:

It expresses x' , y' , z' , t' , in terms of linear homogeneous functions of x , y , z , t , of such a kind that the relation

THE LORENTZ TRANSFORMATION 145

$$x'^2 + y'^2 + z'^2 - c^2 t'^2 = x^2 + y^2 + z^2 - c^2 t^2 . \quad (11a)$$

is satisfied identically. That is to say: If we substitute their expressions in x, y, z, t , in place of x', y', z', t' , on the left-hand side, then the left-hand side of (11a) agrees with the right-hand side.

APPENDIX II

MINKOWSKI'S FOUR—DIMENSIONAL SPACE ("WORLD") [SUPPLEMENTARY TO SECTION XVII]

WE can characterise the Lorentz transformation still more simply if we introduce the imaginary $\sqrt{-1} \cdot ct$ in place of t , as time-variable. If, in accordance with this, we insert

$$\begin{aligned}x_1 &= x \\x_2 &= y \\x_3 &= z \\x_4 &= \sqrt{-1} \cdot ct,\end{aligned}$$

and similarly for the accented system K' , then the condition which is identically satisfied by the transformation can be expressed thus:

$$x_1'^2 + x_2'^2 + x_3'^2 + x_4'^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2. \quad (12).$$

That is, by the afore-mentioned choice of "co-ordinates" (11a) is transformed into this equation.

We see from (12) that the imaginary time co-ordinate x_4 enters into the condition of transformation in exactly the same way as the space co-ordinates x_1, x_2, x_3 . It is due to this fact that, according to the theory of relativity, the "time"

x_4 enters into natural laws in the same form as the space co-ordinates x_1, x_2, x_3 .

A four-dimensional continuum described by the “co-ordinates” x_1, x_2, x_3, x_4 , was called “world” by Minkowski, who also termed a point-event a “world-point.” From a “happening” in three-dimensional space, physics becomes, as it were, an “existence” in the four-dimensional “world.”

This four-dimensional “world” bears a close similarity to the three-dimensional “space” of (Euclidean) analytical geometry. If we introduce into the latter a new Cartesian co-ordinate system (x'_1, x'_2, x'_3) with the same origin, then x'_1, x'_2, x'_3 , are linear homogeneous functions of x_1, x_2, x_3 , which identically satisfy the equation

$$x'_1{}^2 + x'_2{}^2 + x'_3{}^2 = x_1{}^2 + x_2{}^2 + x_3{}^2.$$

The analogy with (12) is a complete one. We can regard Minkowski’s “world” in a formal manner as a four-dimensional Euclidean space (with imaginary time co-ordinate); the Lorentz transformation corresponds to a “rotation” of the co-ordinate system in the four-dimensional “world.”

APPENDIX III

THE EXPERIMENTAL CONFIRMATION OF THE GENERAL THEORY OF RELATIVITY

FROM a systematic theoretical point of view, we may imagine the process of evolution of an empirical science to be a continuous process of induction. Theories are evolved, and are expressed in short compass as statements of a large number of individual observations in the form of empirical laws, from which the general laws can be ascertained by comparison. Regarded in this way, the development of a science bears some resemblance to the compilation of a classified catalogue. It is, as it were, a purely empirical enterprise.

But this point of view by no means embraces the whole of the actual process; for it slurs over the important part played by intuition and deductive thought in the development of an exact science. As soon as a science has emerged from its initial stages, theoretical advances are no longer achieved merely by a process of arrangement. Guided by empirical data, the investigator rather develops a system of thought which, in

general, is built up logically from a small number of fundamental assumptions, the so-called axioms. We call such a system of thought a *theory*. The theory finds the justification for its existence in the fact that it correlates a large number of single observations, and it is just here that the "truth" of the theory lies.

Corresponding to the same complex of empirical data, there may be several theories, which differ from one another to a considerable extent. But as regards the deductions from the theories which are capable of being tested, the agreement between the theories may be so complete, that it becomes difficult to find such deductions in which the two theories differ from each other. As an example, a case of general interest is available in the province of biology, in the Darwinian theory of the development of species by selection in the struggle for existence, and in the theory of development which is based on the hypothesis of the hereditary transmission of acquired characters.

We have another instance of far-reaching agreement between the deductions from two theories in Newtonian mechanics on the one hand, and the general theory of relativity on the other. This agreement goes so far, that up to the present we have been able to find only a few deductions from the general theory of relativity which are

APPENDIX III

capable of investigation, and to which the physics of pre-relativity days does not also lead, and this despite the profound difference in the fundamental assumptions of the two theories. In what follows, we shall again consider these important deductions, and we shall also discuss the empirical evidence appertaining to them which has hitherto been obtained.

(a) MOTION OF THE PERIHELION OF MERCURY

According to Newtonian mechanics and Newton's law of gravitation, a planet which is revolving round the sun would describe an ellipse round the latter, or, more correctly, round the common centre of gravity of the sun and the planet. In such a system, the sun, or the common centre of gravity, lies in one of the foci of the orbital ellipse in such a manner that, in the course of a planet-year, the distance sun-planet grows from a minimum to a maximum, and then decreases again to a minimum. If instead of Newton's law we insert a somewhat different law of attraction into the calculation, we find that, according to this new law, the motion would still take place in such a manner that the distance sun-planet exhibits periodic variations; but in this case the angle described by the line joining sun and planet during such a period (from perihelion — closest

EXPERIMENTAL CONFIRMATION 151

proximity to the sun — to perihelion) would differ from 360° . The line of the orbit would not then be a closed one, but in the course of time it would fill up an annular part of the orbital plane, viz. between the circle of least and the circle of greatest distance of the planet from the sun.

According also to the general theory of relativity, which differs of course from the theory of Newton, a small variation from the Newton-Kepler motion of a planet in its orbit should take place, and in such a way, that the angle described by the radius sun-planet between one perihelion and the next should exceed that corresponding to one complete revolution by an amount given by

$$+ \frac{24\pi^3 a^2}{T^2 c^2 (1 - e^2)}.$$

(N.B. — One complete revolution corresponds to the angle 2π in the absolute angular measure customary in physics, and the above expression gives the amount by which the radius sun-planet exceeds this angle during the interval between one perihelion and the next.) In this expression a represents the major semi-axis of the ellipse, e its eccentricity, c the velocity of light, and T the period of revolution of the planet. Our result may also be stated as follows: According to the general theory of relativity, the major axis of the ellipse rotates round the sun in the same

APPENDIX III

sense as the orbital motion of the planet. Theory requires that this rotation should amount to 43 seconds of arc per century for the planet Mercury, but for the other planets of our solar system its magnitude should be so small that it would necessarily escape detection.¹

In point of fact, astronomers have found that the theory of Newton does not suffice to calculate the observed motion of Mercury with an exactness corresponding to that of the delicacy of observation attainable at the present time. After taking account of all the disturbing influences exerted on Mercury by the remaining planets, it was found (Leverrier — 1859 — and Newcomb — 1895) that an unexplained perihelial movement of the orbit of Mercury remained over, the amount of which does not differ sensibly from the above-mentioned + 43 seconds of arc per century. The uncertainty of the empirical result amounts to a few seconds only.

(b) DEFLECTION OF LIGHT BY A
GRAVITATIONAL FIELD

In Section XXII it has been already mentioned that, according to the general theory of relativity, a ray of light will experience a curvature of its

¹ Especially since the next planet Venus has an orbit that is almost an exact circle, which makes it more difficult to locate the perihelion with precision.

EXPERIMENTAL CONFIRMATION 153

path when passing through a gravitational field, this curvature being similar to that experienced by the path of a body which is projected through a gravitational field. As a result of this theory, we should expect that a ray of light which is passing close to a heavenly body would be deviated towards the latter. For a ray of light which passes the sun at a distance of Δ sun-radii from its centre, the angle of deflection (α) should amount to

$$\alpha = \frac{1.7 \text{ seconds of arc}}{\Delta}.$$

It may be added that, according to the theory, half of this deflection is produced by the Newtonian field of attraction of the sun, and the other half by the geometrical modification ("curvature") of space caused by the sun.

This result admits of an experimental test by means of the photographic registration of stars during a total eclipse of the sun. The only reason why we must wait for a total eclipse is because at every other time the atmosphere is so strongly illuminated

by the light from the sun that the stars situated near the sun's disc are invisible. The predicted effect can be seen clearly from the accompanying

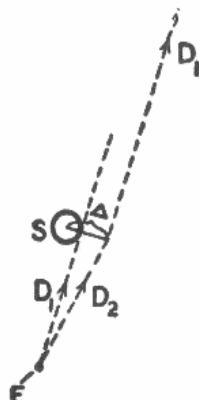


FIG. 5.

APPENDIX III

diagram. If the sun (S) were not present, a star which is practically infinitely distant would be seen in the direction D_1 , as observed from the earth. But as a consequence of the deflection of light from the star by the sun, the star will be seen in the direction D_2 , *i.e.* at a somewhat greater distance from the centre of the sun than corresponds to its real position.

In practice, the question is tested in the following way. The stars in the neighbourhood of the sun are photographed during a solar eclipse.

In addition, a second photograph of the same stars is taken when the sun is situated at another position in the sky, *i.e.* a few months earlier or later. As compared with the standard photograph, the positions of the stars on the eclipse-photograph ought to appear displaced radially outwards (away from the centre of the sun) by an amount corresponding to the angle α .

We are indebted to the Royal Society and to the Royal Astronomical Society for the investigation of this important deduction. Undaunted by the war and by difficulties of both a material and a psychological nature aroused by the war, these societies equipped two expeditions — to Sobral (Brazil) and to the island of Principe (West Africa) — and sent several of Britain's most celebrated astronomers (Eddington, Cottingham, Crommelin, Davidson), in order to obtain

photographs of the solar eclipse of 29th May, 1919. The relative discrepancies to be expected between the stellar photographs obtained during the eclipse and the comparison photographs amounted to a few hundredths of a millimetre only. Thus great accuracy was necessary in making the adjustments required for the taking of the photographs, and in their subsequent measurement.

The results of the measurements confirmed the theory in a thoroughly satisfactory manner. The rectangular components of the observed and of the calculated deviations of the stars (in seconds of arc) are set forth in the following table of results:

Number of the Star.	First Co-ordinate.		Second Co-ordinate.	
	Observed.	Calculated.	Observed.	Calculated.
11 . .	- 0·19	- 0·22	+ 0·16	+ 0·02
5 . .	+ 0·29	+ 0·31	- 0·46	- 0·43
4 . .	+ 0·11	+ 0·10	+ 0·83	+ 0·74
3 . .	+ 0·20	+ 0·12	+ 1·00	+ 0·87
6 . .	+ 0·10	+ 0·04	+ 0·57	+ 0·40
10 . .	- 0·08	+ 0·09	+ 0·35	+ 0·32
2 . .	+ 0·95	+ 0·85	- 0·27	- 0·09

(c) DISPLACEMENT OF SPECTRAL LINES TOWARDS THE RED

In Section XXIII it has been shown that in a system K' which is in rotation with regard to a Galileian system K , clocks of identical construc-

APPENDIX III

tion, and which are considered at rest with respect to the rotating reference-body, go at rates which are dependent on the positions of the clocks. We shall now examine this dependence quantitatively. A clock, which is situated at a distance r from the centre of the disc, has a velocity relative to K which is given by

$$v = \omega r,$$

where ω represents the ^{*} velocity of rotation of the disc K' with respect to K . If v_0 represents the number of ticks of the clock per unit time ("rate" of the clock) relative to K when the clock is at rest, then the "rate" of the clock (v) when it is moving relative to K with a velocity v , but at rest with respect to the disc, will, in accordance with Section XII, be given by

$$v = v_0 \sqrt{1 - \frac{v^2}{c^2}},$$

or with sufficient accuracy by

$$v = v_0 \left(1 - \frac{1}{2} \frac{v^2}{c^2} \right).$$

This expression may also be stated in the following form:

$$v = v_0 \left(1 - \frac{1}{c^2} \frac{\omega^2 r^2}{2} \right).$$

If we represent the difference of potential of the centrifugal force between the position of the clock and the centre of the disc by ϕ , i.e. the work,

[^{*} The word "angular" was inserted here in later editions. — J.M.]

EXPERIMENTAL CONFIRMATION 157

considered negatively, which must be performed on the unit of mass against the centrifugal force in order to transport it from the position of the clock on the rotating disc to the centre of the disc, then we have

$$\phi = -\frac{\omega^2 r^2}{2}.$$

From this it follows that

$$v = v_0 \left(1 + \frac{\phi}{c^2} \right).$$

In the first place, we see from this expression that two clocks of identical construction will go at different rates when situated at different distances from the centre of the disc. This result is also valid from the standpoint of an observer who is rotating with the disc.

Now, as judged from the disc, the latter is in a gravitational field of potential ϕ , hence the result we have obtained will hold quite generally for gravitational fields. Furthermore, we can regard an atom which is emitting spectral lines as a clock, so that the following statement will hold:

An atom absorbs or emits light of a frequency which is dependent on the potential of the gravitational field in which it is situated.

The frequency of an atom situated on the surface of a heavenly body will be somewhat less than the frequency of an atom of the same

APPENDIX III

element which is situated in free space (or on the surface of a smaller celestial body).

Now $\phi = -K \frac{M}{r}$, where K is Newton's constant of gravitation, and M is the mass of the heavenly body. Thus a displacement towards the red ought to take place for spectral lines produced at the surface of stars as compared with the spectral lines of the same element produced at the surface of the earth, the amount of this displacement being

$$\frac{\nu_0 - \nu}{\nu_0} = \frac{K}{c^2} \frac{M}{r}.$$

For the sun, the displacement towards the red predicted by theory amounts to about two millionths of the wave-length. A trustworthy calculation is not possible in the case of the stars, because in general neither the mass M nor the radius r is known.

It is an open question whether or not this effect exists, and at the present time astronomers are working with great zeal towards the solution. Owing to the smallness of the effect in the case of the sun, it is difficult to form an opinion as to its existence. Whereas Grebe and Bachem (Bonn), as a result of their own measurements and those of Evershed and Schwarzschild on the cyanogen bands, have placed the existence of the effect almost beyond doubt, other investigators, par-

ticularly St. John, have been led to the opposite opinion in consequence of their measurements.

Mean displacements of lines towards the less refrangible end of the spectrum are certainly revealed by statistical investigations of the fixed stars; but up to the present the examination of the available data does not allow of any definite decision being arrived at, as to whether or not these displacements are to be referred in reality to the effect of gravitation. The results of observation have been collected together, and discussed in detail from the standpoint of the question which has been engaging our attention here, in a paper by E. Freundlich entitled "Zur Prüfung der allgemeinen Relativitäts-Theorie" (*Die Naturwissenschaften*, 1919, No. 35, p. 520: Julius Springer, Berlin).

At all events, a definite decision will be reached during the next few years. If the displacement of spectral lines towards the red by the gravitational potential does not exist, then the general theory of relativity will be untenable. On the other hand, if the cause of the displacement of spectral lines be definitely traced to the gravitational potential, then the study of this displacement will furnish us with important information as to the mass of the heavenly bodies.

771

FÉUE WHI

BIBLIOGRAPHY

WORKS IN ENGLISH ON EINSTEIN'S THEORY

INTRODUCTORY

The Foundations of Einstein's Theory of Gravitation: Erwin Freundlich (translation by H. L. Brose). Camb. Univ. Press, 1920.

Space and Time in Contemporary Physics: Moritz Schlick (translation by H. L. Brose). Clarendon Press, Oxford, 1920.

THE SPECIAL THEORY

The Principle of Relativity: E. Cunningham. Camb. Univ. Press.

Relativity and the Electron Theory: E. Cunningham, Monographs on Physics. Longmans, Green & Co.

The Theory of Relativity: L. Silberstein. Macmillan & Co.

The Space-Time Manifold of Relativity: E. B. Wilson and G. N. Lewis, *Proc. Amer. Soc. Arts & Science*, vol. xlviii., No. 11, 1912.

THE GENERAL THEORY

Report on the Relativity Theory of Gravitation: A. S. Eddington. Fleetway Press Ltd., Fleet Street, London.

On Einstein's Theory of Gravitation and its Astronomical Consequences: W. de Sitter, *M. N. Roy. Astron. Soc.*, lxxvi. p. 699, 1916; lxxvii. p. 155, 1916; lxxviii. p. 3, 1917.

On Einstein's Theory of Gravitation: H. A. Lorentz, *Proc. Amsterdam Acad.*, vol. xix. p. 1341, 1917.

Space, Time and Gravitation: W. de Sitter: *The Observatory*, No. 505, p. 412. Taylor & Francis, Fleet Street, London.

BIBLIOGRAPHY

The Total Eclipse of 29th May 1919, and the Influence of Gravitation on Light: A. S. Eddington, *ibid.*, March, 1919.

Discussion on the Theory of Relativity: M. N. Roy. *Astron. Soc.*, vol. lxxx., No. 2., p. 96, December 1919.

The Displacement of Spectrum Lines and the Equivalence Hypothesis: W. G. Duffield, *M. N. Roy. Astron. Soc.*, vol. lxxx.; No. 3, p. 262, 1920.

Space, Time and Gravitation: A. S. Eddington. Camb. Univ. Press, 1920.

ALSO, CHAPTERS IN

The Mathematical Theory of Electricity and Magnetism: J. H. Jeans (4th edition). Camb. Univ. Press, 1920.

The Electron Theory of Matter: O. W. Richardson. Camb. Univ. Press.

INDEX

775

FÉUE WHI

INDEX

- Aberration, 59
- Absorption of energy, 54
- Acceleration, 76, 78, 83
- Action at a distance, 57
- Addition of velocities, 19, 46
- Adjacent points, 105
- Aether, 62
 - drift, 62, 63
- Arbitrary substitutions, 116
- Astronomy, 8, 121
- Astronomical day, 12
- Axioms, 2, 149
 - truth of, 2
- Bachem, 158
- Basis of theory, 52
- “Being,” 78, 128
- β -rays, 59
- Biology, 149
- Cartesian system of co-ordinates, 7, 100, 147
- Cathode rays, 59
- Celestial mechanics, 125
- Centrifugal force, 94, 156
- Chest, 78
- Classical mechanics, 9, 13, 16, 20, 36, 52, 84, 121, 123, 150
 - truth of, 15
- Clocks, 11, 28, 95, 96, 112, 114, 117–120, 121, 135, 155
 - rate of, 156
- Conception of mass, 54
 - position, 6
- Conservation of energy, 54, 121
 - impulse, 121
 - mass, 54, 56
- Continuity, 113
- Continuum, 65, 98
 - Continuum, two-dimensional, 112
 - three-dimensional, 67
 - four-dimensional, 106, 108, 109, 112, 147
 - space-time, 93, 108–114
 - Euclidean, 99, 101, 104, 110
 - non-Euclidean, 102, 107
- Co-ordinate differences, 109
 - differentials, 109
 - planes, 38
- Cottingham, 154
- Counter-point, 133
- Co-variant, 51
- Crommelin, 154
- Curvature of light-rays, 124, 152
 - space, 153
- Curvilinear motion, 88
- Cyanogen bands, 158
- Darwinian theory, 149
- Davidson, 154
- Deductive thought, 148
- Derivation of laws, 52

166**INDEX**

- De Sitter, 21
 Displacement of spectral lines, 124, 155
 Distance (line-interval), 3, 5, 8, 34, 35, 99, 104, 129
 — physical interpretation of, 5
 — relativity of, 34
 Doppler principle, 59
 Double stars, 21
 Eclipse of star, 21
 Eddington, 124, 154
 Electricity, 90
 Electrodynamics, 15, 24, 48, 52, 90
 Electromagnetic theory, 58
 — waves, 75
 Electron, 52, 60
 — electrical masses of, 60
 Electrostatics, 90
 Elliptical space, 133
 Empirical laws, 148
 Encounter (space-time coincidence), 113
 Equivalent, 16
 Euclidean geometry, 1, 2, 68, 97, 101, 104, 128, 129, 135, 147
 — propositions of, 3, 8
 — space, 68, 102, 147
 Evershed, 158
 Experience, 59, 70
 Faraday, 56, 74
 FitzGerald, 63
 Fixed stars, 12
 Fizeau, 46, 58, 61
 Fizeau, experiment of, 46
 Frequency of atom, 157
 Galilei, 12
 — transformation, 40, 43, 45, 50, 61
 Galilean system of co-ordinates, 13, 15, 17, 54, 93, 108, 116, 119
 Gauss, 102, 103, 106
 Gaussian co-ordinates, 103–105, 112, 114–118
 General theory of relativity, 69–124, 115
 Geometrical ideas, 2, 3
 — propositions, 1
 — — truth of, 2–4
 Gravitation, 75, 82, 92, 121
 Gravitational field, 75, 79, 87, 91, 111, 116, 119, 120, 136
 — — potential of, 157
 — mass, 76, 81, 121
 Grebe, 158
 Group-density of stars, 126
 Helmholtz, 128
 Heuristic value of relativity, 50
 Induction, 148, 149
 Inertia, 77
 Inertial mass, 55, 76, 81, 120, 121
 Instantaneous photograph (snapshot), 141
 Intensity of gravitational field, 127

INDEX

167

- Intuition, 148
 Ions, 53
 Kepler, 152
 Kinetic energy, 53, 121
 Lattice, 128
 Laws of Galilei-Newton, 15
 Law of inertia, 12, 71, 72, 78^{*}
 Laws of Nature, 70, 84, 118
 Leverrier, 123, 152
 Light-signal, 40, 139, 143
 Light-stimulus, 40
 Limiting velocity (*c*), 43, 44
 Lines of force, 126
 Lorentz, H. A., 24, 48, 52, 58,
 59–63
 — transformation, 39, 46, 50,
 108, 116, 139, 143, 144,
 146
 —— (generalised), 144
 Mach, E., 86
 Magnetic field, 74
 Manifold (*see* Continuum)
 Mass of heavenly bodies, 159
 Matter, 120
 Maxwell, 49, 52, 56–59, 61
 — fundamental equations, 56,
 90
 Measurement of length, 101
 Measuring-rod, 5, 6, 34, 95,
 96, 112, 119, 121, 132,
 135, 141
 Mercury, 123, 152
 — orbit of, 123, 152
 Michelson, 62–64
 Minkowski, 65–68, 108, 147
 Morley, 63, 64
 Motion, 16, 70
 — of heavenly bodies, 16, 17,
 52, 122, 135
 Newcomb, 152
 Newton, 12, 86, 122, 126, 150
 Newton's constant of gravitation, 158
 — law of gravitation, 57, 94,
 127, 149
 — law of motion, 76
 Non-Euclidean geometry, 128
 Non-Galileian reference-bodies, 117
 Non-uniform motion, 72
 Optics, 15, 24, 52
 Organ-pipe, note of, 17
 Parabola, 9, 10
 Path-curve, 10
 Perihelion of Mercury, 150–
 152
 Physics, 8
 — of measurement, 7
 Place specification, 6
 Plane, 1, 128, 129
 Poincaré, 128
 Point, 1
 Point-mass, energy of, 54
 Position, 9
 Principle of relativity, 15–17,
 23, 24, 70
 Processes of Nature, 50
 Propagation of light, 21, 23,
 24, 36, 108, 143
 —— in liquid, 47
 —— in gravitational fields, 88

[^{*} 170 — J.M.]

INDEX

- Quasi-Euclidean universe, 136
 Quasi-spherical universe, 136
- Radiation, 55
 Radioactive substances, 59
 Reference-body, 5, 7, 8–11, 22, 28, 31, 32, 44, 70
 — rotating, 94
 — mollusk, 118–120
 Relative position, 3
 — velocity, 141
 Rest, 17
 Riemann, 102, 128, 132
 Rotation, 95, 147
- Schwarzschild, 158
 Seconds-clock, 44
 Seeliger, 125, 127
 Simultaneity, 26, 29–32, 96
 — relativity of, 31
 Size-relations, 107
 Solar eclipse, 89, 153, 155
 Space, 9, 62, 65, 125
 — conception of, 24
 Space co-ordinates, 66, 96, 118
 Space-interval, 36, 67
 — point, 118
 Space, two-dimensional, 128
 — three-dimensional, 147
 Special theory of Relativity, 1–68, 24
 Spherical surface, 129
 — space, 132, 133
 St. John, 159
 Stellar universe, 126
 — photographs, 153
- Straight line, 1–3, 9, 10, 97, 105, 129
 System of co-ordinates, 5, 10, 11
- Terrestrial space, 18
 Theory, 148
 — truth of, 149
 Three-dimensional, 65
 Time, conception of, 24, 61, 125
 — co-ordinate, 66, 118
 — in Physics, 26, 117, 146
 — of an event, 28, 32
 — — interval, 36, 67
 Trajectory, 10
 “Truth,” 2
- Uniform translation, 14, 69
 Universe (World), structure of, 128, 135
 — circumference of, 131
 Universe, elliptical, 133, 136
 — Euclidean, 130, 132
 — space expanse (radius) of, 137
 — spherical, 132, 136
- Value of π , 97, 130
 Velocity of light, 11, 21, 22, 89, 143
 Venus, 152
 Weight (heaviness), 77
 World, 65, 66, 130, 147
 World-point, 147
 — radius, 133
 — sphere, 130, 131
 Zeeman, 48

' ' JUST THE MATHS ' '

by

A.J. Hobson**TEACHING UNITS - TABLE OF CONTENTS**

(Average number of pages = 1038 ,140 = 7.4 per unit)

All units are presented as .PDF files

[**\(Home\)**](#) [**\(Foreword\)**](#) [**\(About the Author\)**](#)**UNIT 1.1 - ALGEBRA 1 - INTRODUCTION TO ALGEBRA**

- 1.1.1 The Language of Algebra
- 1.1.2 The Laws of Algebra
- 1.1.3 Priorities in Calculations
- 1.1.4 Factors
- 1.1.5 Exercises
- 1.1.6 Answers to exercises (6 pages)

UNIT 1.2 - ALGEBRA 2 - NUMBERWORK

- 1.2.1 Types of number
- 1.2.2 Decimal numbers
- 1.2.3 Use of electronic calculators
- 1.2.4 Scientific notation
- 1.2.5 Percentages
- 1.2.6 Ratio
- 1.2.7 Exercises
- 1.2.8 Answers to exercises (8 pages)

UNIT 1.3 - ALGEBRA 3 - INDICES AND RADICALS (OR SURDS)

- 1.3.1 Indices
- 1.3.2 Radicals (or Surds)
- 1.3.3 Exercises
- 1.3.4 Answers to exercises (8 pages)

UNIT 1.4 - ALGEBRA 4 - LOGARITHMS

- 1.4.1 Common logarithms
- 1.4.2 Logarithms in general
- 1.4.3 Useful Results
- 1.4.4 Properties of logarithms
- 1.4.5 Natural logarithms
- 1.4.6 Graphs of logarithmic and exponential functions
- 1.4.7 Logarithmic scales
- 1.4.8 Exercises
- 1.4.9 Answers to exercises (10 pages)

UNIT 1.5 - ALGEBRA 5 - MANIPULATION OF ALGEBRAIC EXPRESSIONS

- 1.5.1 Simplification of expressions
- 1.5.2 Factorisation

- 1.5.3 Completing the square in a quadratic expression
- 1.5.4 Algebraic Fractions
- 1.5.5 Exercises
- 1.5.6 Answers to exercises (9 pages)

UNIT 1.6 - ALGEBRA 6 - FORMULAE AND ALGEBRAIC EQUATIONS

- 1.6.1 Transposition of formulae
- 1.6.2 Solution of linear equations
- 1.6.3 Solution of quadratic equations
- 1.6.4 Exercises
- 1.6.5 Answers to exercises (7 pages)

UNIT 1.7 - ALGEBRA 7 - SIMULTANEOUS LINEAR EQUATIONS

- 1.7.1 Two simultaneous linear equations in two unknowns
- 1.7.2 Three simultaneous linear equations in three unknowns
- 1.7.3 Ill-conditioned equations
- 1.7.4 Exercises
- 1.7.5 Answers to exercises (6 pages)

UNIT 1.8 - ALGEBRA 8 - POLYNOMIALS

- 1.8.1 The factor theorem
- 1.8.2 Application to quadratic and cubic expressions
- 1.8.3 Cubic equations
- 1.8.4 Long division of polynomials
- 1.8.5 Exercises
- 1.8.6 Answers to exercises (8 pages)

UNIT 1.9 - ALGEBRA 9 - THE THEORY OF PARTIAL FRACTIONS

- 1.9.1 Introduction
- 1.9.2 Standard types of partial fraction problem
- 1.9.3 Exercises
- 1.9.4 Answers to exercises (7 pages)

UNIT 1.10 - ALGEBRA 10 - INEQUALITIES 1

- 1.10.1 Introduction
- 1.10.2 Algebraic rules for inequalities
- 1.10.3 Intervals
- 1.10.4 Exercises
- 1.10.5 Answers to exercises (5 pages)

UNIT 1.11 - ALGEBRA 11 - INEQUALITIES 2

- 1.11.1 Recap on modulus, absolute value or numerical value
- 1.11.2 Interval inequalities
- 1.11.3 Exercises
- 1.11.4 Answers to exercises (5 pages)

UNIT 2.1 - SERIES 1 - ELEMENTARY PROGRESSIONS AND SERIES

- 2.1.1 Arithmetic progressions
- 2.1.2 Arithmetic series
- 2.1.3 Geometric progressions
- 2.1.4 Geometric series
- 2.1.5 More general progressions and series
- 2.1.6 Exercises

2.1.7 Answers to exercises (12 pages)

UNIT 2.2 - SERIES 2 - BINOMIAL SERIES

2.2.1 Pascal's Triangle

2.2.2 Binomial Formulae

2.2.3 Exercises

2.2.4 Answers to exercises (9 pages)

UNIT 2.3 - SERIES 3 - ELEMENTARY CONVERGENCE AND DIVERGENCE

2.3.1 The definitions of convergence and divergence

2.3.2 Tests for convergence and divergence (positive terms)

2.3.3 Exercises

2.3.4 Answers to exercises (13 pages)

UNIT 2.4 - SERIES 4 - FURTHER CONVERGENCE AND DIVERGENCE

2.4.1 Series of positive and negative terms

2.4.2 Absolute and conditional convergence

2.4.3 Tests for absolute convergence

2.4.4 Power series

2.4.5 Exercises

2.4.6 Answers to exercises (9 pages)

UNIT 3.1 - TRIGONOMETRY 1 - ANGLES AND TRIGONOMETRIC FUNCTIONS

3.1.1 Introduction

3.1.2 Angular measure

3.1.3 Trigonometric functions

3.1.4 Exercises

3.1.5 Answers to exercises (6 pages)

UNIT 3.2 - TRIGONOMETRY 2 - GRAPHS OF TRIGONOMETRIC FUNCTIONS

3.2.1 Graphs of elementary trigonometric functions

3.2.2 Graphs of more general trigonometric functions

3.2.3 Exercises

3.2.4 Answers to exercises (7 pages)

UNIT 3.3 - TRIGONOMETRY 3 - APPROXIMATIONS AND INVERSE

FUNCTIONS

3.3.1 Approximations for trigonometric functions

3.3.2 Inverse trigonometric functions

3.3.3 Exercises

3.3.4 Answers to exercises (6 pages)

UNIT 3.4 - TRIGONOMETRY 4 - SOLUTION OF TRIANGLES

3.4.1 Introduction

3.4.2 Right-angled triangles

3.4.3 The sine and cosine rules

3.4.4 Exercises

3.4.5 Answers to exercises (5 pages)

UNIT 3.5 - TRIGONOMETRY 5 - TRIGONOMETRIC IDENTITIES AND WAVE-FORMS

3.5.1 Trigonometric identities

3.5.2 Amplitude, wave-length, frequency and phase-angle

3.5.3 Exercises

3.5.4 Answers to exercises (8 pages)

UNIT 4.1 - HYPERBOLIC FUNCTIONS 1 - DEFINITIONS, GRAPHS AND IDENTITIES

- 4.1.1 Introduction
- 4.1.2 Definitions
- 4.1.3 Graphs of hyperbolic functions
- 4.1.4 Hyperbolic identities
- 4.1.5 Osborn's rule
- 4.1.6 Exercises
- 4.1.7 Answers to exercises (7 pages)

UNIT 4.2 - HYPERBOLIC FUNCTIONS 2 - INVERSE HYPERBOLIC FUNCTIONS

- 4.2.1 Introduction
- 4.2.2 The proofs of the standard formulae
- 4.2.3 Exercises
- 4.2.4 Answers to exercises (6 pages)

UNIT 5.1 - GEOMETRY 1 - CO-ORDINATES, DISTANCE AND GRADIENT

- 5.1.1 Co-ordinates
- 5.1.2 Relationship between polar & cartesian co-ordinates
- 5.1.3 The distance between two points
- 5.1.4 Gradient
- 5.1.5 Exercises
- 5.1.6 Answers to exercises (5 pages)

UNIT 5.2 - GEOMETRY 2 - THE STRAIGHT LINE

- 5.2.1 Preamble
- 5.2.2 Standard equations of a straight line
- 5.2.3 Perpendicular straight lines
- 5.2.4 Change of origin
- 5.2.5 Exercises
- 5.2.6 Answers to exercises (8 pages)

UNIT 5.3 - GEOMETRY 3 - STRAIGHT LINE LAWS

- 5.3.1 Introduction
- 5.3.2 Laws reducible to linear form
- 5.3.3 The use of logarithmic graph paper
- 5.3.4 Exercises
- 5.3.5 Answers to exercises (7 pages)

UNIT 5.4 - GEOMETRY 4 - ELEMENTARY LINEAR PROGRAMMING

- 5.4.1 Feasible Regions
- 5.4.2 Objective functions
- 5.4.3 Exercises
- 5.4.4 Answers to exercises (9 pages)

UNIT 5.5 - GEOMETRY 5 - CONIC SECTIONS (THE CIRCLE)

- 5.5.1 Introduction
- 5.5.2 Standard equations for a circle
- 5.5.3 Exercises
- 5.5.4 Answers to exercises (5 pages)

UNIT 5.6 - GEOMETRY 6 - CONIC SECTIONS (THE PARABOLA)

- 5.6.1 Introduction (the standard parabola)
- 5.6.2 Other forms of the equation of a parabola
- 5.6.3 Exercises
- 5.6.4 Answers to exercises (6 pages)

UNIT 5.7 - GEOMETRY 7 - CONIC SECTIONS (THE ELLIPSE)

- 5.7.1 Introduction (the standard ellipse)
- 5.7.2 A more general form for the equation of an ellipse
- 5.7.2 Exercises
- 5.7.3 Answers to exercises (4 pages)

UNIT 5.8 - GEOMETRY 8 - CONIC SECTIONS (THE HYPERBOLA)

- 5.8.1 Introduction (the standard hyperbola)
- 5.8.2 Asymptotes
- 5.8.3 More general forms for the equation of a hyperbola
- 5.8.4 The rectangular hyperbola
- 5.8.5 Exercises
- 5.8.6 Answers to exercises (8 pages)

UNIT 5.9 - GEOMETRY 9 - CURVE SKETCHING IN GENERAL

- 5.9.1 Symmetry
- 5.9.2 Intersections with the co-ordinate axes
- 5.9.3 Restrictions on the range of either variable
- 5.9.4 The form of the curve near the origin
- 5.9.5 Asymptotes
- 5.9.6 Exercises
- 5.9.7 Answers to exercises (10 pages)

UNIT 5.10 - GEOMETRY 10 - GRAPHICAL SOLUTIONS

- 5.10.1 The graphical solution of linear equations
- 5.10.2 The graphical solution of quadratic equations
- 5.10.3 The graphical solution of simultaneous equations
- 5.10.4 Exercises
- 5.10.5 Answers to exercises (7 pages)

UNIT 5.11 - GEOMETRY 11 - POLAR CURVES

- 5.11.1 Introduction
- 5.11.2 The use of polar graph paper
- 5.11.3 Exercises
- 5.11.4 Answers to exercises (10 pages)

UNIT 6.1 - COMPLEX NUMBERS 1 - DEFINITIONS AND ALGEBRA

- 6.1.1 The definition of a complex number
- 6.1.2 The algebra of complex numbers
- 6.1.3 Exercises
- 6.1.4 Answers to exercises (8 pages)

UNIT 6.2 - COMPLEX NUMBERS 2 - THE ARGAND DIAGRAM

- 6.2.1 Introduction
- 6.2.2 Graphical addition and subtraction
- 6.2.3 Multiplication by j
- 6.2.4 Modulus and argument
- 6.2.5 Exercises

6.2.6 Answers to exercises (7 pages)

UNIT 6.3 - COMPLEX NUMBERS 3 - THE POLAR AND EXPONENTIAL FORMS

6.3.1 The polar form

6.3.2 The exponential form

6.3.3 Products and quotients in polar form

6.3.4 Exercises

6.3.5 Answers to exercises (8 pages)

UNIT 6.4 - COMPLEX NUMBERS 4 - POWERS OF COMPLEX NUMBERS

6.4.1 Positive whole number powers

6.4.2 Negative whole number powers

6.4.3 Fractional powers & De Moivre's Theorem

6.4.4 Exercises

6.4.5 Answers to exercises (5 pages)

UNIT 6.5 - COMPLEX NUMBERS 5 - APPLICATIONS TO TRIGONOMETRIC IDENTITIES

6.5.1 Introduction

6.5.2 Expressions for $\cos n\theta$, $\sin n\theta$ in terms of $\cos \theta$, $\sin \theta$

6.5.3 Expressions for $\cos^n \theta$ and $\sin^n \theta$ in terms of sines and cosines of whole multiples of θ

6.5.4 Exercises

6.5.5 Answers to exercises (5 pages)

UNIT 6.6 - COMPLEX NUMBERS 6 - COMPLEX LOCI

6.6.1 Introduction

6.6.2 The circle

6.6.3 The half-straight-line

6.6.4 More general loci

6.6.5 Exercises

6.6.6 Answers to exercises (6 pages)

UNIT 7.1 - DETERMINANTS 1 - SECOND ORDER DETERMINANTS

7.1.1 Pairs of simultaneous linear equations

7.1.2 The definition of a second order determinant

7.1.3 Cramer's Rule for two simultaneous linear equations

7.1.4 Exercises

7.1.5 Answers to exercises (7 pages)

UNIT 7.2 - DETERMINANTS 2 - CONSISTENCY AND THIRD ORDER DETERMINANTS

7.2.1 Consistency for three simultaneous linear equations in two unknowns

7.2.2 The definition of a third order determinant

7.2.3 The rule of Sarrus

7.2.4 Cramer's rule for three simultaneous linear equations in three unknowns

7.2.5 Exercises

7.2.6 Answers to exercises (10 pages)

UNIT 7.3 - DETERMINANTS 3 - FURTHER EVALUATION OF 3 X 3 DETERMINANTS

7.3.1 Expansion by any row or column

7.3.2 Row and column operations on determinants

7.3.3 Exercises

7.3.4 Answers to exercises (10 pages)

UNIT 7.4 - DETERMINANTS 4 - HOMOGENEOUS LINEAR EQUATIONS

- 7.4.1 Trivial and non-trivial solutions
- 7.4.2 Exercises
- 7.4.3 Answers to exercises (7 pages)

UNIT 8.1 - VECTORS 1 - INTRODUCTION TO VECTOR ALGEBRA

- 8.1.1 Definitions
- 8.1.2 Addition and subtraction of vectors
- 8.1.3 Multiplication of a vector by a scalar
- 8.1.4 Laws of algebra obeyed by vectors
- 8.1.5 Vector proofs of geometrical results
- 8.1.6 Exercises
- 8.1.7 Answers to exercises (7 pages)

UNIT 8.2 - VECTORS 2 - VECTORS IN COMPONENT FORM

- 8.2.1 The components of a vector
- 8.2.2 The magnitude of a vector in component form
- 8.2.3 The sum and difference of vectors in component form
- 8.2.4 The direction cosines of a vector
- 8.2.5 Exercises
- 8.2.6 Answers to exercises (6 pages)

UNIT 8.3 - VECTORS 3 - MULTIPLICATION OF ONE VECTOR BY ANOTHER

- 8.3.1 The scalar product (or 'dot' product)
- 8.3.2 Deductions from the definition of dot product
- 8.3.3 The standard formula for dot product
- 8.3.4 The vector product (or 'cross' product)
- 8.3.5 Deductions from the definition of cross product
- 8.3.6 The standard formula for cross product
- 8.3.7 Exercises
- 8.3.8 Answers to exercises (8 pages)

UNIT 8.4 - VECTORS 4 - TRIPLE PRODUCTS

- 8.4.1 The triple scalar product
- 8.4.2 The triple vector product
- 8.4.3 Exercises
- 8.4.4 Answers to exercises (7 pages)

UNIT 8.5 - VECTORS 5 - VECTOR EQUATIONS OF STRAIGHT LINES

- 8.5.1 Introduction
- 8.5.2 The straight line passing through a given point and parallel to a given vector
- 8.5.3 The straight line passing through two given points
- 8.5.4 The perpendicular distance of a point from a straight line
- 8.5.5 The shortest distance between two parallel straight lines
- 8.5.6 The shortest distance between two skew straight lines
- 8.5.7 Exercises
- 8.5.8 Answers to exercises (14 pages)

UNIT 8.6 - VECTORS 6 - VECTOR EQUATIONS OF PLANES

- 8.6.1 The plane passing through a given point and perpendicular to a given vector
- 8.6.2 The plane passing through three given points
- 8.6.3 The point of intersection of a straight line and a plane
- 8.6.4 The line of intersection of two planes

- 8.6.5 The perpendicular distance of a point from a plane
8.6.6 Exercises
8.6.7 Answers to exercises (9 pages)

UNIT 9.1 - MATRICES 1 - DEFINITIONS AND ELEMENTARY MATRIX ALGEBRA

- 9.1.1 Introduction
9.1.2 Definitions
9.1.3 The algebra of matrices (part one)
9.1.4 Exercises
9.1.5 Answers to exercises (8 pages)

UNIT 9.2 - MATRICES 2 - FURTHER MATRIX ALGEBRA

- 9.2.1 Multiplication by a single number
9.2.2 The product of two matrices
9.2.3 The non-commutativity of matrix products
9.2.4 Multiplicative identity matrices
9.2.5 Exercises
9.2.6 Answers to exercises (6 pages)

UNIT 9.3 - MATRICES 3 - MATRIX INVERSION AND SIMULTANEOUS EQUATIONS

- 9.3.1 Introduction
9.3.2 Matrix representation of simultaneous linear equations
9.3.3 The definition of a multiplicative inverse
9.3.4 The formula for a multiplicative inverse
9.3.5 Exercises
9.3.6 Answers to exercises (11 pages)

UNIT 9.4 - MATRICES 4 - ROW OPERATIONS

- 9.4.1 Matrix inverses by row operations
9.4.2 Gaussian elimination (the elementary version)
9.4.3 Exercises
9.4.4 Answers to exercises (10 pages)

UNIT 9.5 - MATRICES 5 - CONSISTENCY AND RANK

- 9.5.1 The consistency of simultaneous linear equations
9.5.2 The row-echelon form of a matrix
9.5.3 The rank of a matrix
9.5.4 Exercises
9.5.5 Answers to exercises (9 pages)

UNIT 9.6 - MATRICES 6 - EIGENVALUES AND EIGENVECTORS

- 9.6.1 The statement of the problem
9.6.2 The solution of the problem
9.6.3 Exercises
9.6.4 Answers to exercises (9 pages)

UNIT 9.7 - MATRICES 7 - LINEARLY INDEPENDENT AND NORMALISED EIGENVECTORS

- 9.7.1 Linearly independent eigenvectors
9.7.2 Normalised eigenvectors
9.7.3 Exercises
9.7.4 Answers to exercises (5 pages)

UNIT 9.8 - MATRICES 8 - CHARACTERISTIC PROPERTIES AND SIMILARITY

TRANSFORMATIONS

- 9.8.1 Properties of eigenvalues and eigenvectors
- 9.8.2 Similar matrices
- 9.8.3 Exercises
- 9.7.4 Answers to exercises (9 pages)

UNIT 9.9 - MATRICES 9 - MODAL AND SPECTRAL MATRICES

- 9.9.1 Assumptions and definitions
- 9.9.2 Diagonalisation of a matrix
- 9.9.3 Exercises
- 9.9.4 Answers to exercises (9 pages)

UNIT 9.10 - MATRICES 10 - SYMMETRIC MATRICES AND QUADRATIC FORMS

- 9.10.1 Symmetric matrices
- 9.10.2 Quadratic forms
- 9.10.3 Exercises
- 9.10.4 Answers to exercises (7 pages)

UNIT 10.1 - DIFFERENTIATION 1 - FUNTIONS AND LIMITS

- 10.1.1 Functional notation
- 10.1.2 Numerical evaluation of functions
- 10.1.3 Functions of a linear function
- 10.1.4 Composite functions
- 10.1.5 Indeterminate forms
- 10.1.6 Even and odd functions
- 10.1.7 Exercises
- 10.1.8 Answers to exercises (12 pages)

UNIT 10.2 - DIFFERENTIATION 2 - RATES OF CHANGE

- 10.2.1 Introduction
- 10.2.2 Average rates of change
- 10.2.3 Instantaneous rates of change
- 10.2.4 Derivatives
- 10.2.5 Exercises
- 10.2.6 Answers to exercises (7 pages)

UNIT 10.3 - DIFFERENTIATION 3 - ELEMENTARY TECHNIQUES OF DIFFERENTIATION

- 10.3.1 Standard derivatives
- 10.3.2 Rules of differentiation
- 10.3.3 Exercises
- 10.3.4 Answers to exercises (9 pages)

UNIT 10.4 - DIFFERENTIATION 4 - PRODUCTS, QUOTIENTS AND LOGARITHMIC DIFFERENTIATION

- 10.4.1 Products
- 10.4.2 Quotients
- 10.4.3 Logarithmic differentiation
- 10.4.4 Exercises
- 10.4.5 Answers to exercises (10 pages)

UNIT 10.5 - DIFFERENTIATION 5 - IMPLICIT AND PARAMETRIC FUNCTIONS

- 10.5.1 Implicit functions
- 10.5.2 Parametric functions

- 10.5.3 Exercises
- 10.5.4 Answers to exercises (5 pages)

UNIT 10.6 - DIFFERENTIATION 6 - DERIVATIVES OF INVERSE TRIGONOMETRIC FUNCTIONS

- 10.6.1 Summary of results
- 10.6.2 The derivative of an inverse sine
- 10.6.3 The derivative of an inverse cosine
- 10.6.4 The derivative of an inverse tangent
- 10.6.5 Exercises
- 10.6.6 Answers to exercises (7 pages)

UNIT 10.7 - DIFFERENTIATION 7 - DERIVATIVES OF INVERSE HYPERBOLIC FUNCTIONS

- 10.7.1 Summary of results
- 10.7.2 The derivative of an inverse hyperbolic sine
- 10.7.3 The derivative of an inverse hyperbolic cosine
- 10.7.4 The derivative of an inverse hyperbolic tangent
- 10.7.5 Exercises
- 10.7.6 Answers to exercises (7 pages)

UNIT 10.8 - DIFFERENTIATION 8 - HIGHER DERIVATIVES

- 10.8.1 The theory
- 10.8.2 Exercises
- 10.8.3 Answers to exercises (4 pages)

UNIT 11.1 - DIFFERENTIATION APPLICATIONS 1 - TANGENTS AND NORMALS

- 11.1.1 Tangents
- 11.1.2 Normals
- 11.1.3 Exercises
- 11.1.4 Answers to exercises (5 pages)

UNIT 11.2 - DIFFERENTIATION APPLICATIONS 2 - LOCAL MAXIMA, LOCAL MINIMA AND POINTS OF INFLEXION

- 11.2.1 Introduction
- 11.2.2 Local maxima
- 11.2.3 Local minima
- 11.2.4 Points of inflexion
- 11.2.5 The location of stationary points and their nature
- 11.2.6 Exercises
- 11.2.7 Answers to exercises (14 pages)

UNIT 11.3 - DIFFERENTIATION APPLICATIONS 3 - CURVATURE

- 11.3.1 Introduction
- 11.3.2 Curvature in cartesian co-ordinates
- 11.3.3 Exercises
- 11.3.4 Answers to exercises (6 pages)

UNIT 11.4 - DIFFERENTIATION APPLICATIONS 4 - CIRCLE, RADIUS AND CENTRE OF CURVATURE

- 11.4.1 Introduction
- 11.4.2 Radius of curvature
- 11.4.3 Centre of curvature
- 11.4.4 Exercises
- 11.4.5 Answers to exercises (5 pages)

UNIT 11.5 - DIFFERENTIATION APPLICATIONS 5 - MACLAURIN'S AND TAYLOR'S SERIES

- 11.5.1 Maclaurin's series
- 11.5.2 Standard series
- 11.5.3 Taylor's series
- 11.5.4 Exercises
- 11.5.5 Answers to exercises (10 pages)

UNIT 11.6 - DIFFERENTIATION APPLICATIONS 6 - SMALL INCREMENTS AND SMALL ERRORS

- 11.6.1 Small increments
- 11.6.2 Small errors
- 11.6.3 Exercises
- 11.6.4 Answers to exercises (8 pages)

UNIT 12.1 - INTEGRATION 1 - ELEMENTARY INDEFINITE INTEGRALS

- 12.1.1 The definition of an integral
- 12.1.2 Elementary techniques of integration
- 12.1.3 Exercises
- 12.1.4 Answers to exercises (11 pages)

UNIT 12.2 - INTEGRATION 2 - INTRODUCTION TO DEFINITE INTEGRALS

- 12.2.1 Definition and examples
- 12.2.2 Exercises
- 12.2.3 Answers to exercises (3 pages)

UNIT 12.3 - INTEGRATION 3 - THE METHOD OF COMPLETING THE SQUARE

- 12.3.1 Introduction and examples
- 12.3.2 Exercises
- 12.3.3 Answers to exercises (4 pages)

UNIT 12.4 - INTEGRATION 4 - INTEGRATION BY SUBSTITUTION IN GENERAL

- 12.4.1 Examples using the standard formula
- 12.4.2 Integrals involving a function and its derivative
- 12.4.3 Exercises
- 12.4.4 Answers to exercises (5 pages)

UNIT 12.5 - INTEGRATION 5 - INTEGRATION BY PARTS

- 12.5.1 The standard formula
- 12.5.2 Exercises
- 12.5.3 Answers to exercises (6 pages)

UNIT 12.6 - INTEGRATION 6 - INTEGRATION BY PARTIAL FRACTIONS

- 12.6.1 Introduction and illustrations
- 12.6.2 Exercises
- 12.6.3 Answers to exercises (4 pages)

UNIT 12.7 - INTEGRATION 7 - FURTHER TRIGONOMETRIC FUNCTIONS

- 12.7.1 Products of sines and cosines
- 12.7.2 Powers of sines and cosines
- 12.7.3 Exercises
- 12.7.4 Answers to exercises (7 pages)

UNIT 12.8 - INTEGRATION 8 - THE TANGENT SUBSTITUTIONS

- 12.8.1 The substitution $t = \tan x$
- 12.8.2 The substitution $t = \tan(x/2)$

- 12.8.3 Exercises
- 12.8.4 Answers to exercises (5 pages)

UNIT 12.9 - INTEGRATION 9 - REDUCTION FORMULAE

- 12.9.1 Indefinite integrals
- 12.9.2 Definite integrals
- 12.9.3 Exercises
- 12.9.4 Answers to exercises (7 pages)

UNIT 12.10 - INTEGRATION 10 - FURTHER REDUCTION FORMULAE

- 12.10.1 Integer powers of a sine
- 12.10.2 Integer powers of a cosine
- 12.10.3 Wallis's formulae
- 12.10.4 Combinations of sines and cosines
- 12.10.5 Exercises
- 12.10.6 Answers to exercises (8 pages)

UNIT 13.1 - INTEGRATION APPLICATIONS 1 - THE AREA UNDER A CURVE

- 13.1.1 The elementary formula
- 13.1.2 Definite integration as a summation
- 13.1.3 Exercises
- 13.1.4 Answers to exercises (6 pages)

UNIT 13.2 - INTEGRATION APPLICATIONS 2 - MEAN AND ROOT MEAN SQUARE VALUES

- 13.2.1 Mean values
- 13.2.2 Root mean square values
- 13.2.3 Exercises
- 13.2.4 Answers to exercises (4 pages)

UNIT 13.3 - INTEGRATION APPLICATIONS 3 - VOLUMES OF REVOLUTION

- 13.3.1 Volumes of revolution about the x-axis
- 13.3.2 Volumes of revolution about the y-axis
- 13.3.3 Exercises
- 13.3.4 Answers to exercises (7 pages)

UNIT 13.4 - INTEGRATION APPLICATIONS 4 - LENGTHS OF CURVES

- 13.4.1 The standard formulae
- 13.4.2 Exercises
- 13.4.3 Answers to exercises (5 pages)

UNIT 13.5 - INTEGRATION APPLICATIONS 5 - SURFACES OF REVOLUTION

- 13.5.1 Surfaces of revolution about the x-axis
- 13.5.2 Surfaces of revolution about the y-axis
- 13.5.3 Exercises
- 13.5.4 Answers to exercises (7 pages)

UNIT 13.6 - INTEGRATION APPLICATIONS 6 - FIRST MOMENTS OF AN ARC

- 13.6.1 Introduction
- 13.6.2 First moment of an arc about the y-axis
- 13.6.3 First moment of an arc about the x-axis
- 13.6.4 The centroid of an arc
- 13.6.5 Exercises
- 13.6.6 Answers to exercises (11 pages)

UNIT 13.7 - INTEGRATION APPLICATIONS 7 - FIRST MOMENTS OF AN AREA

- 13.7.1 Introduction
- 13.7.2 First moment of an area about the y-axis
- 13.7.3 First moment of an area about the x-axis
- 13.7.4 The centroid of an area
- 13.7.5 Exercises
- 13.7.6 Answers to exercises (12 pages)

UNIT 13.8 - INTEGRATION APPLICATIONS 8 - FIRST MOMENTS OF A VOLUME

- 13.8.1 Introduction
- 13.8.2 First moment of a volume of revolution about a plane through the origin, perpendicular to the x-axis
- 13.8.3 The centroid of a volume
- 13.8.4 Exercises
- 13.8.5 Answers to exercises (10 pages)

UNIT 13.9 - INTEGRATION APPLICATIONS 9 - FIRST MOMENTS OF A SURFACE OF REVOLUTION

- 13.9.1 Introduction
- 13.9.2 Integration formulae for first moments
- 13.9.3 The centroid of a surface of revolution
- 13.9.4 Exercises
- 13.9.5 Answers to exercises (11 pages)

UNIT 13.10 - INTEGRATION APPLICATIONS 10 - SECOND MOMENTS OF AN ARC

- 13.10.1 Introduction
- 13.10.2 The second moment of an arc about the y-axis
- 13.10.3 The second moment of an arc about the x-axis
- 13.10.4 The radius of gyration of an arc
- 13.10.5 Exercises
- 13.10.6 Answers to exercises (11 pages)

UNIT 13.11 - INTEGRATION APPLICATIONS 11 - SECOND MOMENTS OF AN AREA (A)

- 13.11.1 Introduction
- 13.11.2 The second moment of an area about the y-axis
- 13.11.3 The second moment of an area about the x-axis
- 13.11.4 Exercises
- 13.11.5 Answers to exercises (8 pages)

UNIT 13.12 - INTEGRATION APPLICATIONS 12 - SECOND MOMENTS OF AN AREA (B)

- 13.12.1 The parallel axis theorem
- 13.12.2 The perpendicular axis theorem
- 13.12.3 The radius of gyration of an area
- 13.12.4 Exercises
- 13.12.5 Answers to exercises (8 pages)

UNIT 13.13 - INTEGRATION APPLICATIONS 13 - SECOND MOMENTS OF A VOLUME (A)

- 13.13.1 Introduction
- 13.13.2 The second moment of a volume of revolution about the y-axis
- 13.13.3 The second moment of a volume of revolution about the x-axis
- 13.13.4 Exercises
- 13.13.5 Answers to exercises (8 pages)

UNIT 13.14 - INTEGRATION APPLICATIONS 14 - SECOND MOMENTS OF A VOLUME (B)

- 13.14.1 The parallel axis theorem
- 13.14.2 The radius of gyration of a volume
- 13.14.3 Exercises
- 13.14.4 Answers to exercises (6 pages)

UNIT 13.15 - INTEGRATION APPLICATIONS 15 - SECOND MOMENTS OF A SURFACE OF REVOLUTION

- 13.15.1 Introduction
- 13.15.2 Integration formulae for second moments
- 13.15.3 The radius of gyration of a surface of revolution
- 13.15.4 Exercises
- 13.15.5 Answers to exercises (9 pages)

UNIT 13.16 - INTEGRATION APPLICATIONS 16 - CENTRES OF PRESSURE

- 13.16.1 The pressure at a point in a liquid
- 13.16.2 The pressure on an immersed plate
- 13.16.3 The depth of the centre of pressure
- 13.16.4 Exercises
- 13.16.5 Answers to exercises (9 pages)

UNIT 14.1 - PARTIAL DIFFERENTIATION 1 - PARTIAL DERIVATIVES OF THE FIRST ORDER

- 14.1.1 Functions of several variables
- 14.1.2 The definition of a partial derivative
- 14.1.3 Exercises
- 14.1.4 Answers to exercises (7 pages)

UNIT 14.2 - PARTIAL DIFFERENTIATION 2 - PARTIAL DERIVATIVES OF THE SECOND AND HIGHER ORDERS

- 14.2.1 Standard notations and their meanings
- 14.2.2 Exercises
- 14.2.3 Answers to exercises (5 pages)

UNIT 14.3 - PARTIAL DIFFERENTIATION 3 - SMALL INCREMENTS AND SMALL ERRORS

- 14.3.1 Functions of one independent variable - a recap
- 14.3.2 Functions of more than one independent variable
- 14.3.3 The logarithmic method
- 14.3.4 Exercises
- 14.3.5 Answers to exercises (10 pages)

UNIT 14.4 - PARTIAL DIFFERENTIATION 4 - EXACT DIFFERENTIALS

- 14.4.1 Total differentials
- 14.4.2 Testing for exact differentials
- 14.4.3 Integration of exact differentials
- 14.4.4 Exercises
- 14.4.5 Answers to exercises (9 pages)

UNIT 14.5 - PARTIAL DIFFERENTIATION 5 - PARTIAL DERIVATIVES OF COMPOSITE FUNCTIONS

- 14.5.1 Single independent variables
- 14.5.2 Several independent variables
- 14.5.3 Exercises
- 14.5.4 Answers to exercises (8 pages)

UNIT 14.6 - PARTIAL DIFFERENTIATION 6 - IMPLICIT FUNCTIONS

- 14.6.1 Functions of two variables
- 14.6.2 Functions of three variables
- 14.6.3 Exercises
- 14.6.4 Answers to exercises (6 pages)

UNIT 14.7 - PARTIAL DIFFERENTIATION 7 - CHANGE OF INDEPENDENT VARIABLE

- 14.7.1 Illustrations of the method
- 14.7.2 Exercises
- 14.7.3 Answers to exercises (5 pages)

UNIT 14.8 - PARTIAL DIFFERENTIATION 8 - DEPENDENT AND INDEPENDENT FUNCTIONS

- 14.8.1 The Jacobian
- 14.8.2 Exercises
- 14.8.3 Answers to exercises (8 pages)

UNIT 14.9 - PARTIAL DIFFERENTIATION 9 - TAYLOR'S SERIES FOR FUNCTIONS OF SEVERAL VARIABLES

- 14.9.1 The theory and formula
- 14.9.2 Exercises (8 pages)

UNIT 14.10 - PARTIAL DIFFERENTIATION 10 - STATIONARY VALUES FOR FUNCTIONS OF TWO VARIABLES

- 14.10.1 Introduction
- 14.10.2 Sufficient conditions for maxima and minima
- 14.10.3 Exercises
- 14.10.4 Answers to exercises (9 pages)

UNIT 14.11 - PARTIAL DIFFERENTIATION 11 - CONSTRAINED MAXIMA AND MINIMA

- 14.11.1 The substitution method
- 14.11.2 The method of Lagrange multipliers
- 14.11.3 Exercises
- 14.11.4 Answers to exercises (11 pages)

UNIT 14.12 - PARTIAL DIFFERENTIATION 12 - THE PRINCIPLE OF LEAST SQUARES

- 14.12.1 The normal equations
- 14.12.2 Simplified calculation of regression lines
- 14.12.3 Exercises
- 14.12.4 Answers to exercises (9 pages)

UNIT 15.1 - ORDINARY DIFFERENTIAL EQUATIONS 1 - FIRST ORDER EQUATIONS (A)

- 15.1.1 Introduction and definitions
- 15.1.2 Exact equations
- 15.1.3 The method of separation of the variables
- 15.1.4 Exercises
- 15.1.5 Answers to exercises (8 pages)

UNIT 15.2 - ORDINARY DIFFERENTIAL EQUATIONS 2 - FIRST ORDER EQUATIONS (B)

- 15.2.1 Homogeneous equations
- 15.2.2 The standard method
- 15.2.3 Exercises
- 15.2.4 Answers to exercises (6 pages)

UNIT 15.3 - ORDINARY DIFFERENTIAL EQUATIONS 3 - FIRST ORDER EQUATIONS (C)

- 15.3.1 Linear equations
- 15.3.2 Bernoulli's equation

- 15.3.3 Exercises
- 15.3.4 Answers to exercises (9 pages)

UNIT 15.4 - ORDINARY DIFFERENTIAL EQUATIONS 4 - SECOND ORDER EQUATIONS (A)

- 15.4.1 Introduction
- 15.4.2 Second order homogeneous equations
- 15.4.3 Special cases of the auxiliary equation
- 15.4.4 Exercises
- 15.4.5 Answers to exercises (9 pages)

UNIT 15.5 - ORDINARY DIFFERENTIAL EQUATIONS 5 - SECOND ORDER EQUATIONS (B)

- 15.5.1 Non-homogeneous differential equations
- 15.5.2 Determination of simple particular integrals
- 15.5.3 Exercises
- 15.5.4 Answers to exercises (6 pages)

UNIT 15.6 - ORDINARY DIFFERENTIAL EQUATIONS 6 - SECOND ORDER EQUATIONS (C)

- 15.6.1 Recap
- 15.6.2 Further types of particular integral
- 15.6.3 Exercises
- 15.6.4 Answers to exercises (7 pages)

UNIT 15.7 - ORDINARY DIFFERENTIAL EQUATIONS 7 - SECOND ORDER EQUATIONS (D)

- 15.7.1 Problematic cases of particular integrals
- 15.7.2 Exercises
- 15.7.3 Answers to exercises (6 pages)

UNIT 15.8 - ORDINARY DIFFERENTIAL EQUATIONS 8 - SIMULTANEOUS EQUATIONS (A)

- 15.8.1 The substitution method
- 15.8.2 Exercises
- 15.8.3 Answers to exercises (5 pages)

UNIT 15.9 - ORDINARY DIFFERENTIAL EQUATIONS 9 - SIMULTANEOUS EQUATIONS (B)

- 15.9.1 Introduction
- 15.9.2 Matrix methods for homogeneous systems
- 15.9.3 Exercises
- 15.9.4 Answers to exercises (8 pages)

UNIT 15.10 - ORDINARY DIFFERENTIAL EQUATIONS 10 - SIMULTANEOUS EQUATIONS (C)

- 15.10.1 Matrix methods for non-homogeneous systems
- 15.10.2 Exercises
- 15.10.3 Answers to exercises (10 pages)

UNIT 16.1 - LAPLACE TRANSFORMS 1 - DEFINITIONS AND RULES

- 16.1.1 Introduction
- 16.1.2 Laplace Transforms of simple functions
- 16.1.3 Elementary Laplace Transform rules
- 16.1.4 Further Laplace Transform rules
- 16.1.5 Exercises
- 16.1.6 Answers to exercises (10 pages)

UNIT 16.2 - LAPLACE TRANSFORMS 2 - INVERSE LAPLACE TRANSFORMS

- 16.2.1 The definition of an inverse Laplace Transform
- 16.2.2 Methods of determining an inverse Laplace Transform

- 16.2.3 Exercises
- 16.2.4 Answers to exercises (8 pages)

UNIT 16.3 - LAPLACE TRANSFORMS 3 - DIFFERENTIAL EQUATIONS

- 16.3.1 Examples of solving differential equations
- 16.3.2 The general solution of a differential equation
- 16.3.3 Exercises
- 16.3.4 Answers to exercises (7 pages)

UNIT 16.4 - LAPLACE TRANSFORMS 4 - SIMULTANEOUS DIFFERENTIAL EQUATIONS

- 16.4.1 An example of solving simultaneous linear differential equations
- 16.4.2 Exercises
- 16.4.3 Answers to exercises (5 pages)

UNIT 16.5 - LAPLACE TRANSFORMS 5 - THE HEAVISIDE STEP FUNCTION

- 16.5.1 The definition of the Heaviside step function
- 16.5.2 The Laplace Transform of $H(t - T)$
- 16.5.3 Pulse functions
- 16.5.4 The second shifting theorem
- 16.5.5 Exercises
- 16.5.6 Answers to exercises (8 pages)

UNIT 16.6 - LAPLACE TRANSFORMS 6 - THE DIRAC UNIT IMPULSE FUNCTION

- 16.6.1 The definition of the Dirac unit impulse function
- 16.6.2 The Laplace Transform of the Dirac unit impulse function
- 16.6.3 Transfer functions
- 16.6.4 Steady-state response to a single frequency input
- 16.6.5 Exercises
- 16.6.6 Answers to exercises (11 pages)

UNIT 16.7 - LAPLACE TRANSFORMS 7 - (AN APPENDIX)

- One view of how Laplace Transforms might have arisen (4 pages)

UNIT 16.8 - Z-TRANSFORMS 1 - DEFINITION AND RULES

- 16.8.1 Introduction
- 16.8.2 Standard Z-Transform definition and results
- 16.8.3 Properties of Z-Transforms
- 16.8.4 Exercises
- 16.8.5 Answers to exercises (10 pages)

UNIT 16.9 - Z-TRANSFORMS 2 - INVERSE Z-TRANSFORMS

- 16.9.1 The use of partial fractions
- 16.9.2 Exercises
- 16.9.3 Answers to exercises (6 pages)

UNIT 16.10 - Z-TRANSFORMS 3 - SOLUTION OF LINEAR DIFFERENCE EQUATIONS

- 16.10.1 First order linear difference equations
- 16.10.2 Second order linear difference equations
- 16.10.3 Exercises
- 16.10.4 Answers to exercises (9 pages)

UNIT 17.1 - NUMERICAL MATHEMATICS 1 - THE APPROXIMATE SOLUTION OF ALGEBRAIC EQUATIONS

- 17.1.1 Introduction

- 17.1.2 The Bisection method
- 17.1.3 The rule of false position
- 17.1.4 The Newton-Raphson method
- 17.1.5 Exercises
- 17.1.6 Answers to exercises (8 pages)

UNIT 17.2 - NUMERICAL MATHEMATICS 2 - APPROXIMATE INTEGRATION (A)

- 17.2.1 The trapezoidal rule
- 17.2.2 Exercises
- 17.2.3 Answers to exercises (4 pages)

UNIT 17.3 - NUMERICAL MATHEMATICS 3 - APPROXIMATE INTEGRATION (B)

- 17.3.1 Simpson's rule
- 17.3.2 Exercises
- 17.3.3 Answers to exercises (6 pages)

UNIT 17.4 - NUMERICAL MATHEMATICS 4 - FURTHER GAUSSIAN ELIMINATION

- 17.4.1 Gaussian elimination by ``partial pivoting'' with a check column
- 17.4.2 Exercises
- 17.4.3 Answers to exercises (4 pages)

UNIT 17.5 - NUMERICAL MATHEMATICS 5 - ITERATIVE METHODS FOR SOLVING SIMULTANEOUS LINEAR EQUATIONS

- 17.5.1 Introduction
- 17.5.2 The Gauss-Jacobi iteration
- 17.5.3 The Gauss-Seidel iteration
- 17.5.4 Exercises
- 17.5.5 Answers to exercises (7 pages)

UNIT 17.6 - NUMERICAL MATHEMATICS 6 - NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS (A)

- 17.6.1 Euler's unmodified method
- 17.6.2 Euler's modified method
- 17.6.3 Exercises
- 17.6.4 Answers to exercises (6 pages)

UNIT 17.7 - NUMERICAL MATHEMATICS 7 - NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS (B)

- 17.7.1 Picard's method
- 17.7.2 Exercises
- 17.7.3 Answers to exercises (6 pages)

UNIT 17.8 - NUMERICAL MATHEMATICS 8 - NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS (C)

- 17.8.1 Runge's method
- 17.8.2 Exercises
- 17.8.3 Answers to exercises (5 pages)

UNIT 18.1 - STATISTICS 1 - THE PRESENTATION OF DATA

- 18.1.1 Introduction
- 18.1.2 The tabulation of data
- 18.1.3 The graphical representation of data
- 18.1.4 Exercises

18.1.5 Selected answers to exercises (8 pages)

UNIT 18.2 - STATISTICS 2 - MEASURES OF CENTRAL TENDENCY

- 18.2.1 Introduction
- 18.2.2 The arithmetic mean (by coding)
- 18.2.3 The median
- 18.2.4 The mode
- 18.2.5 Quantiles
- 18.2.6 Exercises
- 18.2.7 Answers to exercises (9 pages)

UNIT 18.3 - STATISTICS 3 - MEASURES OF DISPERSION (OR SCATTER)

- 18.3.1 Introduction
- 18.3.2 The mean deviation
- 18.3.3 Practical calculation of the mean deviation
- 18.3.4 The root mean square (or standard) deviation
- 18.3.5 Practical calculation of the standard deviation
- 18.3.6 Other measures of dispersion
- 18.3.7 Exercises
- 18.3.8 Answers to exercises (6 pages)

UNIT 18.4 - STATISTICS 4 - THE PRINCIPLE OF LEAST SQUARES

- 18.4.1 The normal equations
- 18.4.2 Simplified calculation of regression lines
- 18.4.3 Exercises
- 18.4.4 Answers to exercises (6 pages)

UNIT 19.1 - PROBABILITY 1 - DEFINITIONS AND RULES

- 19.1.1 Introduction
- 19.1.2 Application of probability to games of chance
- 19.1.3 Empirical probability
- 19.1.4 Types of event
- 19.1.5 Rules of probability
- 19.1.6 Conditional probabilities
- 19.1.7 Exercises
- 19.1.8 Answers to exercises (5 pages)

UNIT 19.2 - PROBABILITY 2 - PERMUTATIONS AND COMBINATIONS

- 19.2.1 Introduction
- 19.2.2 Rules of permutations and combinations
- 19.2.3 Permutations of sets with some objects alike
- 19.2.4 Exercises
- 19.2.5 Answers to exercises (7 pages)

UNIT 19.3 - PROBABILITY 3 - RANDOM VARIABLES

- 19.3.1 Defining random variables
- 19.3.2 Probability distribution and probability density functions
- 19.3.3 Exercises
- 19.3.4 Answers to exercises (9 pages)

UNIT 19.4 - PROBABILITY 4 - MEASURES OF LOCATION AND DISPERSION

- 19.4.1 Common types of measure

19.4.2 Exercises

19.4.3 Answers to exercises (6 pages)

UNIT 19.5 - PROBABILITY 5 - THE BINOMIAL DISTRIBUTION

19.5.1 Introduction and theory

19.5.2 Exercises

19.5.3 Answers to exercises (5 pages)

UNIT 19.6 - PROBABILITY 6 - STATISTICS FOR THE BINOMIAL DISTRIBUTION

19.6.1 Construction of histograms

19.6.2 Mean and standard deviation of a binomial distribution

19.6.3 Exercises

19.6.4 Answers to exercises (10 pages)

UNIT 19.7 - PROBABILITY 7 - THE POISSON DISTRIBUTION

19.7.1 The theory

19.7.2 Exercises

19.7.3 Answers to exercises (5 pages)

UNIT 19.8 - PROBABILITY 8 - THE NORMAL DISTRIBUTION

19.8.1 Limiting position of a frequency polygon

19.8.2 Area under the normal curve

19.8.3 Normal distribution for continuous variables

19.8.4 Exercises

19.8.5 Answers to exercises (10 pages)

Page last changed: 3 October 2002

Contact for this page: [C J Judd](#)

FOREWORD

[**\(Home\)**](#) [**\(About the Author\)**](#) [**\(Teaching Units\)**](#) [**\(Teaching Slides\)**](#)

In 35 years of teaching mathematics to Engineers and Scientists, I have frequently been made aware (by students) of a common cry for help. "We're coping, generally, with our courses", they may say, "but it's Just the Maths". This is the title chosen for the package herein.

Traditional text-books and programmed learning texts can sometimes include a large amount of material which is not always needed for a particular course; and which can leave students feeling that there is too much to cope with. Many such texts are biased towards the mathematics required for specific engineering or scientific disciplines and emphasise the associated practical applications in their lists of tutorial examples. There can also be a higher degree of mathematical rigor than would be required by students who are not intending to follow a career in mathematics itself.

"Just the Maths" is a collection of separate units, in chronological topic-order, intended to service foundation level and first year degree level courses in higher education, especially those delivered in a modular style. Each unit represents, on average, the work to be covered in a typical two-hour session consisting of a lecture and a tutorial. However, since each unit attempts to deal with self-contained and, where possible, independent topics, it may sometimes require either more than or less than two hours spent on it.

"Just the Maths" does not have the format of a traditional text-book or a course of programmed learning; but it is written in a traditional pure-mathematics style with the minimum amount of formal rigor. By making use of the well-worn phrase, "it can be shown that", it is able to concentrate on the core mathematical techniques required by any scientist or engineer. The techniques are demonstrated by worked examples and reinforced by exercises that are few enough in number to allow completion, or near-completion, in a one-hour tutorial session. Answers to exercises are supplied at the end of each unit of work.

A.J. Hobson
January 2002

Page last changed: 9 September 2002
Contact for this page: [C J Judd](#)

ABOUT THE AUTHOR

[**\(Home\)**](#) [**\(Foreword\)**](#) [**\(Teaching Units\)**](#) [**\(Teaching Slides\)**](#)

Tony Hobson was, until retirement in November 2001, a Senior Lecturer in Mathematics of the School of Mathematical and Information Sciences at Coventry University. He graduated from the University College of Wales, Aberystwyth in 1964, with a BSc. Degree 2(i) in Pure Mathematics, and from Birmingham University in 1965, with an MSc. Degree in Pure Mathematics. His Dissertation for the MSc. Degree consisted of an investigation into the newer styles teaching Mathematics in the secondary schools of the 1960's with the advent of experiments such as the Midland Mathematics Experiment and the School Mathematics Project. His teaching career began in 1965 at the Rugby College of Engineering Technology where, as well as involvement with the teaching of Analysis and Projective Geometry to the London External Degree in Mathematics, he soon developed a particular interest in the teaching of Mathematics to Science and Engineering Students. This interest continued after the creation of the Polytechnics in 1971 and a subsequent move to the Coventry Polytechnic, later to become Coventry University. It was his main teaching interest throughout the thirty six years of his career; and it meant that much of the time he spent on research and personal development was in the area of curriculum development. In 1982 he became a Non-stipendiary Priest in the Church of England, an interest he maintained throughout his retirement. Tony Hobson died in December 2002.

The set of teaching units for "Just the Maths" has been the result of a pruning, honing and computer-processing exercise (over some four or five years) of **many** years' personal teaching materials, into a form which may be easily accessible to students of Science and Engineering in the future.

Page last changed: 17 January 2003

Contact for this page: [C.J.Judd](#)

“JUST THE MATHS”

UNIT NUMBER

1.1

ALGEBRA 1 (Introduction to algebra)

by

A.J. Hobson

- 1.1.1 The Language of Algebra
- 1.1.2 The Laws of Algebra
- 1.1.3 Priorities in Calculations
- 1.1.4 Factors
- 1.1.5 Exercises
- 1.1.5 Answers to exercises

UNIT 1.1 - ALGEBRA 1 - INTRODUCTION TO ALGEBRA

DEFINITION

An “**Algebra**” is any Mathematical system which uses the concepts of Equality ($=$), Addition (+), Subtraction (-), Multiplication (\times or .) and Division (\div).

Note:

The Algebra of Numbers is what we normally call “**Arithmetic**” and, as far as this unit is concerned, it is only the algebra of numbers which we shall be concerned with.

1.1.1 THE LANGUAGE OF ALGEBRA

Suppose we use the symbols a , b and c to denote numbers of arithmetic; then

(a) $a + b$ is called the “**sum of a and b** ”.

Note:

$a + a$ is usually abbreviated to $2a$,

$a + a + a$ is usually abbreviated to $3a$ and so on.

(b) $a - b$ is called the “**difference between a and b** ”.

(c) $a \times b$, $a.b$ or even just ab is called the “**product**” of a and b .

Notes:

(i)

$a.a$ is usually abbreviated to a^2 ,

$a.a.a$ is usually abbreviated to a^3 and so on.

(ii) $-1 \times a$ is usually abbreviated to $-a$ and is called the “**negation**” of a .

(d) $a \div b$ or $\frac{a}{b}$ is called the “**quotient**” or “**ratio**” of a and b .

(e) $\frac{1}{a}$, [also written a^{-1}], is called the “**reciprocal**” of a .

(f) $|a|$ is called the “**modulus**”, “**absolute value**” or “**numerical value**” of a . It can be defined by the two statements

$|a| = a$ when a is positive or zero;

$|a| = -a$ when a is negative or zero.

Note:

Further work on fractions (ratios) will appear later, but we state here for reference the rules for combining fractions together:

Rules for combining fractions together

1.

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

2.

$$\frac{a}{b} - \frac{c}{d} = \frac{ad - bc}{bd}$$

3.

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot c}{b \cdot d}$$

4.

$$\frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{a \cdot d}{b \cdot c}$$

EXAMPLES

1. How much more than the difference of 127 and 59 is the sum of 127 and 59 ?

Solution

The difference of 127 and 59 is $127 - 59 = 68$ and the sum of 127 and 59 is $127 + 59 = 186$. The sum exceeds the difference by $186 - 68 = 118$.

2. What is the reciprocal of the number which is 5 multiplied by the difference of 8 and 2 ?

Solution

We require the reciprocal of $5(8 - 2)$; that is, the reciprocal of 30. The answer is therefore $\frac{1}{30}$.

3. Calculate the value of $4\frac{2}{3} - 5\frac{1}{9}$ expressing the answer as a fraction.

Solution

Converting both numbers to a single fraction, we require

$$\frac{14}{3} - \frac{46}{9} = \frac{126 - 138}{27} = -\frac{12}{27} = -\frac{4}{9}.$$

We could also have observed that the ‘lowest common multiple’ (see later) of the two denominators, 3 and 9, is 9; hence we could write the alternative solution

$$\frac{42}{9} - \frac{46}{9} = -\frac{4}{9}.$$

4. Remove the modulus signs from the expression $|a - 2|$ in the cases when (i) a is greater than (or equal to) 2 and (ii) a is less than 2.

Solution

- (i) If a is greater than or equal to 2,

$$|a - 2| = a - 2;$$

- (ii) If a is less than 2,

$$|a - 2| = -(a - 2) = 2 - a.$$

1.1.2 THE LAWS OF ALGEBRA

If the symbols a , b and c denote numbers of arithmetic, then the following Laws are obeyed by them:

- (a) The Commutative Law of Addition $a + b = b + a$
- (b) The Associative Law of Addition $a + (b + c) = (a + b) + c$
- (c) The Commutative Law of Multiplication $a.b = b.a$
- (d) The Associative Law of Multiplication $a.(b.c) = (a.b).c$
- (e) The Distributive Laws $a.(b + c) = a.b + a.c$ and $(a + b).c = a.c + b.c$

Notes:

- (i) A consequence of the Distributive Laws is the rule for multiplying together a pair of bracketted expressions. It will be encountered more formally later, but we state it here for reference:

$$(a + b).(c + d) = a.c + b.c + a.d + b.d$$

- (ii) The alphabetical letters so far used for numbers in arithmetic have been taken from the **beginning** of the alphabet. These tend to be reserved for fixed quantities called **constants**. Letters from the **end** of the alphabet, such as w , x , y , z are normally used for quantities which may take many values, and are called **variables**.

1.1.3 PRIORITIES IN CALCULATIONS

Suppose that we encountered the expression $5 \times 6 - 4$. It would seem to be ambiguous, meaning either $30 - 4 = 26$ or $5 \times 2 = 10$.

However, we may remove the ambiguity by using brackets where necessary, together with a rule for precedence between the use of the brackets and the symbols $+$, $-$, \times and \div .

The rule is summarised in the abbreviation

B.O.D.M.A.S.

which means that the order of precedence is

B	brackets	()	First Priority
O	of	\times	Joint Second Priority
D	division	\div	Joint Second Priority
M	multiplication	\times	Joint Second Priority
A	addition	$+$	Joint Third Priority
S	subtraction	$-$	Joint Third Priority

$$\text{Thus, } 5 \times (6 - 4) = 5 \times 2 = 10$$

$$\text{but } 5 \times 6 - 4 = 30 - 4 = 26.$$

$$\text{Similarly, } 12 \div 3 - 1 = 4 - 1 = 3$$

$$\text{whereas } 12 \div (3 - 1) = 12 \div 2 = 6.$$

1.1.4 FACTORS

If a number can be expressed as a product of other numbers, each of those other numbers is called a “factor” of the original number.

EXAMPLES

1. We may observe that

$$70 = 2 \times 7 \times 5$$

so that the number 70 has factors of 2, 7 and 5. These three cannot be broken down into factors themselves because they are what are known as “prime” numbers (numbers whose only factors are themselves and 1). Hence the only factors of 70, apart from 70 and 1, are 2, 7 and 5.

2. Show that the numbers 78 and 182 have two common factors which are prime numbers.

The two factorisations are as follows:

$$78 = 2 \times 3 \times 13,$$

$$182 = 2 \times 7 \times 13.$$

The common factors are thus 2 and 13, both of which are prime numbers.

Notes:

(i) If two or more numbers have been expressed as a product of their prime factors, we may easily identify the prime factors which are common to all the numbers and hence obtain the “**highest common factor**”, h.c.f.

For example, $90 = 2 \times 3 \times 3 \times 5$ and $108 = 2 \times 2 \times 3 \times 3 \times 3$. Hence the h.c.f. = $2 \times 3 \times 3 = 18$

(ii) If two or more numbers have been expressed as a product of their prime factors, we may also identify the “**lowest common multiple**”, l.c.m.

For example, $15 = 3 \times 5$ and $20 = 2 \times 2 \times 5$. Hence the smallest number into which both 15 and 20 will divide requires two factors of 2 (for 20), one factor of 5 (for both 15 and 20) and one factor of 3 (for 15). The l.c.m. is thus $2 \times 2 \times 3 \times 5 = 60$.

(iii) If the numerator and denominator of a fraction have factors in common, then such factors may be cancelled to leave the fraction in its “**lowest terms**”.

For example $\frac{15}{105} = \frac{3 \times 5}{3 \times 5 \times 7} = \frac{1}{7}$.

1.1.5 EXERCISES

1. Find the sum and product of

- (a) 3 and 6; (b) 10 and 7; (c) 2, 3 and 6;
- (d) $\frac{3}{2}$ and $\frac{4}{11}$; (e) $1\frac{2}{5}$ and $7\frac{3}{4}$; (f) $2\frac{1}{7}$ and $5\frac{4}{21}$.

2. Find the difference between and quotient of

- (a) 18 and 9; (b) 20 and 5; (c) 100 and 20;
- (d) $\frac{3}{5}$ and $\frac{7}{10}$; (e) $3\frac{1}{4}$ and $2\frac{2}{9}$; (f) $1\frac{2}{3}$ and $5\frac{5}{6}$.

3. Evaluate the following expressions:

- (a) $6 - 2 \times 2$; (b) $(6 - 2) \times 2$;
- (c) $6 \div 2 - 2$; (d) $(6 \div 2) - 2$;
- (e) $6 - 2 + 3 \times 2$; (f) $6 - (2 + 3) \times 2$;
- (g) $(6 - 2) + 3 \times 2$; (h) $\frac{16}{-2}$; (i) $\frac{-24}{-3}$; (j) $(-6) \times (-2)$.

4. Place brackets in the following to make them correct:

- (a) $6 \times 12 - 3 + 1 = 55$; (b) $6 \times 12 - 3 + 1 = 68$;
- (c) $6 \times 12 - 3 + 1 = 60$; (d) $5 \times 4 - 3 + 2 = 7$;
- (e) $5 \times 4 - 3 + 2 = 15$; (f) $5 \times 4 - 3 + 2 = -5$.

5. Express the following as a product of prime factors:

- (a) 26; (b) 100; (c) 27; (d) 71;
- (e) 64; (f) 87; (g) 437; (h) 899.

6. Find the h.c.f of

- (a) 12, 15 and 21; (b) 16, 24 and 40; (c) 28, 70, 120 and 160;
- (d) 35, 38 and 42; (e) 96, 120 and 144.

7. Find the l.c.m of

- (a) 5, 6, and 8; (b) 20 and 30; (c) 7, 9 and 12;
- (d) 100, 150 and 235; (e) 96, 120 and 144.

1.1.6 ANSWERS TO EXERCISES

1. (a) 9, 18; (b) 17,70; (c) 11,36; (d) $\frac{41}{22}, \frac{6}{11}$; (e) $\frac{183}{20}, \frac{217}{20}$; (f) $\frac{154}{21}, \frac{545}{49}$.
2. (a) 9,2; (b) 15,4; (c) 80,5; (d) $-\frac{1}{10}, \frac{6}{7}$; (e) $\frac{37}{36}, \frac{117}{80}$; (f) $-\frac{25}{6}, \frac{2}{7}$.
3. (a) 2; (b) 8; (c) 1; (d) 1; (e) 10;
(f) -4; (g) 10; (h) -8; (i) 8; (j) 12;
4. (a) $6 \times (12 - 3) + 1 = 55$; (b) $6 \times 12 - (3 + 1) = 68$;
(c) $6 \times (12 - 3 + 1) = 60$; (d) $5 \times (4 - 3) + 2 = 7$;
(e) $5 \times 4 - (3 + 2) = 15$; (f) $5 \times (4 - [3 + 2]) = -5$.
5. (a) 2×13 ; (b) $2 \times 2 \times 5 \times 5$; (c) $3 \times 3 \times 3$; (d) 71×1 ;
(e) $2 \times 2 \times 2 \times 2 \times 2 \times 2$; (f) 3×29 ; (g) 19×23 ; (h) 29×31 .
6. (a) 3; (b) 8; (c) 2; (d) 1; (e) 24.
7. (a) 120; (b) 60; (c) 252; (d) 14100; (e) 1440.

“JUST THE MATHS”

UNIT NUMBER

1.2

ALGEBRA 2 (Numberwork)

by

A.J. Hobson

- 1.2.1 Types of number
- 1.2.2 Decimal numbers
- 1.2.3 Use of electronic calculators
- 1.2.4 Scientific notation
- 1.2.5 Percentages
- 1.2.6 Ratio
- 1.2.7 Exercises
- 1.2.8 Answers to exercises

UNIT 1.2 - - ALGEBRA 2 - NUMBERWORK

1.2.1 TYPES OF NUMBER

In this section (and elsewhere) the meaning of the following types of numerical quantity will need to be appreciated:

(a) NATURAL NUMBERS

These are the counting numbers 1, 2, 3, 4,

(b) INTEGERS

These are the positive and negative whole numbers and zero;
i.e.-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5,

(c) RATIONALS

These are the numbers which can be expressed as the ratio of two integers but can also be written as a terminating or recurring decimal (see also next section)

For example

$$\frac{2}{5} = 0.4$$

and

$$\frac{3}{7} = 0.428714287142871....$$

(d) IRRATIONALS

These are the numbers which cannot be expressed as either the ratio of two integers or a recurring decimal (see also next section)

Typical examples are numbers like

$$\pi \simeq 3.1415926.....$$

$$e \simeq 2.71828.....$$

$$\sqrt{2} \simeq 1.4142135....$$

$$\sqrt{5} \simeq 2.2360679....$$

The above four types of number form the system of “real numbers”.

1.2.2 DECIMAL NUMBERS

(a) Rounding to a specified number of decimal places

Most decimal quantities used in scientific work need to be approximated by “**rounding**” them (up or down as appropriate) to a specified number of decimal places, depending on the accuracy required.

When rounding to n decimal places, the digit in the n -th place is left as it is when the one after it is below 5; otherwise it is taken up by one digit.

EXAMPLES

1. $362.5863 = 362.586$ to 3 decimal places;
 $362.5863 = 362.59$ to 2 decimal places;
 $362.5863 = 362.6$ to 1 decimal place;
 $362.5863 = 363$ to the nearest whole number.
2. $0.02158 = 0.0216$ to 4 decimal places;
 $0.02158 = 0.022$ to three decimal places;
 $0.02158 = 0.02$ to 2 decimal places.

(b) Rounding to a specified number of significant figures

The first significant figure of a decimal quantity is the first non-zero digit from the left, whether it be before or after the decimal point.

Hence when rounding to a specified number of significant figures, we use the same principle as in (a), but starting from the first significant figure, then working to the right.

EXAMPLES

1. $362.5863 = 362.59$ to 5 significant figures;
 $362.5863 = 362.6$ to 4 significant figures;
 $362.5863 = 363$ to 3 significant figures;
 $362.5863 = 360$ to 2 significant figures;
 $362.5863 = 400$ to 1 significant figure.
2. $0.02158 = 0.0216$ to 3 significant figures; $0.02158 = 0.022$ to 2 significant figures;
 $0.02158 = 0.02$ to 1 significant figure.

1.2.3 THE USE OF ELECTRONIC CALCULATORS

(a) B.O.D.M.A.S.

The student will normally need to work to the instruction manual for the particular calculator being used; but care must be taken to remember the B.O.D.M.A.S. rule for priorities in calculations when pressing the appropriate buttons.

For example, in working out $7.25 + 3.75 \times 8.32$, the multiplication should be carried out first, then the addition. The answer is 38.45, not 91.52.

Similarly, in working out $6.95 \div [2.43 - 1.62]$, it is best to evaluate $2.43 - 1.62$, then generate its reciprocal with the $\frac{1}{x}$ button, then multiply by 6.95. The answer is 8.58, not 1.24

(b) Other Useful Numerical Functions

Other useful functions to become familiar with for scientific work with numbers are those indicated by labels such as \sqrt{x} , x^2 , x^y and $x^{\frac{1}{y}}$, using, where necessary, the “shift” control to bring the correct function into operation.

For example:

$$\sqrt{173} \simeq 13.153;$$

$$173^2 = 29929;$$

$$23^3 = 12167;$$

$$23^{\frac{1}{3}} \simeq 2.844$$

(c) The Calculator Memory

Familiarity with the calculator's memory facility will be essential for more complicated calculations in which various parts need to be stored temporarily while the different steps are being carried out.

For example, in order to evaluate

$$(1.4)^3 - 2(1.4)^2 + 5(1.4) - 3 \simeq 2.824$$

we need to store each of the four terms in the calculation (positively or negatively) then recall their total sum at the end.

1.2.4 SCIENTIFIC NOTATION

(a) Very large numbers, especially decimal numbers are customarily written in the form

$$a \times 10^n$$

where n is a positive integer and a lies between 1 and 10.

For instance,

$$521983677.103 = 5.21983677103 \times 10^8.$$

(b) Very small decimal numbers are customarily written in the form

$$a \times 10^{-n}$$

where n is a positive integer and a lies between 1 and 10.

For instance,

$$0.00045938 = 4.5938 \times 10^{-4}.$$

Note:

An electronic calculator will allow you to enter numbers in scientific notation by using the **EXP** or **EE** buttons.

EXAMPLES

1. Key in the number 3.90816×10^{57} on a calculator.

Press **3.90816 EXP 57**

In the display there will now be $3.90816\ 57$ or 3.90816×10^{57} .

2. Key in the number 1.5×10^{-27} on a calculator

Press **1.5 EXP 27 +/-**

In the display there will now be $1.5\ - 27$ or 1.5×10^{-27} .

Notes:

- (i) On a calculator or computer, scientific notation is also called *floating point notation*.
- (ii) When performing a calculation involving decimal numbers, it is always a good idea to check that the result is reasonable and that a major arithmetical error has not been made with the calculator.

For example,

$$69.845 \times 196.574 = 6.9845 \times 10^1 \times 1.96574 \times 10^3.$$

This product can be **estimated** for reasonableness as:

$$7 \times 2 \times 1000 = 14000.$$

The answer obtained by calculator is 13729.71 to two decimal places which is 14000 when rounded to the nearest 1000, indicating that the exact result could be reasonably expected.

(iii) If a set of measurements is made with an accuracy to a given number of significant figures, then it may be shown that any calculation involving those measurements will be accurate only to one significant figure more than the least number of significant figures in any measurement.

For example, the edges of a rectangular piece of cardboard are measured as 12.5cm and 33.43cm respectively and hence the area may be evaluated as

$$12.5 \times 33.43 = 417.875\text{cm}^2.$$

Since one of the edges is measured only to three significant figures, the area result is accurate only to four significant figures and hence must be stated as 417.9cm².

1.2.5 PERCENTAGES

Definition

A percentage is a fraction whose denominator is 100. We use the per-cent symbol, %, to represent a percentage.

For instance, the fraction $\frac{17}{100}$ may be written 17%

EXAMPLES

- Express $\frac{2}{5}$ as a percentage.

Solution

$$\frac{2}{5} = \frac{2}{5} \times \frac{20}{20} = \frac{40}{100} = 40\%$$

- Calculate 27% of 90.

Solution

$$27\% \text{ of } 90 = \frac{27}{100} \times 90 = \frac{27}{10} \times 9 = 24.3$$

3. Express 30% as a decimal.

Solution

$$30\% = \frac{30}{100} = 0.3$$

1.2.6 RATIO

Sometimes, a more convenient way of expressing the ratio of two numbers is to use a colon (:) in place of either the standard division sign (\div) or the standard notation for fractions.

For instance, the expression 7:3 could be used instead of either $7 \div 3$ or $\frac{7}{3}$. It denotes that two quantities are “in the ratio 7 to 3” which implies that the first number is seven thirds times the second number or, alternatively, the second number is three sevenths times the first number. Although more cumbersome, the ratio 7:3 could also be written $\frac{7}{3}:1$ or $1:\frac{3}{7}$.

EXAMPLES

1. Divide 170 in the ratio 3:2

Solution

We may consider that 170 is made up of $3 + 2 = 5$ parts, each of value $\frac{170}{5} = 34$.

Three of these make up a value of $3 \times 34 = 102$ and two of them make up a value of $2 \times 34 = 68$.

Thus 170 needs to be divided into 102 and 68.

2. Divide 250 in the ratio 1:3:4

Solution

This time, we consider that 250 is made up of $1 + 3 + 4 = 8$ parts, each of value $\frac{250}{8} = 31.25$. Three of these make up a value of $3 \times 31.25 = 93.75$ and four of them make up a value of $4 \times 31.25 = 125$.

Thus 250 needs to be divided into 31.25, 93.75 and 125.

1.2.7 EXERCISES

1. Write to 3 s.f.
 (a) 6962; (b) 70.406; (c) 0.0123;
 (d) 0.010991; (e) 45.607; (f) 2345.
2. Write 65.999 to
 (a) 4 s.f. (b) 3 s.f. (c) 2 s.f.
 (d) 1 s.f. (e) 2 d.p. (f) 1 d.p.
3. Compute the following in scientific notation:
 (a) $(0.003)^2 \times (0.00004) \times (0.00006) \times 5,000,000,000$;
 (b) $800 \times (0.00001)^2 \div (200,000)^4$.
4. Assuming that the following contain numbers obtained by measurement, use a calculator to determine their value and state the expected level of accuracy:
 - (a)
$$\frac{(13.261)^{0.5}(1.2)}{(5.632)^3};$$
 - (b)
$$\frac{(8.342)(-9.456)^3}{(3.25)^4}.$$
5. Calculate 23% of 124.
6. Express the following as percentages:
 (a) $\frac{9}{11}$; (b) $\frac{15}{20}$; (c) $\frac{9}{10}$; (d) $\frac{45}{50}$; (e) $\frac{75}{90}$.
7. A worker earns £400 a week, then receives a 6% increase. Calculate the new weekly wage.
8. Express the following percentages as decimals:
 (a) 50% (b) 36% (c) 75% (d) 100% (e) 12.5%
9. Divide 180 in the ratio 8:1:3
10. Divide 930 in the ratio 1:1:3
11. Divide 6 in the ratio 2:3:4
12. Divide 1200 in the ratio 1:2:3:4
13. A sum of £2600 is to be divided in the ratio $2\frac{3}{4} : 1\frac{1}{2} : 2\frac{1}{4}$. Calculate the amount of money in each part of the division.

1.2.8. ANSWERS TO EXERCISES

1. (a) 6960; (b) 70.4; (c) 0.0123;
(d) 0.0110; (e) 45.6; (f) 2350.
2. (a) 66.00; (b) 66.0; (c) 66;
(d) 70; (e) 66.00; (f) 66.0
3. (a) $1.08 - 04$ or 1.08×10^{-4} (b) $5 - 29$ or 5×10^{-29} ;
4. (a) 0.0245, accurate to three sig. figs. (b) -63.22 , accurate to four sig. figs.
5. 28.52
6. (a) 81.82% (b) 75% (c) 90% (d) 90% (e) 83.33%
7. £424.
8. (a) 0.5; (b) 0.36; (c) 0.75; (d) 1; (e) 0.125
9. 120, 15, 45.
10. 186, 186, 558.
11. 1.33, 2, 2.67
12. 120, 240, 360, 480
13. £1100, £600 £900.

“JUST THE MATHS”

UNIT NUMBER

1.3

ALGEBRA 3 (Indices and radicals (or surds))

by

A.J.Hobson

- 1.3.1 Indices
- 1.3.2 Radicals (or Surds)
- 1.3.3 Exercises
- 1.3.4 Answers to exercises

UNIT 1.3 - ALGEBRA 3 - INDICES AND RADICALS (or Surds)

1.3.1 INDICES

(a) Positive Integer Indices

It was seen earlier that, for any number a , a^2 denotes $a \cdot a$, a^3 denotes $a \cdot a \cdot a$, a^4 denotes $a \cdot a \cdot a \cdot a$ and so on.

Suppose now that a and b are arbitrary numbers and that m and n are natural numbers (i.e. positive whole numbers)

Then the following rules are the basic Laws of Indices:

Law No. 1

$$a^m \times a^n = a^{m+n}$$

Law No. 2

$$a^m \div a^n = a^{m-n}$$

assuming, for the moment, that m is greater than n .

Note:

It is natural to use this rule to give a definition to a^0 which would otherwise be meaningless.

Clearly $\frac{a^m}{a^m} = 1$ but the present rule for indices suggests that $\frac{a^m}{a^m} = a^{m-m} = a^0$. Hence, we **define** a^0 to be equal to 1.

Law No. 3

$$(a^m)^n = a^{mn}$$

$$a^m b^m = (ab)^m$$

EXAMPLE

Simplify the expression,

$$\frac{x^2 y^3}{z} \div \frac{xy}{z^5}.$$

Solution

The expression becomes

$$\frac{x^2 y^3}{z} \times \frac{z^5}{xy} = xy^2 z^4.$$

(b) Negative Integer Indices**Law No. 4**

$$a^{-1} = \frac{1}{a}$$

Note:

It has already been mentioned that a^{-1} means the same as $\frac{1}{a}$; and the logic behind this statement is to maintain the basic Laws of Indices for negative indices as well as positive ones.

For example $\frac{a^m}{a^{m+1}}$ is clearly the same as $\frac{1}{a}$ but, using Law No. 2 above, it could also be thought of as $a^{m-[m+1]} = a^{-1}$.

Law No. 5

$$a^{-n} = \frac{1}{a^n}$$

Note:

This time, we may observe that $\frac{a^m}{a^{m+n}}$ is clearly the same as $\frac{1}{a^n}$; but we could also use Law No. 2 to interpret it as $a^{m-[m+n]} = a^{-n}$

Law No. 6

$$a^{-\infty} = 0$$

Note:

Strictly speaking, no power of a number can ever be equal to zero, but Law No. 6 asserts that a very large negative power of a number a gives a very small value; the larger the negative power, the smaller will be the value.

EXAMPLE

Simplify the expression,

$$\frac{x^5y^2z^{-3}}{x^{-1}y^4z^5} \div \frac{z^2x^2}{y^{-1}}.$$

Solution

The expression becomes

$$x^5y^2z^{-3}xy^{-4}z^{-5}y^{-1}z^{-2}x^{-2} = x^4y^{-3}z^{-10}.$$

(c) Rational Indices

(i) Indices of the form $\frac{1}{n}$ where n is a natural number.

In order to preserve Law No. 3, we interpret $a^{\frac{1}{n}}$ to mean a number which gives the value a when it is raised to the power n . It is called an “ n -th Root of a ” and, sometimes there is more than one value.

ILLUSTRATION

$$81^{\frac{1}{4}} = \pm 3 \text{ but } (-27)^{\frac{1}{3}} = -3 \text{ only.}$$

(ii) Indices of the form $\frac{m}{n}$ where m and n are natural numbers with no common factor.

The expression $y^{\frac{m}{n}}$ may be interpreted in two ways as either $(y^m)^{\frac{1}{n}}$ or $(y^{\frac{1}{n}})^m$. It may be shown that both interpretations give the same result but, sometimes, the arithmetic is shorter with one rather than the other.

ILLUSTRATION

$$27^{\frac{2}{3}} = 3^2 = 9 \text{ or } 27^{\frac{2}{3}} = 729^{\frac{1}{3}} = 9.$$

Note:

It may be shown that all of the standard laws of indices may be used for fractional indices.

1.3.2 RADICALS (or Surds)

The symbol “ $\sqrt{}$ ” is called a “**radical**” (or “**surd**”). It is used to indicate the positive or “**principal**” square root of a number. Thus $\sqrt{16} = 4$ and $\sqrt{25} = 5$.

The number under the radical is called the “**radicand**”.

Most of our work on radicals will deal with square roots, but we may have occasion to use other roots of a number. For instance the **principal n-th root** of a number a is denoted by $\sqrt[n]{a}$, and is a number x such that $x^n = a$. The number n is called the **index** of the radical but, of course, when $n = 2$ we usually leave the index out.

ILLUSTRATIONS

1. $\sqrt[3]{64} = 4$ since $4^3 = 64$.
2. $\sqrt[3]{-64} = -4$ since $(-4)^3 = -64$.
3. $\sqrt[4]{81} = 3$ since $3^4 = 81$.
4. $\sqrt[5]{32} = 2$ since $2^5 = 32$.
5. $\sqrt[5]{-32} = -2$ since $(-2)^5 = -32$.

Note:

If the index of the radical is an odd number, then the radicand may be positive or negative; but if the index of the radical is an even number, then the radicand may not be negative since no even power of a negative number will ever give a negative result.

(a) Rules for Square Roots

In preparation for work which will follow in the next section, we list here the standard rules for square roots:

- (i) $(\sqrt{a})^2 = a$
- (ii) $\sqrt{a^2} = |a|$
- (iii) $\sqrt{ab} = \sqrt{a}\sqrt{b}$
- (iv) $\sqrt{\frac{a}{b}} = \frac{\sqrt{a}}{\sqrt{b}}$

assuming that all of the radicals can be evaluated.

ILLUSTRATIONS

1. $\sqrt{9 \times 4} = \sqrt{36} = 6$ and $\sqrt{9} \times \sqrt{4} = 3 \times 2 = 6$.
2. $\sqrt{\frac{144}{36}} = \sqrt{4} = 2$ and $\frac{\sqrt{144}}{\sqrt{36}} = \frac{12}{6} = 2$.

(b) Rationalisation of Radical (or Surd) Expressions.

It is often desirable to eliminate expressions containing radicals from the denominator of a quotient. This process is called

rationalising the denominator.

The process involves multiplying numerator and denominator of the quotient by the same amount - an amount which eliminates the radicals in the denominator (often using the fact that the square root of a number multiplied by itself gives just the number; i.e. $\sqrt{a} \cdot \sqrt{a} = a$). We illustrate with examples:

EXAMPLES

1. Rationalise the surd form $\frac{5}{4\sqrt{3}}$

Solution

We simply multiply numerator and denominator by $\sqrt{3}$ to give

$$\frac{5}{4\sqrt{3}} = \frac{5}{4\sqrt{3}} \times \frac{\sqrt{3}}{\sqrt{3}} = \frac{5\sqrt{3}}{12}.$$

2. Rationalise the surd form $\frac{3\sqrt{a}}{3\sqrt{b}}$

Solution

Here we observe that, if we can convert the denominator into the cube root of b^n , where n is a whole multiple of 3, then the square root sign will disappear.

We have

$$\frac{3\sqrt{a}}{3\sqrt{b}} = \frac{3\sqrt{a}}{3\sqrt{b}} \times \frac{3\sqrt{b^2}}{3\sqrt{b^2}} = \frac{3\sqrt{ab^2}}{3\sqrt{b^3}} = \frac{3\sqrt{ab^2}}{b}.$$

If the denominator is of the form $\sqrt{a} + \sqrt{b}$, we multiply the numerator and the denominator by the expression $\sqrt{a} - \sqrt{b}$ because

$$(\sqrt{a} + \sqrt{b})(\sqrt{a} - \sqrt{b}) = a - b.$$

3. Rationalise the surd form $\frac{4}{\sqrt{5}+\sqrt{2}}$.

Solution

Multiplying numerator and denominator by $\sqrt{5} - \sqrt{2}$ gives

$$\frac{4}{\sqrt{5} + \sqrt{2}} \times \frac{\sqrt{5} - \sqrt{2}}{\sqrt{5} - \sqrt{2}} = \frac{4\sqrt{5} - 4\sqrt{2}}{3}.$$

4. Rationalise the surd form $\frac{1}{\sqrt{3}-1}$.

Solution

Multiplying numerator and denominator by $\sqrt{3} + 1$ gives

$$\frac{1}{\sqrt{3}-1} \times \frac{\sqrt{3}+1}{\sqrt{3}+1} = \frac{\sqrt{3}+1}{2}.$$

(c) Changing numbers to and from radical form

The modulus of any number of the form $a^{\frac{m}{n}}$ can be regarded as the principal n -th root of a^m ; i.e.

$$| a^{\frac{m}{n}} | = \sqrt[n]{a^m}.$$

If a number of the type shown on the left is converted to the type on the right, we are said to have expressed it in radical form.

If a number of the type on the right is converted to the type on the left, we are said to have expressed it in exponential form.

Note:

The word “**exponent**” is just another word for “**power**” or “**index**” and the standard rules of indices will need to be used in questions of the type discussed here.

EXAMPLES

1. Express the number $x^{\frac{2}{5}}$ in radical form.

Solution

The answer is just

$$\sqrt[5]{x^2}.$$

2. Express the number $^3\sqrt{a^5b^4}$ in exponential form.

Solution

Here we have

$$^3\sqrt{a^5b^4} = (a^5b^4)^{\frac{1}{3}} = a^{\frac{5}{3}}b^{\frac{4}{3}}.$$

1.3.3 EXERCISES

1. Simplify

(a) $5^7 \times 5^{13}$; (b) $9^8 \times 9^5$; (c) $11^2 \times 11^3 \times 11^4$.

2. Simplify

(a) $\frac{15^3}{15^2}$; (b) $\frac{4^{18}}{4^9}$; (c) $\frac{5^{20}}{5^{19}}$.

3. Simplify

(a) a^7a^3 ; (b) a^4a^5 ;
 (c) $b^{11}b^{10}b$; (d) $3x^6 \times 5x^9$.

4. Simplify

(a) $(7^3)^2$; (b) $(4^2)^8$; (c) $(7^9)^2$.

5. Simplify

(a) $(x^2y^3)(x^3y^2)$; (b) $(2x^2)(3x^4)$;
 (c) $(a^2bc^2)(b^2ca)$; (d) $\frac{6c^2d^3}{3cd^2}$.

6. Simplify

(a) $(4^{-3})^2$ (b) $a^{13}a^{-2}$;
 (c) $x^{-9}x^{-7}$; (d) $x^{-21}x^2x$;
 (e) $\frac{x^2y^{-1}}{z^3} \div \frac{z^2}{x^{-1}y^3}$.

7. Without using a calculator, evaluate the following:

(a) $\frac{4^{-8}}{4^{-6}}$; (b) $\frac{3^{-5}}{3^{-8}}$.

8. Evaluate the following:

(a) $64^{\frac{1}{3}}$; (b) $144^{\frac{1}{2}}$;
 (c) $16^{-\frac{1}{4}}$; (d) $25^{-\frac{1}{2}}$;
 (e) $16^{\frac{3}{2}}$; (f) $125^{-\frac{2}{3}}$.

9. Simplify the following radicals:

(a) $-\sqrt[3]{-8}$; (b) $\sqrt{36x^4}$; (c) $\sqrt{\frac{9a^2}{36b^2}}$.

10. Rationalise the following surd forms:

(a) $\frac{\sqrt{2}}{\sqrt{3}}$; (b) $\frac{\sqrt[3]{18}}{\sqrt{2}}$; (c) $\frac{2+\sqrt{5}}{\sqrt{3}-2}$; (d) $\frac{\sqrt{a}}{\sqrt{a}+3\sqrt{b}}$.

11. Change the following to exponential form:

(a) $4\sqrt{7^2}$; (b) $5\sqrt{a^2b}$; (c) $3\sqrt{9^5}$.

12. Change the following to radical form:
 (a) $b^{\frac{3}{5}}$; (b) $r^{\frac{5}{3}}$; (c) $s^{\frac{7}{3}}$.

1.3.4 ANSWERS TO EXERCISES

1. (a) 5^{20} ; (b) 9^{13} ; (c) 11^9 .
2. (a) 15; (b) 4^9 ; (c) 5.
3. (a) a^{10} ; (b) a^9 ; (c) b^{22} ; (d) $15x^{15}$.
4. (a) 7^6 ; (b) 4^{16} ; (c) 7^{18} .
5. (a) x^5y^5 ; (b) $6x^6$; (c) $a^3b^3c^3$; (d) $2cd$.
6. (a) 4^{-6} ; (b) a^{11} ; (c) x^{-16} ; (d) x^{-18} ; (e) xy^2z^{-5} .
7. (a) $\frac{1}{16}$; (b) 27.
8. (a) 4; (b) ± 12 ; (c) $\pm \frac{1}{2}$;
 (d) $\pm \frac{1}{5}$; (e) ± 64 ; (f) $\frac{1}{25}$;
9. (a) 2; (b) $6x^2$; (c) $\left| \frac{a}{2b} \right|$.
10. (a) $\frac{\sqrt[3]{6}}{3}$; (b) $\frac{\sqrt[3]{72}}{2} = \sqrt[3]{9}$; (c) $-(2 + \sqrt{5})(2 + \sqrt{3})$; (d) $\frac{a - 3\sqrt{ab}}{a - 9b}$.
11. (a) $\left| 7^{\frac{1}{2}} \right|$; (b) $a^{\frac{2}{5}}b^{\frac{1}{5}}$; (c) $9^{\frac{5}{3}}$.
12. (a) $5\sqrt{b^3}$; (b) $3\sqrt{r^5}$; (c) $3\sqrt{s^7}$.

“JUST THE MATHS”

UNIT NUMBER

1.4

ALGEBRA 4 (Logarithms)

by

A.J.Hobson

- 1.4.1 Common logarithms
- 1.4.2 Logarithms in general
- 1.4.3 Useful Results
- 1.4.4 Properties of logarithms
- 1.4.5 Natural logarithms
- 1.4.6 Graphs of logarithmic and exponential functions
- 1.4.7 Logarithmic scales
- 1.4.8 Exercises
- 1.4.9 Answers to exercises

UNIT 1.4 - ALGEBRA 4 - LOGARITHMS

1.4.1 COMMON LOGARITHMS

The system of numbers with which we normally count and calculate has a base of 10; this means that each of the successive digits of a particular number correspond to that digit multiplied by a certain power of 10.

For example

$$73,520 = 7 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 2 \times 10^1.$$

Note:

Other systems (not discussed here) are sometimes used - such as the binary system which uses successive powers of 2.

The question now arises as to whether a given number can be expressed as a single power of 10, not necessarily an integer power. It will certainly need to be a **positive** number since powers of 10 are not normally negative (or even zero).

It can easily be verified by calculator, for instance that

$$1.99526 \simeq 10^{0.3}$$

and

$$2 \simeq 10^{0.30103}.$$

DEFINITION

In general, when it occurs that

$$x = 10^y,$$

for some positive number x , we say that y is the "**logarithm to base 10**" of x (or "**common logarithm**" of x) and we write

$$y = \log_{10} x.$$

EXAMPLES

1. $\log_{10} 1.99526 = 0.3$ from the illustrations above.
2. $\log_{10} 2 = 0.30103$ from the illustrations above.
3. $\log_{10} 1 = 0$ simply because $10^0 = 1$.

1.4.2 LOGARITHMS IN GENERAL

In practice, with scientific work, only two bases of logarithms are ever used; but it will be useful to include here a general discussion of the definition and properties of logarithms to **any** base so that unnecessary repetition may be avoided. We consider only positive bases of logarithms in the general discussion.

DEFINITION

If B is a fixed positive number and x is another positive number such that

$$x = B^y,$$

we say that y is the “**logarithm to base B** ” of x and we write

$$y = \log_B x.$$

EXAMPLES

1. $\log_B 1 = 0$ simply because $B^0 = 1$.
2. $\log_B B = 1$ simply because $B^1 = B$.
3. $\log_B 0$ doesn’t really exist because no power of B could ever be equal to zero. But, since a very large negative power of B will be a very small positive number, we usually write

$$\log_B 0 = -\infty.$$

1.4.3 USEFUL RESULTS

In preparation for the general properties of logarithms, we note the following two results which can be obtained directly from the definition of a logarithm:

- (a) For any positive number x ,

$$x = B^{\log_B x}.$$

In other words, any positive number can be expressed as a power of B without necessarily using a calculator.

We have simply replaced the y in the statement $x = B^y$ by $\log_B x$ in the equivalent statement $y = \log_B x$.

- (b) For any number y ,

$$y = \log_B B^y.$$

In other words, any number can be expressed in the form of a logarithm without necessarily using a calculator.

We have simply replaced x in the statement $y = \log_B x$ by B^y in the equivalent statement $x = B^y$.

1.4.4 PROPERTIES OF LOGARITHMS

The following properties were once necessary for performing numerical calculations before electronic calculators came into use. We do not use logarithms for this purpose nowadays; but we do need their properties for various topics in scientific mathematics.

(a) The Logarithm of Product.

$$\log_B p \cdot q = \log_B p + \log_B q.$$

Proof:

We need to show that, when $p \cdot q$ is expressed as a power of B , that power is the expression on the right hand side of the above formula.

From Result (a) of the previous section,

$$p \cdot q = B^{\log_B p} \cdot B^{\log_B q} = B^{\log_B p + \log_B q},$$

by elementary properties of indices.

The result therefore follows.

(b) The Logarithm of a Quotient

$$\log_B \frac{p}{q} = \log_B p - \log_B q.$$

Proof:

The proof is along similar lines to that in (i).

From Result (a) of the previous section,

$$\frac{p}{q} = \frac{B^{\log_B p}}{B^{\log_B q}} = B^{\log_B p - \log_B q},$$

by elementary properties of indices.

The result therefore follows.

(c) The Logarithm of an Exponential

$$\log_B p^n = n \log_B p,$$

where n need not be an integer.

Proof:

From Result (a) of the previous section,

$$p^n = (B^{\log_B p})^n = B^{n \log_B p},$$

by elementary properties of indices.

(d) The Logarithm of a Reciprocal

$$\log_B \frac{1}{q} = -\log_B q.$$

Proof:

This property may be proved in two ways as follows:

Method 1.

The left-hand side $= \log_B 1 - \log_B q = 0 - \log_B q = -\log_B q$.

Method 2.

The left-hand side $= \log_B q^{-1} = -\log_B q$.

(e) Change of Base

$$\log_B x = \frac{\log_A x}{\log_A B}.$$

Proof:

Suppose $y = \log_B x$, then $x = B^y$ and hence

$$\log_A x = \log_A B^y = y \log_A B.$$

Thus,

$$y = \frac{\log_A x}{\log_A B}$$

and the result follows.

Note:

The result shows that the logarithms of any set of numbers to a given base will be directly

proportional to the logarithms of the same set of numbers to another given base. This is simply because the number $\log_A B$ is a constant.

1.4.5 NATURAL LOGARITHMS

It was mentioned earlier that, in scientific work, only two bases of logarithms are ever used. One of these is base 10 and the other is a base which arises **naturally** out of elementary calculus when discussing the simplest available result for the “derivative” (rate of change) of a logarithm.

This other base turns out to be a non-recurring, non-terminating decimal quantity (irrational number) which is equal to 2.71828.....and clearly this would be inconvenient to write into the logarithm notation.

We therefore denote it by e to give the “**natural logarithm**” of a number, x , in the form $\log_e x$, although most scientific books use the alternative notation $\ln x$.

Note:

From the earlier change of base formula we can say that

$$\log_{10} x = \frac{\log_e x}{\log_e 10} \quad \text{and} \quad \log_e x = \frac{\log_{10} x}{\log_{10} e}.$$

EXAMPLES

1. Solve for x the indicial equation

$$4^{3x-2} = 26^{x+1}.$$

Solution

The secret of solving an equation where an unknown quantity appears in a power (or index or exponent) is to take logarithms of both sides first.

Here we obtain

$$(3x - 2) \log_{10} 4 = (x + 1) \log_{10} 26;$$

$$(3x - 2)0.6021 = (x + 1)1.4150;$$

$$1.8063x - 1.2042 = 1.4150x + 1.4150;$$

$$(1.8603 - 1.4150)x = 1.4150 + 1.2042;$$

$$0.3913x = 2.6192;$$

$$x = \frac{2.6192}{0.3913} \simeq 6.6936$$

2. Rewrite the expression

$$4x + \log_{10}(x+1) - \log_{10}x - \frac{1}{2}\log_{10}(x^3 + 2x^2 - x)$$

as the common logarithm of a single mathematical expression.

Solution

The secret here is to make sure that every term in the given expression is converted, where necessary, to a logarithm with no multiple in front of it or behind it. In this case, we need first to write $4x = \log_{10} 10^{4x}$ and $\frac{1}{2}\log_{10}(x^3 + 2x^2 - x) = \log_{10}(x^3 + 2x^2 - x)^{\frac{1}{2}}$.

We can then use the results for the logarithms of a product and a quotient to give

$$\log_{10} \frac{10^{4x}(x+1)}{x\sqrt{(x^3 + 2x^2 - x)}}.$$

3. Rewrite without logarithms the equation

$$2x + \ln x = \ln(x-7).$$

Solution

This time, we need to convert both sides to the natural logarithm of a single mathematical expression in order to remove the logarithms completely.

$$2x + \ln x = \ln e^{2x} + \ln x = \ln xe^{2x}.$$

Hence,

$$xe^{2x} = x - 7.$$

4. Solve for x the equation

$$6 \ln 4 + \ln 2 = 3 + \ln x.$$

Solution

In view of the facts that $6 \ln 4 = \ln 4^6$ and $3 = \ln e^3$, the equation can be written

$$\ln 2(4^6) = \ln xe^3.$$

Hence,

$$2(4^6) = xe^3,$$

so that

$$x = \frac{2(4^6)}{e^3} \simeq 407.856$$