Astroparticle Physics 43 (2013) 56–70

Contents lists available at SciVerse ScienceDirect

Astroparticle Physics

journal homepage: www.elsevier.com/locate/astropart

Cosmic ray acceleration

A.R. Bell

Clarendon Laboratory, University of Oxford, Parks Road, Oxford, OX1 3PU, UK

ARTICLE INFO

Article history:
Available online 12 June 2012

Keywords:
Cosmic rays
Particle acceleration
Shocks

ABSTRACT

This review describes the basic theory of cosmic ray acceleration by shocks including the plasma instabilities confining cosmic rays near the shock, the effect of the magnetic field orientation, the maximum cosmic ray energy and the shape of the cosmic ray spectrum. Attention is directed mainly towards Galactic cosmic rays accelerated by supernova remnants.

1. Introduction

Until the later part of the twentieth century, cosmic ray (CR) physics was a subject with its own distinct nature that separated it from much of astrophysics. It appeared that the greater part of astrophysics could be understood without reference to cosmic rays. Cosmic rays first became important to mainstream astrophysics with the development of radio telescopes since synchrotron radiation requires energetic electrons, which have to be accelerated, and magnetic field, which is often in close energy equipartition with the energetic electrons. With the expansion of observational techniques into the whole range of the electromagnetic expansion, high energy astrophysics embraced cosmic rays as an essential part of astrophysics since a substantial fraction of the available energy is often channelled into cosmic rays. Cosmic rays are also diagnostically important as a source of radiation. Cosmic ray electrons are responsible for synchrotron and inverse Compton emission in many parts of the electromagnetic spectrum, and it appears likely, although not yet completely certain, that gamma-rays generated through pion decay provide a direct window into in situ acceleration of TeV-PeV protons.

CR arriving at the Earth with energies up to and beyond 10^{20} eV have a Larmor radius greater than the size of the Galaxy and must have their origin in extreme conditions beyond our Galaxy. Arrival directions measured by the Auger array suggest an origin correlated with AGN for the highest energy CR [7], but this is far from certain and more data are needed [8]. Cosmic rays contribute a large fraction of the energy content of explosive environments from stellar to galactic scales, and they appear to play an important role in the generation of magnetic field. Since cosmic rays are important both dynamically and diagnostically it is essential that we understand their acceleration, transport, radiative emissions, and interaction with other components of astrophysical environments.

Cosmic rays arriving at the Earth consist mainly of high energy protons with smaller numbers of electrons and other nuclei. A century of detector development since their discovery has defined the shape and extent of the CR energy spectrum. The differential energy spectrum in the Galaxy extends as a E^{-s} power law from GeV energies to a few PeV with a spectral index $s = 2.6 - 2.7$ [55]. The spectrum steepens slightly at a few PeV before flattening at about an EeV and then turning over and terminating at a few 100 EeV [77]. This is usually referred to as a knee-ankle structure with the 'knee' at a few PeV and the 'ankle' at the less well defined energy of a few EeV. The Larmor radius of a proton with energy E_{PeV} in PeV gyrating in a magnetic field $B_{\mu G}$ in μG is $r_g = E_{PeV}/B_{\mu G}$ parsec. It is mostly assumed with good reason that CR with energies above the 'ankle' must be extragalactic in origin since their Larmor radius is much larger than the Galaxy, and conversely that CR with energies below the knee must originate within the Galaxy. Although much debated, the common picture is that the transition from a Galactic to an extragalactic origin occurs somewhere between the knee and the ankle. A heavy nucleus with charge Z has a Larmor radius Z times smaller than a proton with the same energy, so if protons can be accelerated within the Galaxy to a few PeV then it is reasonable to suppose that a heavy nucleus can be accelerated to a few times ZPeV within the Galaxy. Composition studies support this, but heavy nuclei cannot easily account for all CR in the energy range 1–100 PeV, and a further population of protons accelerated in the Galaxy beyond the knee may be needed [56,57]. Proton acceleration to the knee pushes shock acceleration theory to its limits when applied to shell-type supernova remnants (SNR) of the kind observed in our Galaxy, so CR between the knee and the ankle pose a severe challenge to our understanding.

The arrival directions of CR at all except the highest energies are scrambled by deflection in the interstellar or intergalactic magnetic field and give no information on their source, but the

E-mail address: t.bell1@physics.ox.ac.uk

0927-6505
http://dx.doi.org/10.1016/j.astropartphys.2012.05.022

large total CR energy in the Galaxy places severe limitations on possible sources. Supernova remnants (SNR) provide the largest energy input into the interstellar medium and, as discussed below, their associated shocks are natural particle accelerators, so it is usually assumed that Galactic CR are accelerated by SNR. To account for the Galactic CR energy budget, at least a few percent of the total SN energy has to be given to CR [10,2].

The development of Cherenkov telescopes sensitive to gamma-rays with energies approaching 100 TeV has dramatically expanded our knowledge of CR origins. Since gamma-rays are undeflected as they propagate from the source, Cherenkov telescopes point directly to the CR producing the gamma-rays. Observations by the HESS Cherenkov telescope of the supernova remnant RX J1713.7–3946 detect gamma-rays up to nearly 100 TeV with an angular resolution of 0.06 degrees [3,4]. A quantitative analysis of radiation from SNR at TeV energies can be found in Drury et al. [33]. This provides direct evidence that at least some SNR accelerate cosmic rays to within an order of magnitude of the energy of the knee since gamma-rays are generated by cosmic rays of about 7 times the gamma-ray energy, depending on the emission process [3]. It is not completely decided whether gamma-rays from SNR are emitted by ions as well as electrons, but discrimination between electron and ion sources is within reach through the measurement of the gamma-ray spectrum over a range of photon energies. When Cherenkov measurements are united with lower energy gamma-ray detections by the FERMI satellite it is becoming possible to distinguish between gamma-ray spectra characteristic of inverse-Compton and pion processes. Latest results indicate that pion processes dominate in at least one SNR [52]. CR electrons are also detectable through their synchrotron emission from radio to X-ray energies. X-ray synchrotron emission is mapped in great detail in some supernova remnants, giving direct evidence of in situ acceleration of TeV electrons.

Previous reviews of cosmic ray acceleration include Drury [32], Blandford and Eichler [24], Jones and Ellison [62] and Malkov and Drury [74]. Only a small fraction of the extensive literature on diffusive shock acceleration and related astrophysics can be covered in a review of the present length which considers mainly the plasma physics of CR acceleration by SNR. This review does not cover heliospheric shocks (e.g. [118]) which are generally weaker and more dependent on the local context due to their smaller spatial extent. Injection of initially thermal particles into the acceleration process and their subsequent acceleration to non-relativistic energies (e.g. [92]) are important aspects of heliospheric shocks not discussed here. Also, this review does not consider CR acceleration at shocks with relativistic velocities (e.g. [64,1,103,82,83]). A comprehensive review of observational signatures of cosmic ray acceleration can be found in Helder et al. [54].

2. Diffusive shock acceleration

More than 60 years ago Fermi [46] suggested that CR gain their energy from large scale fluid motions in the interstellar medium. In the second order Fermi process, CR are reflected elastically by moving magnetic field structures that might be anchored in interstellar clouds. A 'head-on' encounter between a CR and a cloud leads to the CR gaining energy. A 'tail-on' encounter leads to an energy loss, but the CR gains energy on average because head-on encounters are more frequent than tail-on encounters. If the typical cloud velocity is $\mathbf{u}$ and CR move at the speed of light c, the average fractional energy exchange on each encounter is of order u/c, and the excess of head-on over tail-on encounters is also u/c, making the mean energy gain on each encounter of order $(u/c)^2$, which makes the process second order. The second order Fermi process may play a role in CR acceleration in older SNR [80], but attention has moved to a faster first order Fermi process that operates in the environment of shocks. In the rest frame of a shock, the upstream plasma moves into the shock at the shock velocity u_s and exits downstream at a lower velocity u_s/r where r is the density compression ratio at the shock. $r = 4$ for a non-relativistic shock with a high Mach number. If a CR reflects back and forth between magnetic structures upstream and downstream of the shock, when viewed in an appropriate frame every encounter is head-on with a mean fractional increase in energy of order u_s/c producing relatively rapid acceleration. The key to the operation of first order Fermi shock acceleration is that CR trajectories are scattered by fluctuations in the magnetic field. If the local magnetic field were uniform, CR would easily escape the shock environment by streaming along magnetic field lines. It was shown around 1970 [66,117,99–101] that CR streaming excites fluctuations in the magnetic field in the form of Alfven waves propagating along the field lines with wavelengths on the scale of a CR Larmor radius. The resonance between the Larmor radius and the wavelength of the fluctuation produces a strong interaction that scatters CR so they propagate diffusively along the field lines [94]. CR perform a random walk along field lines which can lead to a particular CR crossing and recrossing the shock many times, gaining energy on each crossing.

The resulting CR spectrum can be derived either from solution of the Boltzmann equation for a CR distribution near a shock [65,9,25] or equivalently from the statistics of a random walk by a single particle [14,15]. These equivalent derivations were proposed independently, and they can be outlined as follows.

The derivation from single particle [14] can be separated into two steps:

(i) The first step is to derive an expression for the probability of a CR crossing and recrossing the shock m times before escaping downstream. From simple kinetic theory the rate at which CR cross from upstream to downstream is $n_s c/4$ where n_s is the CR number density at the shock. The rate at which CR are carried away downstream at velocity u_s/r by background fluid motions is $n_\infty(u_s/r)$ where n_∞ is the CR number density far downstream. In the diffusion approximation, valid for $u_s \ll c$, $n_s = n_\infty$. The ratio of the two rates prescribes that on average a CR crosses and recrosses the shock $rc/4u_s$ times. Since the CR has equal probability of being carried away downstream on each adventure into the downstream plasma, the probability of escape during one excursion into the downstream plasma is $4u_s/rc$. The probability of making at least m shock crossings is therefore $(1 - 4u_s/rc)^m$. Assuming a strong shock, $r = 4$ and the number of CR still present at the shock after m crossings is

$$N_m = (1 - u_s/c)^m N_0$$

where N_0 is the number of CR initially encountering the shock.

(ii) The second step in the derivation is to find an expression for the average energy gained after m shock crossings. By averaging over the various angles at which relativistic CR cross and re-cross a strong non-relativistic shock we find that the average energy gain ΔE on one cycle of crossing from upstream to downstream and back to upstream is $\Delta E = (u_s/c)E$, and the average energy after m such cycles is

$$E_m = (1 + u_s/c)^m E_0$$

where E_0 is the initial CR energy. Since (u_s/c) is small, $(1 + u_s/c)^m \approx 1/(1 - u_s/c)^m$, $(N_m/N_0) = (E_m/E_0)^{-1}$ and $N(E) \approx (E/E_0)^{-1} N_0$. $N(E) \propto E^{-1}$ is the integral energy spectrum of CR reaching at least energy E, so the differential energy spectrum $n(E) = -dN/dE$ is

$$n(E) \propto E^{-2}.$$

It can be shown by a similar argument that the equivalent p^{-2} momentum spectrum applies to non-relativistic as well as relativis-

tic CR. The corresponding energy spectrum of non-relativistic CR is proportional to E^{-1}.

The CR spectrum was also derived from the Boltzmann equation by Krymsky [65], Axford et al. [9] and Blandford and Ostriker [25] in independent papers. Consider a shock at position $x = 0$ in its rest frame with plasma flowing through the shock in the x direction. The time-dependent Boltzmann equation for the CR distribution function $f(x, \mathbf{p})$ in position and momentum $\mathbf{p}$ space is

$$\frac{\partial f}{\partial t} + (v_x + u)\frac{\partial f}{\partial x} - \frac{\partial u}{\partial x} p_x \frac{\partial f}{\partial p_x} + e\mathbf{v} \times \mathbf{B}.\frac{\partial f}{\partial \mathbf{p}} = C(f) \quad (1)$$

where the background fluid velocity u is considered small, f is defined in the local fluid rest frame, $\mathbf{B}$ is the large-scale unperturbed magnetic field, and $C(f)$ represents scattering by small scale fluctuations in the magnetic field. If the distribution function is assumed to consist of an isotropic part f_0 and a small diffusive drift $\mathbf{f}_1$ relative to the background, $f = f_0 + \mathbf{f}_1.\mathbf{p}/p$, the zeroth order isotropic moment of the Boltzmann or Vlasov–Fokker–Planck VFP equation is

$$u\frac{\partial f_0}{\partial x} + \frac{c}{3}\frac{\partial f_x}{\partial x} - \frac{1}{3}\frac{\partial u}{\partial x} p \frac{\partial f_0}{\partial p} = 0$$

where f_x is the x component of the vector $\mathbf{f}_1$. Integration in x from immediately downstream to immediately upstream of the shock yields

$$\frac{c}{3}(f_{xu} - f_{xd}) = \frac{u_u - u_d}{3} p \frac{\partial f_0}{\partial p}$$

where subscripts u and d refer to upstream and downstream values, respectively. Downstream of the shock CR advect with the background fluid, so $f_{xd} = 0$. CR cannot escape far upstream of the shock so the diffusive drift $f_{du}c/3$ must balance the advective flow $u_u f_0$, giving $f_{xu} = -3(u_u/c)f_0$, giving

$$\frac{p}{f_0}\frac{\partial f_0}{\partial p} = -\frac{u_u}{u_u - u_d} = -\frac{3r}{r-1}$$

since $u_u = u_s$ and $u_d = u_s/r$. For a strong shock $r = 4$ giving $f_0 \propto p^{-4}$ in momentum space which is equivalent to $n(E) \propto E^{-2}$ for relativistic energies as found above from single particle statistics.

Note that the details of the CR scattering process in the upstream and downstream plasmas appear in neither derivation. The key assumptions are that the shock velocity is much less than the CR velocity, the CR distribution is nearly isotropic at the shock, CR are unable to escape far upstream, and the CR density at the shock is equal to that far downstream. Because the number of assumptions is so small, the result of an E^{-2} spectrum for relativistic CR has very wide application. This predicted CR spectrum is a little flatter than the measured Galactic CR spectrum $n(E) \propto E^{-2.6}$. The difference in spectral index is discussed below, but when other factors are taken account of, the agreement between theory and observations is pleasingly close, and the theoretical result is robust and relatively model independent.

3. The maximum CR energy

The previous section demonstrated that shock acceleration gives a power law energy spectrum close to that of CR arriving at the Earth and of CR electrons responsible for synchrotron radiation in radio sources. For shock acceleration to be a credible theory it also has to explain how CR are accelerated to at least 1 PeV in the Galaxy and to 100's of EeV beyond the Galaxy.

In the case of SNR the maximum energy is probably determined by a balance between the acceleration rate and the lifetime of the remnant, as shown by Lagage and Cesarsky [67,68]. Here we present a 'back of the envelope' derivation of their result. We suppose that CR in a small energy range with number density n_{cr} upstream of the shock have a spatially uniform diffusion coefficient D. The balance between the diffusive flux $-D\partial n_{cr}/\partial x$ and the advective flux un_{cr}, produces an exponential precursor ahead of the shock: $n_{cr} = n_s \exp(-x/L)$ where x is the distance ahead of the shock and $L = D/u_s$ is the precursor scaleheight. n_s is the value of n_{cr} at the shock. The total number of CR in the precursor is then $n_s L$. The rate at which CR cross the shock from upstream to downstream and from downstream to upstream is $n_s c/4$, so the average time a CR spends upstream between shock crossings is $\Delta t = 4L/c = 4D/cu_s$. The average fractional energy gain on each shock crossing is $\Delta E/E = u_s/c$ as shown above. The acceleration rate is then $dE/dt = \Delta E/\Delta t = Eu_s^2/4D$. The characteristic acceleration time to energy E is $\tau_{accel} = E/(dE/dt) = 4D/u_s^2$. This ignores the time spent downstream. Lagage and Cesarsky include the downstream dwell time to show that $\tau_{accel} = 4D_u/u_s^2 + 4D_d/u_d^2$. The downstream diffusion coefficient D_d is probably much smaller than the upstream coefficient D_u because the downstream magnetic field is larger due to compression in the shock and probably highly turbulent. So the inclusion of the downstream dwell time probably does not increase the acceleration time τ_{accel} by more than a factor of 2 and we proceed by assuming that $\tau_{accel} = 8D_u/u_s^2$.

Assuming that CR are accelerated by SNR, a maximum CR energy can be estimated by setting the acceleration time equal to the age τ of the SNR: $8D/u_s^2 = \tau$. Writing the diffusion coefficient as $D = c\lambda/3$, where λ is the CR scattering mean free path, the acceleration time is equal to the SNR lifetime when $\lambda = 3u_s^2\tau/8c$. Consequently acceleration is only possible for CR with mean free paths less than $\lambda = 4 \times 10^{15} u_7^2 \tau_{1000}$ m, where u_7 is the SNR shock velocity in units of 10,000 km s^{-1}, and τ_{1000} is the SNR age in units of 1000 years. Higher energy CR have longer mean free paths because they are less easily deflected by magnetic field, and this sets an upper limit on the CR energy. The mean free path λ cannot be smaller than r_g, so acceleration to an energy E_{PeV} requires a magnetic field of at least $8u_7^{-2}\tau_{1000}^{-1}E_{PeV}$ μG, where the Larmor radius of a CR with energy E_{PeV} in units of PeV is $r_g = 3 \times 10^{16} E_{PeV} B_{\mu G}^{-1}$ m.

In the case of a typical historical SNR such as Tycho, Kepler or Cas A ($u_7 \sim 0.5$ & $\tau_{1000} \sim 0.4$) CR acceleration to the knee at a few PeV is only possible if the SNR shock propagates into a magnetic field greater than 100 μG, which is much larger than a typical interstellar magnetic field of a few μG. In this approximate model, the maximum CR energy is

$$E_{max} \approx 10^{14}(\lambda/r_g)^{-1} B_{\mu G}\tau_{1000}u_7^2 \text{ eV} \quad (2)$$

The most favourable assumption is that $\lambda/r_g \sim 1$ in which case the historical SNR expanding into a magnetic field of 3 μG could accelerate CR to only $\sim 3 \times 10^{13}$ eV, which is far short of the energy of the knee.

Hillas [55] proposed the parameter uBR as a measure of the maximum attainable CR energy in eV in a wide range of acceleration scenarios, where u, B and R are the characteristic velocity, magnetic field and length scale, respectively. The maximum CR energy derived by Lagage and Cesarsky is consistent with the Hillas argument in the case of Bohm diffusion $\lambda = p/eB$. Equating the acceleration time $8\lambda c/3u_s^2$ to the SNR lifetime R/u_s gives a maximum CR energy $E_{max} = (3/8)u_s BR$ in eV as in Eq. (2).

A crucial question is whether the maximum CR energy is limited by space or time. A spatial limit on the energy is given by the condition that the precursor scaleheight D/u_s must be less than the SNR radius R. The timescale limit of Lagage and Cesarsky is given by the condition that the acceleration time $8D/u_s^2$, or $4D/u_s^2$ if the CR downstream dwell time is negligible, must be less than the SNR lifetime R/u_s. The two conditions only differ by a numerical factor of 4 or 8, but the difference has important consequences. In the timescale limit, the precursor scaleheight L of the highest energy CR is smaller than the SNR radius R by 4–8 in which case CR

do not escape upstream from the shock into the interstellar medium. This is because CR with energy E_{max} typically diffuse only a distance L into the precursor before being overtaken by the advancing shock. CR with an energy much less than E_{max} stand even less chance of escaping upstream since their precursor scaleheight is correspondingly smaller. In contrast, if the spatial limit were to apply, CR with energy E_{max} would by definition escape into the interstellar medium. Since the timescale limit is the more stringent, it appears that CR may be unable to escape upstream, and this raises the question of how the Galactic CR population can be replenished by SNR. This important issue is given further consideration below. As we shall see, the assumptions that CR transport is diffusive and that the mean free path is proportional to the Larmor radius are open to question. Breakdown of the diffusion approximation or the inability of CR to generate magnetic turbulence may result in escape upstream at high CR energies.

Acceleration to the knee is only possible if the magnetic field exceeds typical interstellar values. Unless the magnetic field greatly exceeds 100 μG, acceleration to PeV in SNR requires a CR mean free path not much larger than the Larmor radius. The case of $\lambda/r_g \sim 1$, known as 'Bohm' diffusion, is only achieved if the magnetic field is disordered on the scale of a CR Larmor radius. Recent theories of magnetic field generation suggest that the magnetic field is in fact suitably tangled (see below) but even with the assumption of Bohm diffusion it is clear that acceleration to the knee in the historical SNR is only possible if the magnetic field is strongly amplified above typical interstellar values.

The acceleration rate is increased if the magnetic field lies in the plane of the shock surface (known as a 'perpendicular shock'), but in this case the maximum CR energy is limited by spatial constraints to a very similar value and the maximum CR energy is very similar to that in Eq. (2). This will be discussed in Section 6.

Older remnants in the Sedov blast wave phase have larger values of τ_{1000} but the shock velocity decreases with age, $u_s \propto \tau^{-3/5}$, so the maximum CR energy actually decreases with time according to the above formula: $E_{max} \propto \tau^{-1/5}$. The energy of an individual CR always increases with time if it stays with the shock, so a CR has the greatest opportunity to gain energy if it is injected into the acceleration process during the free expansion phase and stays with the shock front into and throughout the Sedov phase. However, a CR can only stay with the shock through the Sedov phase if the magnetic field is strong enough to confine it. We have shown above that fields much greater than typical interstellar fields are needed and these may not be available during the slow expansion of the Sedov phase, in which case the maximum CR energy may only be reached during the pre-Sedov expansion. The estimates of the maximum CR energy made by Lagage and Cesarsky presented a serious obstacle to an explanation of Galactic CR acceleration until it was understood that CR themselves can amplify a seed interstellar magnetic field to ~ 100 μG as part of the acceleration process (see below), thereby self-consistently generating the magnetic field needed for acceleration to PeV energies.

The above estimate of the maximum CR energy applies to CR protons. The maximum energy for electron acceleration is usually limited by synchrotron losses. Acceleration is terminated when the acceleration time is equal to the energy loss time due to synchrotron cooling. The synchrotron cooling time is $\tau_{sync} = 4 \times 10^{11} E_{\rm PeV}^{-1} B_{\mu\rm G}^{-2}$ s [71]. In comparison, the acceleration time, assumed to be $8D/u_s^2$, is $\tau_{accel} = 2.7 \times 10^{11} (\lambda/r_g) E_{\rm PeV} B_{\mu\rm G}^{-1} u_7^{-2}$s. The cooling time decreases as an electron gains energy and acceleration terminates when $\tau_{sync} = \tau_{accel}$ which limits the CR electron energy to $E_{\rm PeV} = 1.2(\lambda/r_g)^{-1/2} B_{\mu\rm G}^{-1/2} u_7$ resulting in a corresponding turnover in the synchrotron spectrum. The characteristic energy of synchrotron photons emitted by electrons of energy $E_{\rm PeV}$ in a magnetic field $B_{\mu\rm G}$ is $h\nu \sim 50 B_{\mu\rm G} E_{\rm PeV}^2$ keV. The photon energy is proportional to $B_{\mu\rm G} E_{\rm PeV}^2$ and the maximum CR energy is proportional to $B_{\mu\rm G}^{-1/2}$ (see above), so the turnover of the synchrotron spectrum is independent of the magnetic field: $h\nu \sim 70 u_7^2 (\lambda/r_g)^{-1}$ keV. Since u_7 and the synchrotron spectrum turnover are both measurable for many SNR, this relation can be used to estimate the parameter λ/r_g. However, the expression for $h\nu$ is uncertain to a numerical factor because the synchrotron emission is not confined to a single frequency and the acceleration time depends on the details of the upstream and downstream diffusion coefficients (see for example [2]). Stage et al. [105] arrive at the equation $h\nu_c \sim 30 u_7^2 (\lambda/r_g)^{-1}$ keV where ν_c is the frequency at which the synchrotron emission cuts off. They applied this formula to observations of Cas A with the welcome conclusion that $\lambda/r_g \sim 1$. Their results increase confidence in the assumption that Bohm diffusion applies during shock acceleration. A similar conclusion is reached by Uchiyama [106].

The synchrotron cooling time can also be used to estimate the magnetic field at SNR shocks [109,22,112]. Filaments of X-ray synchrotron emission are observed at the shock. If the emission extends a distance x downstream of the shock, the synchrotron cooling time is $\sim 4x/u_s$ if CR are advected away from the shock at the post-shock fluid velocity $u_s/4$.

4. Magnetic field generation

As discussed above, the outer shocks of SNR can only accelerate CR to a few PeV if the shock propagates into a magnetic field much larger than a typical interstellar field. Fortunately, plasma streaming instabilities provide a natural way for the CR precursor to amplify the magnetic field in the upstream plasma. The instabilities set the upstream plasma into turbulent motion. This stretches the magnetic field lines which are frozen into the plasma, producing a tangled magnetic field with a magnitude much larger than the initial seed field. Further field enhancement can take place downstream of the shock either as a result of the vorticity generated by the upstream instability or because the shock overtakes an already clumpy medium [51,119]. As mentioned above, the presence of large magnetic field at shocks is confirmed by X-ray observations of thin lines of synchrotron emission coincident with the outer shocks of young SNR. Analysis of young SNR indicates fields in the range 100 μG to mG [22,108], which matches the field required to accelerate CR to PeV. Here we describe instabilities that directly generate magnetic field but other instabilities in the precursor [34] may play a role in field generation through turbulence [73].

4.1. The Weibel instability

It has been known for decades that anisotropic fluxes of charged particles are unstable to the exponential growth of small perturbations. There are a number of a streaming instabilities, but one of the most important in this context is the Weibel instability [116]. The instability occurs in various forms, but it applies particularly to shocks and cosmic rays in the form of a filamentation instability. The Weibel instability separates a spatially uniform drifting population of charged particles into current filaments. Each filament carries an excess electric current surrounded by a corresponding magnetic field that further focuses current into the filament, causing an increase in the local current and a further increase in the magnetic field in a positive feedback loop. In its simplest form the instability grows most rapidly on the scale of an electron skin depth c/ω_{pe} or the ion skin depth c/ω_{pi}. In a plasma density of one particle per cm^3 the electron skin depth is only 5 km and the ion skin depth is 200 km so the Weibel instability can be effective in providing the internal dissipation needed to sustain a collisionless shock, and it can provide the turbulence needed to

scatter low energy CR with mildly suprathermal energies, particularly in relativistic shocks [102–104,69]. However it is unlikely that it can generate the large scale magnetic fields needed to scatter high energy CR with Larmor radii many orders of magnitude larger than c/ω_{pe}.

4.2. The Alfven instability

In order to accelerate a charged particle even to a GeV, magnetic field is needed on the scale of at least 10^5 km corresponding to the CR Larmor radius in a 300 μG magnetic field. This scale is very much larger than c/ω_{pe} and also very much larger than the Larmor radius of thermal particles which are therefore strongly magnetized and fixed to magnetic field lines on this scale. Hence for high energy CR to be confined near the shock during acceleration the confining magnetic field must be generated by a process in which the background plasma behaves as an MHD fluid. The Alfven instability meets this requirement. The Alfven instability results in the exponential growth of MHD Alfven waves with wavelengths resonantly matched to the Larmor radius of the CR which stream through a background MHD plasma [66,117]. As cosmic rays propagate upstream of the shock they resonantly excite Alfven waves and thereby generate fluctuations in the magnetic field. Because the Alfven waves are excited with a wavelength comparable with the CR Larmor radius they are ideally suited to scatter the CR which then execute a diffusive random walk in the shock environment and are accelerated as described in Section 2.

Self-consistent equations for CR transport and Alfven wave generation were set out by Skilling [99–101]. The realisation that the Alfven instability could result in strong CR scattering opened the way for the development of the theory of diffusive shock acceleration, but it did not show how the magnetic field could be amplified to the large magnitude needed to accelerate CR to a few PeV. It was commonly supposed that the Alfven instability would saturate when the fluctuating part of the field $\delta\mathbf{B}$ becomes comparable with the unperturbed field $\mathbf{B}$ present in the upstream plasma. It seems intuitively reasonable that unstable growth ceases once $\delta\mathbf{B} \sim \mathbf{B}$ and the spatial resonance between the wavelength and CR gyration is lost, but this is not the case.

4.3. The non-resonant hybrid (NRH) instability

The presumption that CR streaming instabilities saturate at $\delta\mathbf{B}/\mathbf{B} \sim 1$ was tested by Lucek and Bell [72] who simulated CR as particles flowing along magnetic field lines frozen into a fluid modelled with an MHD code. They found that in fact the field grows far beyond $\delta\mathbf{B}/\mathbf{B} \sim 1$ as shown in Fig. 1. Bell [17,18] then showed that the strongest field amplification results not from the Alfven instability but from the non-linear growth of a previously unknown non-resonant hybrid (NRH) instability. The instability is non-resonant because it does not require a match between the CR Larmor radius and the instability wavelength, and it is a hybrid instability because, like the Alfven instability, it results from a mixture of kinetic and hydrodynamic behaviour. The thermal particles behave magnetohydrodynamically since they are strongly magnetized, but CR behave kinetically since their Larmor radii are at least as large as the instability wavelength. This contrasts with the Weibel instability in which all particles behave kinetically. The NRH instability does not saturate at $\delta\mathbf{B}/\mathbf{B} \sim 1$, but continues growing non-linearly to produce fields much larger than the field in which the instability begins to grow.

The features of the NRH instability are easily understood by deriving its growth rate in its simplest configuration. Consider instabilities operating on scales much smaller than the CR Larmor radius. In this case CR trajectories are negligibly deflected by the perturbed magnetic field and we can suppose that an electric current $\mathbf{j}_{cr}$ carried by the CR, consisting mainly of protons, is uniform in space, unvarying in time, and directed in the z direction. We consider the case in which the background plasma is cold in the sense that forces due to the pressure gradient can be neglected in comparison with the $\mathbf{j}_{cr} \times \mathbf{B}$ force exerted by the cosmic rays, and that the magnetic forces $\mathbf{B} \times (\nabla \times \mathbf{B})/\mu_0$ are similarly negligible. To preserve quasi-neutrality the CR current must be balanced by an electric current $\mathbf{j}_t$ carried by the background thermal plasma. The instability evolves on a fluid rather than an electromagnetic timescale so the displacement current can be neglected and the CR and thermal currents are related by $\nabla \times \mathbf{B} = \mu_0(\mathbf{j}_{cr} + \mathbf{j}_t)$. $\nabla \times \mathbf{B}$ can also be neglected as a consequence of the assumption that the magnetic force is small, leaving $\mathbf{j}_t = -\mathbf{j}_{cr}$ in this approximation, that is, the thermal current $\mathbf{j}_t$ locally balances the CR current $\mathbf{j}_{cr}$. The momentum equation for the background plasma moving with velocity $\mathbf{v}$ is $\rho \mathrm{d}\mathbf{v}/\mathrm{d}t = \mathbf{j}_t \times \mathbf{B} = -\mathbf{j}_{cr} \times \mathbf{B}$. The equations controlling the instability are then

$$\rho\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\mathbf{j}_{cr} \times \mathbf{B} \quad \frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{v} \times \mathbf{B}) \quad \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho\nabla.\mathbf{v} \tag{3}$$

After some manipulation, these equations reduce to

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\mathbf{j}_{cr} \times \left(\frac{\mathbf{B}}{\rho}\right) \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{B}}{\rho}\right) = \left(\frac{\mathbf{B}}{\rho}\right).\nabla\mathbf{v}$$

The equations are transverse and linear without further approximation if (i) CR stream along the unperturbed field lines in the z direction of the initial field (ii) all quantities vary only in the direction parallel to $\mathbf{j}_{cr}$ and the initial unperturbed field $\mathbf{B}_0$. The equations then further reduce to

$$B_z = B_0 \quad \rho = \rho_0$$

$$\frac{\partial \mathbf{v}}{\partial t} = -\mathbf{j}_{cr} \times \left(\frac{\mathbf{B}}{\rho}\right) \quad \frac{\partial}{\partial t}\left(\frac{\mathbf{B}}{\rho}\right) = \left(\frac{B_0}{\rho_0}\right)\frac{\partial \mathbf{v}}{\partial z}$$

These linear equations in $\mathbf{v}$ and $\mathbf{B}/\rho$ have been derived without making the linear assumption that second order quantities can be neglected. They are correct both linearly and non-linearly in their natural configuration. This is why the instability does not saturate at $\delta\mathbf{B}/\mathbf{B} \sim 1$ but continues growing to produce amplified magnetic field much larger than the initial field.

The equations have an exponentially growing circularly polarised solution. The magnetic field lines take the form of spirals with constant wavelength $(2\pi/k)$ along the spiral and exponentially growing amplitude. The growth rate is

$$\gamma = \left(\frac{k j_{cr} B_0}{\rho_0}\right)^{1/2} \tag{4}$$

The plane wave assumption that quantities only vary in the z direction parallel to the CR current is incorrect in realistic cases, but the spirals grow most rapidly with their axes aligned with the large scale magnetic field for which the difference between $\mathrm{d}/\mathrm{d}t$ and $\partial/\partial t$ is small. In reality, spatially separated spirals with different wavelengths start growing independently. A spiral growing in one part of the plasma eventually expands into another spiral growing about a separate axis. Exponential growth of interacting spirals is slowed as they run into each other, but non-linear simulations [17] show that the spirals expand further either by coalescing into a single larger spiral or by the larger spiral pushing aside the smaller spiral. In the non-linear configuration the field lines evolve into wandering spirals as in Fig. 1.

As a spiral expands it evacuates a low density cavity on axis since the plasma is frozen to field lines. The instability grows non-linearly into a structure consisting of cavities of low density and low magnetic field surrounded by walls of large magnetic field accompanied by high density as shown in Figs. 2 and 3. Similar re-

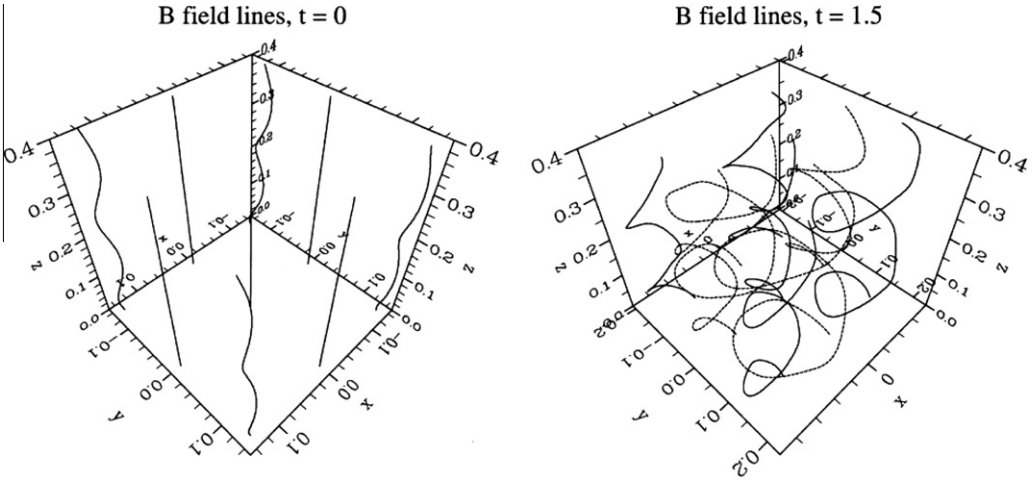

Fig. 1. Results of a simulation by Lucek and Bell [72] of magnetic field amplification driven by streaming cosmic rays. The cosmic rays are treated as charged particles initially streaming in the vertical direction. The left hand plot shows the initial magnetic field which is frozen into a three-dimensional fluid simulated by ideal MHD and self-consistently coupled to the cosmic rays through the CR current. The right hand plot shows how the field lines are stretched into spirals at a later time.

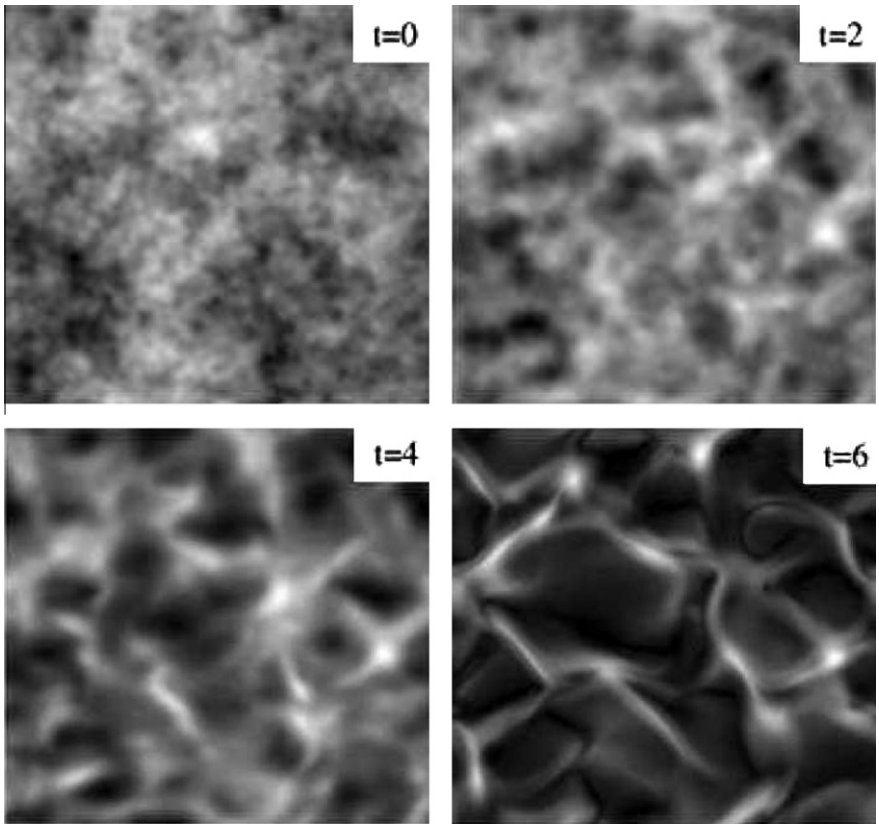

Fig. 2. Magnitude of magnetic field generated by the non-resonant hybrid instability. Grey-scale plots of slices in x, y of a three-dimensional MHD simulation of magnetic amplification driven by a fixed uniform CR current in the z direction. Reproduced from Bell [17] where further details can be found. The grey-scale minima (black) and maxima (white) at each time as bracketed pairs (minimum,maximum) are (0.81,1.22) at $t = 0$, (0.69,1.35) at $t = 2$, (0.40,2.30) at $t = 4$ and (0.20,12.01) at $t = 6$. The maximum growth rate is $\gamma_{max} = 1.26$ in these units. The plots show how initially small scale structure grows into a wall-cavity structure corresponding to the spiral magnetic field lines seen in Fig. 1. The time and distance scales are different in the separate calculations displayed in Fig. 1.

sults have been found in MHD simulations by Reville et al. [86], Zirakashvili and Ptuskin [119], and Zirakshvili et al. [120]. Reville et al. [86] further show that test particles inserted into the calculated magnetic field have their diffusion inhibited by the amplified field. Strong field amplification is seen in extensive particle-in-cell simulations (PIC) by Riquelme and Spitkovsky [90,91]. PIC simulations by Ohira [79] also produce strong field amplification. In contrast Niemiec et al. [78] find that the instability stops growing at $\delta B/B \sim 1$ because of the back-reaction onto the CR current in their PIC simulations.

Because $\gamma \propto k^{1/2}$ small scale structures grow most rapidly as apparent in Fig. 2. The upper limit on k is determined by tension in the magnetic field lines which was neglected in Eq. 3. Including magnetic forces gives

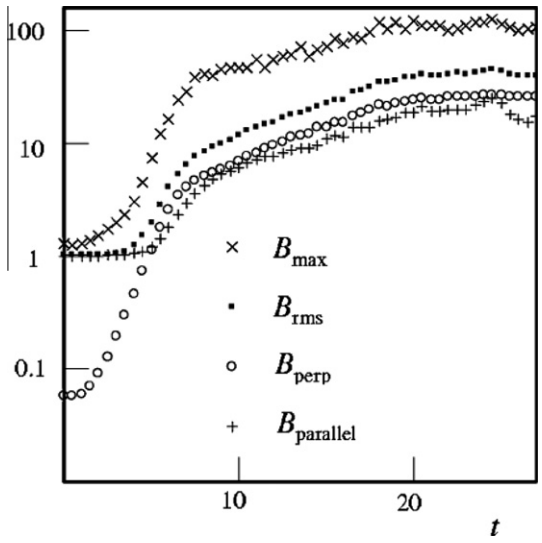

Fig. 3. Magnetic field amplification due to the non-resonant hybrid instability. The plots show the maximum magnetic field, the root mean square of the total magnetic field and magnitudes of the components perpendicular and parallel to the CR current. The maximum growth rate is $\gamma_{max} = 1.26$ in these units. Reproduced from Bell [17].

$$\rho \frac{d\mathbf{v}}{dt} = -\mathbf{j}_{cr} \times \mathbf{B} - \mathbf{B} \times (\nabla \times \mathbf{B})/\mu_0$$

The first term on the right hand side drives the instability, but the second term acts towards stability because it opposes the stretching and bending of magnetic field lines. The ratio of the two terms on the right hand side is the ratio of the CR current $\mathbf{j}_{cr}$ to $\nabla \times \mathbf{B}/\mu_0$. The second term overpowers the first at short wavelengths $k > k_0$ where $k_0 = \mu_0 j_{cr}/B_0$, and this determines the shortest unstable wavelength. Since short wavelengths grow most rapidly, the shortest unstable wavelength is also the fastest growing. More precisely, the maximum growth rate of the NRH instability is $\gamma_{max} = (\mu_0/\rho)^{1/2} j_{cr}/2$, and the fastest growing mode has a wavenumber $k_{max} = k_0/2$. In the units used in Figs. 2 and 3 the maximum growth rate is $\gamma_{max} = 1.26$.

The maximum linear growth rate is independent of the magnitude of the zeroth order magnetic field. It depends only on the CR current density j_{cr} and the background fluid density ρ. In order to find the maximum growth rate of the NRH instability we need to estimate the CR current density $\mathbf{j}_{cr}$, and this is determined by the CR flux needed to carry the CR energy density into the precursor ahead of the shock. If the CR energy density at the shock is $\eta\rho_0 u_s^2$ where η is an efficiency factor in the production of CR, then the CR energy flux in the upstream rest frame is $\eta\rho_0 u_s^3$. For simplicity, consider a monoenergetic CR distribution with CR energy pc and carrying an electric current $\mathbf{j}_{cr}$. The CR number flux is $\mathbf{j}_{cr}/e$, assuming the CR are protons, and the CR energy flux is $pc\mathbf{j}_{cr}/e$. Equating $\eta\rho_0 u_s^3$ and $pc\mathbf{j}_{cr}/e$ gives

$$\mathbf{j}_{cr} = \frac{\eta e \rho_0 u_s^3}{pc} \tag{5}$$

and a maximum growth rate

$$\gamma_{max} = (\mu_0\rho_0)^{1/2} \frac{\eta u_s^3}{2E} \tag{6}$$

where E is the CR energy in eV. Immediately upstream of the shock most of the CR current is carried by GeV protons so it is appropriate to set $E = 10^9$ eV, but GeV protons do not penetrate far upstream of the shock and cannot generate magnetic field capable of confining TeV or PeV protons. The magnetic field required to confine PeV protons can only be generated by the PeV protons themselves since only protons with this energy penetrate sufficiently far upstream. Hence the instability growth rate relevant to PeV protons is best estimated by setting $E = 10^{15}$ eV. η is not easily estimated, but $\eta = 10^{-2}$ represents 1% of the shock energy being given to PeV protons [17]. With these assumptions, an upstream plasma density of 1 cm^{-3} and a shock velocity of 10,000 km s^{-1}, the growth time γ_{max}^{-1} of the fastest growing mode is 120 years. $\eta = 10^{-2}$ is a relatively conservative assumption, but on the other hand the historical SNR such as Cas A, Tycho and Kepler expand typically at only 5000 km s^{-1}. It appears that the NRH instability amplifies the magnetic field sufficiently to explain acceleration to the knee, but not with a large margin to spare. The growth rate of the NRH instability appears not to be sufficient for acceleration beyond the knee in the historical SNR. Since $\gamma_{max} \propto u_s^3$, acceleration to PeV by slowly expanding SNR in the Sedov phase appears unlikely. Acceleration in the Sedov phase must rely upon the interstellar magnetic field ($\sim$3 μG) into which SNR expand and this is not sufficient for CR to reach 1PeV. For PeV acceleration we need to look to relatively young SNR, preferably expanding at high velocity into a dense medium.

The NRH instability grows most rapidly on scales smaller than the CR Larmor radius. With the assumptions made in the previous paragraph and a seed magnetic field of 3 μG, PeV protons generate magnetic field that initially grows on a scale k_{max}^{-1} of $\sim 2 \times 10^{13}$ m which is smaller than the CR Larmor radius of $\sim 3 \times 10^{14}$ m in a magnetic field of 100 μG. Fortunately, Fig. 2 shows that the scale size grows rapidly during the non-linear phase. The characteristic scale at $t = 6$ ($\gamma_{max} t = 7.6$) is an order of magnitude greater than that at $t = 2$ ($\gamma_{max} t = 2.5$.) Plots of magnetic field at $t = 10$ and $t = 20$ in Bell [17] show that the characteristic scale continues to grow although the relevance of the plots is questionable because the radii of the spiralling magnetic fields approach the size of the computational box by these times. The NRH instability amplifies the magnetic field by stretching field lines to ever larger scales so the mismatch in scale between the most rapidly growing wavelength and the CR Larmor radius may not be a problem except maybe under extreme circumstances, but further work is needed to investigate this.

4.4. Long scalelength instabilities and filamentation

Bohm diffusion and effective CR confinement near the shock is only possible if magnetic structures are generated on scales comparable with or greater than the CR Larmor radius. As remarked in the previous section, this may happen naturally as the wall-cavity structure expands during non-linear growth. Here we discuss two alternative pathways to large scale structure: (i) linearly unstable growth from noise of large scale perturbations (ii) filamentation of the CR currents.

The growth rate of the linear NRH instability falls away rapidly at wavelengths longer than the CR Larmor radius. At long wavelengths CR trajectories follow the local magnetic field line in the absence of scattering. The $\mathbf{j}_{cr} \times \mathbf{B}$ force then tends to zero because the CR current is parallel to the local magnetic field. However, the situation changes once it is recognised that CR are scattered by rapidly growing instabilities on scales less than the Larmor radius. Scattering frees the CR from spiralling purely along the large scale field lines. In the limit of strong scattering the CR current is aligned with the CR pressure gradient instead of the magnetic field. Since CR are now able to diffuse across the field the $\mathbf{j} \times \mathbf{B}$ force is non-negligible and the instability returns on scales larger than the CR Larmor radius [97]. The analysis of this problem is complicated because it is non-linear and turbulence is important on scales both greater and less than the CR Larmor radius. Bykov et al. [28] use a mean-field turbulence approach to show that the long wavelength instability grows more rapidly than the Alfven or NRH instability at the same long wavelength. However, the model chosen by

Schure and Bell [97] does not replicate the rapid growth. This is a challenging and important issue since the structure of the magnetic field on scales greater than the CR Larmor radius probably governs the overall CR confinement near to, or escape from, the shock.

Large scale structures may also grow as a result of a filamentation instability that 'piggy-backs' on the NRH instability [88]. The NRH wall-cavity structure evolves in line with the expansion of spirals in the magnetic field. The $\mathbf{j_{cr}} \times \mathbf{B}$ force expands the spirals, but by momentum conservation it also deflects the CR trajectories onto the axis of the spiral. This suggests that a fully self-consistent simulation of CR trajectories and magnetic field amplification might produce a non-linear configuration in which CR propagate as filaments through cavities. This has received provisional confirmation from two-dimensional self-consistent simulations [88] that demonstrate filament formation on scales greater than that of the wall-cavity structure generated by the NRH instability. These simulations of planar slices perpendicular to the zeroth order CR current assume a geometry in which filaments are infinite and uniform in extent along the direction of CR propagation.

Filaments might either increase or reduce CR confinement. On the one hand, the concentration of the CR current along the axis of the filament could enhance the $\mathbf{j_{cr}} \times \mathbf{B}$ force and increase the rate at which magnetic field is amplified, thereby increasing CR confinement. On the other hand, the filaments could provide a pathway for CR to escape upstream of the shock by propagation through low-field cavities. Three-dimensional self-consistent calculations are needed to resolve this issue.

4.5. A distinction between the Alfven and NRH instabilities

The linear growth rate of the non-resonant hybrid (NRH) instability is greatest at wavelengths less than the CR Larmor radius, but it also grows at wavelengths comparable with the Larmor radius, $kr_g \sim 1$. The Alfven and NRH instabilities have similar growth rates at $kr_g \sim 1$ where the Alfven instability has its largest growth rate, but the two instabilities are distinguished by their polarisations. In the Alfven instability a CR proton gyrates in the same direction as the magnetic field spiral so that streaming CR remain in phase with the wave as they propagate. In contrast, in the NRH instability a CR proton gyrates in the opposite direction.

A related distinction lies in the $\mathbf{j}_{cr} \times \mathbf{B}$ that drives the instabilities in the linear limit. The Alfven instability is driven by the vector product of the perturbed CR current and the unperturbed zeroth order magnetic field. In contrast, the NRH instability is driven by the vector product of the unperturbed CR current and the perturbed magnetic field.

4.6. Saturation of the NRH instability

The NRH instability leading to magnetic field amplification is active provided there is a range of k-space between a lower limit $k_{min} = eB/p$ equal to the inverse CR Larmor radius and an upper limit $k_{max} = \mu_0 j_{cr}/B$ imposed by tension in the magnetic field as discussed above. The upper and lower limits in k-space coincide and close off the instability when $\mu_0 j_{cr}/B = eB/p$ which occurs when the magnetic energy density grows to $B^2/\mu_0 = p j_{cr}/e$. It was shown above that $\mathbf{j}_{cr} = \eta e \rho_0 u_s^3/pc$ and this imposes a limit on the magnetic energy density, $B^2/\mu_0 = \eta \rho_0 u_s^3/c$ [19]. If we assume that this limit applies when the NRH instability grows non-linearly as well as linearly, it gives an estimate of the magnetic field at which the instability reaches saturation. Allowing for magnetic field compression at the shock, this leads to a saturated magnetic field in surprisingly good agreement with SNR observations as analysed by Vink [108,110] who fits the data with

$$B \sim 700 (u_s/10^7 \text{ m s}^{-1})^{3/2} (n_e/\text{cm}^{-3})^{1/2} \ \mu\text{G} \quad (7)$$

4.7. Observations of precursor microphysics

The Larmor radius of a CR with energy E_{PeV} in PeV in a magnetic field B_{100} in units of 100 μG is $r_g = 3 \times 10^{14} E_{\text{PeV}} B_{100}^{-1}$ m. At a distance from the Earth of R_{kpc} kpc, the Larmor radius subtends an angle on the sky of $2 E_{\text{PeV}} B_{100}^{-1} R_{\text{kpc}}^{-1}$ arc s suggesting that it might be possible to resolve the turbulence generated by CR streaming in nearby SNR.

Eriksen et al. [44] have suggested that the separation of stripes seen in CHANDRA observations of Tycho's SNR correlates with the Larmor radius of high energy CR. Alternative interpretations of the stripe separation have been proposed by Bykov et al. [27], where the stripes correspond to peaks in magnetic turbulence and by Malkov et al. [75], where the stripes develop as solitons in the upstream plasma.

Time-variation offers a different observational diagnostic of the microphysics underlying shock acceleration as shown by Uchiyama et al. [106] in RXJ1713.6–3946 and Patnaude and Fesen [81] in Cas A. The expanding shock maps out a 2D surface advancing into the magnetic field. It takes one year for a shock expanding at 10,000 km s^{-1} to propagate a distance of 3×10^{14} m characteristic of the Larmor radius of a PeV proton. This would be consistent with variations over a few years observed by Uchiyama et al. and by Patnaude and Fesen.

The precursor ahead of the shock is $\sim c/u_s$ times larger than the CR Larmor radius but detection is challenging. Rakowski et al. [84] have interpreted the X-ray emission from protuberances ahead of the shock in SN1006 as possibly resulting from the interaction of the upstream NRH instability with the Rayleigh–Taylor instability. Spectral evidence for a precursor is seen by Raymond et al. [85] in Hα lines at SNR shocks where a narrow spectral feature is interpreted as emission from low temperature upstream hydrogen excited by a CR precursor.

5. Non-linear self-regulation

Diffusive shock acceleration has to be very efficient if it is to generate the Galactic CR population, but the total CR pressure and energy density at a shock cannot exceed ρu_s^2. The injection of CR into the acceleration process must be finely balanced to ensure that acceleration is efficient but not impossibly efficient. Barring fortunate coincidences, a feedback process is required. For a strong shock the test particle CR distribution in energy E is proportional to E^{-1} for non-relativistic CR and E^{-2} for relativistic CR, so CR protons above 1 GeV contribute nearly all the CR energy density. For effective feedback, the pressure of >GeV protons must regulate the injection of protons at mildly suprathermal energies which are of the order of an MeV for SNR shocks [38]. Feedback is also evident in weaker shocks found in the heliosphere (e.g. [43]). The essential features of feedback processes at strong shocks were set out by Drury and Völk [35] and Völk et al. [114] who showed that self-regulation can operate through the CR pressure modifying the flow ahead of the shock.

Fig. 4, reproduced from Drury and Völk [35], illustrates the effect of CR pressure on the shock structure. The CR pressure precursor ahead of the shock is depicted in the right hand plots. If injection is strong the CR pressure P_{cr} is comparable with ρu_s^2 and the CR pressure gradient ∇P_{cr} pushes on the background plasma ahead of the shock. Or when considered in the shock rest frame, the CR pressure slows the plasma as it flows into the shock as shown in the left hand plot. The magnitude of the velocity jump at the shock discontinuity is then reduced, less energy is processed in the transition, and the post-shock temperature is reduced. Consequently the energy at which CR are injected is reduced and fewer

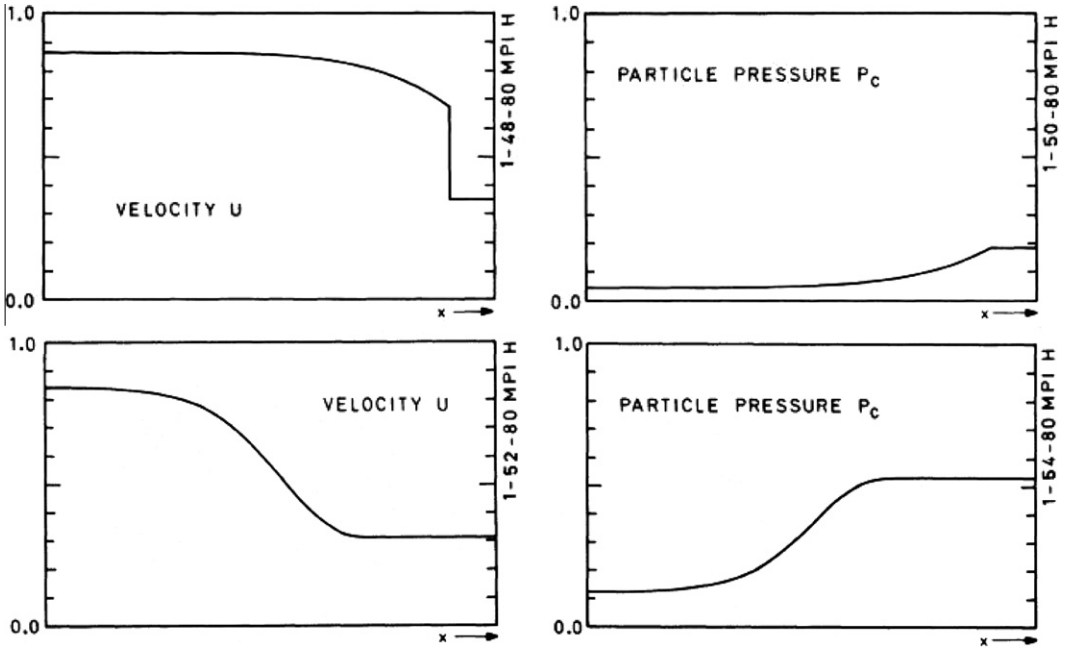

Fig. 4. Non-linear feedback. The shock structure modified by the CR pressure. Top row: moderate CR pressure with a sharp shock transition. Bottom row: high CR pressure with a smooth shock. Reproduced from Figs. 2 and 3 of Drury and Völk [35].

CR are accelerated to relativistic energies, thereby reducing the CR pressure at the shock and completing the feedback loop. Additionally a further reduction in acceleration efficiency follows if the upstream plasma is not completely cold. If the CR pressure reduces the shock velocity to a magnitude not much larger than the upstream sound speed, the Mach number is decreased and reduces the compression ratio at the shock, thereby steepening the spectrum such that fewer CR are accelerated to high energy. In extreme cases of high acceleration efficiency the discontinuous sub-shock may disappear so that the upstream and downstream states are connected by a smooth transition as in the lower row of plots in Fig. 4. However, any smooth transition would occur on the scale of the precursor scaleheight of mildly relativistic protons which is too small to be observable in SNR.

The self-consistent interaction between CR acceleration, injection and hydrodynamic modification was modelled in detail by Ellison and Eichler [42], Bell [16], Falle and Giddings [45] and Blasi et al. [26]. As expected, the feedback processes limit injection and

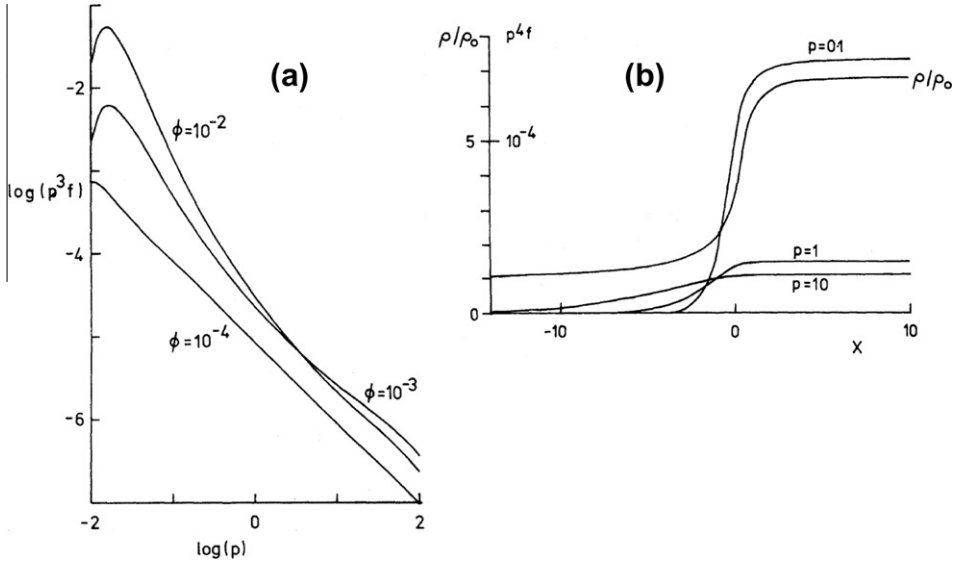

Fig. 5. Non-linear feedback. The CR spectrum and its effect on shock structure, reproduced from Bell [16]. The spectrum is plotted in (a) for different fractions ϕ of CR injected into the acceleration process. For a low injection rate $\phi = 10^{-4}$, CR behave as test particles with no feedback onto the shock structure. As the injection rate is increased to $\phi = 10^{-3}$, and then $\phi = 10^{-2}$, the CR pressure in the precursor accelerates the upstream ahead of the shock and smooths out the hydrodynamic transition as shown in the density (ρ/ρ_0) plot in (b) for $\phi = 10^{-2}$. At high injection rates the spectrum has a strong concave curvature. Plot (b) also shows the CR spatial profiles at different momenta p. In this calculation it is assumed that the diffusion coefficient D, and therefore the characteristic precursor scaleheight, is proportional to $p^{1/2}$. $D \propto p$ might be more realistic but the weaker dependence was assumed for computational convenience.

proton acceleration to GeV energies. Fewer CR reach high energy, but those that do so experience a compression greater than 4 which produces a high energy spectrum flatter than p^{-4}. A compression factor up to 7 is possible for a CR-dominated shock because of the smaller relativistic ratio of specific heats. Overall, the spectrum steepens at low energy and flattens at high energy to produce a characteristic concave spectrum. The non-linear response to different injection rates as calculated by Bell [16] is shown in Fig. 5. As the injection rate ϕ is increased the spectrum becomes increasingly concave. In Fig. 5a this is mainly seen as a steepening in the low energy part of the CR spectrum. Fig. 5b shows how high energy CR with long mean free paths experience the full compression between upstream and downstream while low energy CR experience only part of the total compression. This is a complicated problem in which the self-consistent inclusion of magnetic field amplification and the consequent heating of the upstream plasma can be important [6].

Concavity in SNR emission spectra has been identified at radio [89] and X-ray [5,107] wavelengths. Non-linear self-regulation and efficient CR acceleration can also be deduced from observations of low post-shock temperatures [53,59]. Efficient energy transfer to CR reduces the energy available to heat the post-shock plasma. A concomitant consequence of a low post-shock temperature, the lower ratio of specific heats of relativistic CR, and possible energy loss to escaping CR is a compression factor greater than four at the shock. This affects the overall structure of the SNR, notably reducing the gap between the forward shock and the contact discontinuity. In some SNR the contact discontinuity and the reverse shock are closer to the forward shock than would be expected in the absence of efficient CR acceleration, thus providing further evidence for non-linear self-regulation and efficient acceleration [30,115,81,95]. None of this evidence is completely incontrovertible. For example, spectral concavity might result from spatial integration over different regions with differing spectral indices, low measured post-shock temperatures might apply to not all components of the downstream plasma, and the distance to the forward shock might be reduced by the protrusion of Rayleigh–Taylor fingers ahead of the contact discontinuity. However, the cumulative effect of the evidence, allied with the requirement that SNR must accelerate CR efficiently to explain the Galactic CR energy density, indicates that non-linear effects are significant in diffusive shock acceleration.

6. Perpendicular shocks

The theory of diffusive shock acceleration as set out in Sections 2 and 3 made no mention of the effect of a large scale magnetic field. The analysis was developed for the case of isotropic diffusion in which the CR diffusion coefficient is the same in all directions. If the plasma supports a large scale magnetic field the diffusion coefficients are in fact different along and across the magnetic field. The assumptions in Sections 2 and 3 are correct in the case of a parallel shock, that is, one in which the shock propagates parallel to the field. For a parallel shock, only diffusion along the field in the direction of the shock normal matters and it is irrelevant if CR are unable to diffuse in the plane of the shock. The theory of Sections 2 and 3 also works for shocks which are 'quasi-parallel'. For quasi-parallel shocks, CR may diffuse preferentially along magnetic field lines at an angle to the shock normal but the analysis is unaffected [14]. However, perpendicular and quasi-perpendicular shocks are different. The dividing line between parallel and perpendicular shocks occurs at a critical angle θ_c between the shock normal and the magnetic field satisfying $c\cos\theta_c = u_s$. At the angle θ_c CR are marginally able to escape upstream of the shock by travelling at the speed of light along a magnetic field line and acceleration is marginally possible without strong scattering. $\theta_c = 88^\circ$ for a shock velocity of 10,000 km s^{-1}, so most SNR shocks are in this sense quasi-parallel rather than quasi-perpendicular, but perpendicular shocks are an important case. Shocks encountering the field at angles greater than θ_c are often referred to as 'super-luminal' even if $u_s \ll c$ because CR would have to travel faster than the speed of light to escape upstream of the shock along a magnetic field line. In the next section we consider general oblique shocks and show that CR can be strongly affected by the angle of the shock even for angles much less than θ_c. Here we consider the special case of an exactly perpendicular shock.

Acceleration by perpendicular shocks only takes place if CR are able to diffuse across magnetic field lines. Fig. 6 classifies three dif-

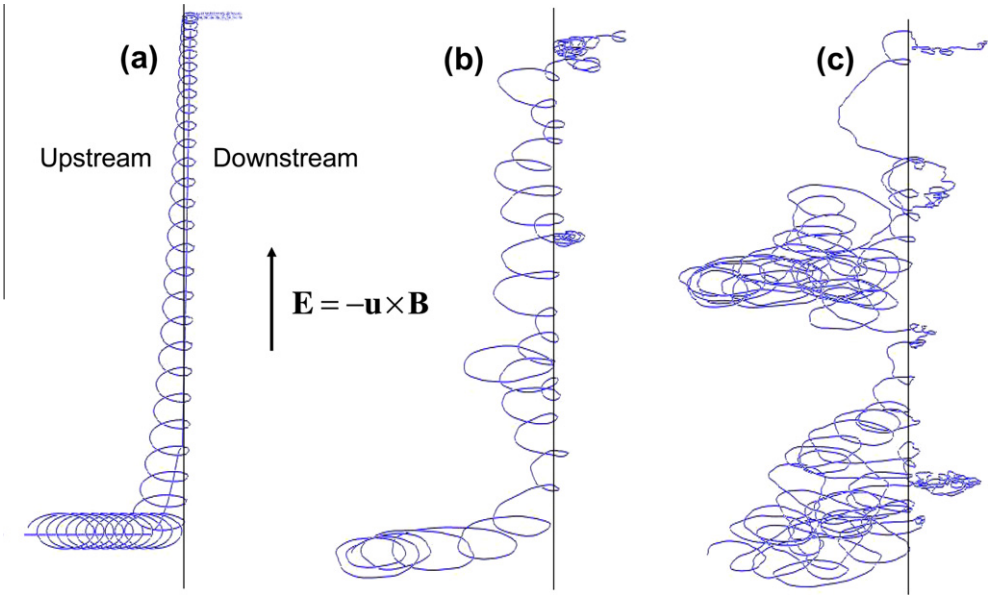

Fig. 6. Illustration of CR trajectories across a perpendicular shock (black line). (a) In the absence of scattering all CR advect across the shock with only a small, approximately adiabatic, increase in energy. (b) Weak scattering allows some CR to gain energy by staying with shock longer without returning fully into the upstream plasma. (c) Strong scattering allows CR to cross and recross the shock as in the case of an unmagnetised shock. The plots are illustrative and not to scale. In none of the cases pictured does the CR gain much energy.

ferent cases. In case (a), scattering is weak and CR are unable to cross field lines. Consequently, CR are tied to a particular field line and are advected through the shock with that field line. They are unable to recross the shock into the upstream plasma and their energy gain is small, approximately corresponding to the adiabatic gain resulting from compression at the shock [58]. In the opposite case, (c), scattering is sufficiently strong that CR are not tied to particular field lines. CR are free to cross and re-cross the shock many times with a statistical chance of large energy gain. The isotropic diffusion theory of Section 2 then gives the correct spectrum. In the intermediate case, (b), CR are able to diffuse across field lines but their opportunity to escape fully into the upstream plasma is limited. Nevertheless, CR have a chance of staying at the shock longer than in case (a). Diffusion of the gyrocentre may allow the CR to remain in contact with the shock for a longer period of time and the CR gains energy as long as its gyrocentre remains within one Larmor radius of the shock. By gyrating between upstream and downstream CR gain energy, but a large energy gain is statistically less probable resulting in an energy spectrum steeper than E^{-2}.

Acceleration by perpendicular shocks is only possible if CR are strongly scattered by large fluctuations in the magnetic field as in cases (b) or (c) in Fig. 6. Acceleration in case (b) is marginal. Nevertheless, if the required conditions are met perpendicular shocks can accelerate CR more rapidly than parallel shocks. As shown in Section 4, the acceleration rate is proportional to u_s^2/D. The diffusion coefficient across a magnetic field is less than that along a magnetic field line by a factor $(r_g/\lambda)^2$ where r_g is the CR Larmor radius, $\lambda = c/\nu$ is the scattering mean free path, and ν is the scattering frequency due to deflection by small scale fluctuations in the magnetic field. The cross-field diffusion coefficient is therefore $D_\perp = (r_g/\lambda)^2 D_\parallel$ where $D_\parallel$ is the diffusion coefficient in an unmagnetised plasma or parallel to a magnetic field. Acceleration at a perpendicular shock is correspondingly $(\lambda/r_g)^2$ faster than at a parallel shock. For favourable scattering lengths, $r_g < \lambda < r_g(c/3u_s)$, perpendicular shocks are rapid CR accelerators [60,61,76].

In principal this suggests that CR can be accelerated to higher energies by perpendicular shocks. Following Lagage and Cesarsky, as set out in Section 3, the maximum CR energy is limited by the condition that the acceleration time $\tau_{accel} = 8D/u_s^2$ should not exceed the age of the SNR. The smaller diffusion coefficient and more rapid acceleration implies an increase in the maximum CR energy when accelerated by a perpendicular shock. As discussed above, CR acceleration is only possible if the precursor scaleheight $L = D/u_s$ exceeds the CR Larmor radius r_g, so the minimum possible diffusion coefficient is $D = r_g u_s$. Most rapid acceleration occurs when the CR trajectory is close to that shown in Fig. 6(b). In Fig. 6(c), the CR spends extended periods upstream or downstream of the shock and acceleration is not as rapid. In Fig. 6(a) the scaleheight for diffusion of the gyrocentre of the CR motion is smaller than a Larmor radius and the CR quickly through the shock at a single pass without opportunity to return from downstream.

Under the optimal conditions, close to those in Fig. 6(b), where $\lambda/r_g \approx c/3u_s$, the acceleration time is $\tau_{accel} = 8r_g/u_s$. This is less than the age τ_{SNR} of a SNR if $r_g < u_s\tau_{SNR}/8$. Since $r_g = E/cB$ where E is the CR energy in eV, the maximum CR energy is $cBR/8$ where $R = u_s\tau_{SNR}$. For historical SNR, $u_s \sim 5\times10^6$ m s^{-1} and $\tau \sim 400$ years, implying a maximum CR energy of $\sim 7\times10^{14}$ eV in the absence of magnetic field amplification ($B \sim 3$ μG) if the maximum CR energy is limited by timescales as in Lagage and Cesarsky [67,68].

Based on a comparison of acceleration timescales with the age of SNR, perpendicular shocks appear to offer acceleration to energies close to the knee without a need for magnetic field amplification [61]. With magnetic field amplification, acceleration beyond the knee appears possible. However, the temporal limit on the maximum CR energy is replaced by a more stringent spatial limit that is not easily overcome. The origin of the spatial limit emerges from an understanding of CR acceleration in terms of CR trajectories in an electric field. Acceleration at any shock, whether perpendicular or parallel, proceeds from CR motion in electric field. Magnetic field by itself can only deflect, not accelerate, particles. The usual understanding of acceleration at parallel shocks hides the effect of the electric field by always considering CR trajectories in the local fluid rest frame where the electric field is zero. The accelerating effect of the electric field is replaced by the frame transformation when a CR crosses between upstream and downstream of the shock.

An alternative perspective on energy gain at a perpendicular shock is obtained by calculating the energy gained from the electric field $\mathbf{E} = -\mathbf{u}\times\mathbf{B}$ where $\mathbf{u}$ and $\mathbf{B}$ are the local fluid velocity and magnetic field. When viewed in the shock rest frame there is a large scale uniform electric field $\mathbf{E} = -\mathbf{u}_s\times\mathbf{B}_0$ where $\mathbf{B}_0$ is the far upstream magnetic field. In the frame in which the shock is at rest and the plasma velocity is normal to the shock, the electric field is the same both upstream and downstream of the shock and $\mathbf{E}$ is perpendicular to the shock normal. CR gain energy from the electric field by propagating along the surface of the shock while crossing the shock as in Fig. 6. This form of 'shock drift' acceleration arises from the different CR Larmor radii upstream and downstream of shock. The field is compressed by the shock so the Larmor radius is smaller downstream causing CR to surf along the surface of the shock. The maximum CR energy gain is then limited by the distance CR can drift along the shock surface. If the maximum distance a CR can drift along the shock surface is approximately equal to the radius R of the SNR, the maximum energy it gains in eV is $|\mathbf{E}|R = u_sB_0R$. This estimate is the same within a numerical factor as that derived for parallel shocks in Section 3. This illustrates the general applicability of the 'Hillas parameter' uBR as a good estimate of the maximum CR energy (in eV) in a variety of acceleration scenarios.

The timescale analysis of Lagage and Cesarsky [67,68] suggests that larger CR energies are possible at perpendicular shocks, but these could only be achieved if the CR surfs along the shock surface a distance much greater than the SNR radius. There are geometries in which this might be possible. If an SNR expands into a uniform magnetic field a CR might gain energy by circulating many times around the equator of the SNR, but possibilities for escape along field lines in the polar direction would make this a rare occurrence and the CR spectrum would steepen at an energy of eu_sB_0R in accord with the Hillas analysis.

The situation regarding perpendicular shocks is uncertain. They undoubtedly have the potential for more rapid CR acceleration, but they require the scattering length λ to lie in a favourable regime, and even then the maximum CR energy is expected to be the same as for parallel shocks because of the spatial limitation. Synchroton X-ray emission from TeV electrons in SN1006 exhibits a polar pattern consistent with preferential CR acceleration at parallel shocks. It has been suggested that this is due to difficulties in injecting and accelerating electrons at a perpendicular shock [70,111,93].

7. Oblique shocks

The special case of perpendicular shocks was discussed in the previous section. Related effects can be important more generally at oblique shocks, especially if they are nearly perpendicular or if the shock velocity is $c/30$ or greater. Previously it has been appreciated that shock obliquity is important at relativistic shocks, where CR have difficulty passing from downstream to upstream (e.g. [13,98]), and in its effect on injection and acceleration out of the thermal pool (e.g. [11,39,49,50]). Here we summarise work reported in Bell et al. [21] which shows that obliquity can have a

strong effect at SNR shocks and thereby affect the steepness of the Galactic CR spectrum.

The diffusion approximation, used in Section 2, breaks down at high velocity oblique shocks and it is necessary to find a solution that allows for higher order anisotropies in the CR distribution. In the diffusion approximation the CR distribution is assumed to take the form

$$f(\mathbf{r},\mathbf{p}) = f_0(\mathbf{r},|\mathbf{p}|) + \mathbf{f}_1(\mathbf{r},|\mathbf{p}|).\frac{\mathbf{p}}{|\mathbf{p}|} \tag{8}$$

where f_0 is the isotropic part of the distribution and $\mathbf{f}_1$ is a diffusive drift. This approximation is adequate at low shock velocities ($u_s \ll c$) or for parallel shocks. The approximation is based on a 'Chapman–Enskog' expansion in the small parameter λ/L where λ is the scattering mean free path and L is the characteristic spatial scalelength. Characteristically $\lambda/L \sim u_s/c$ in the precursor of a parallel shock and the diffusion approximation is valid. The very short scalelength L_s over which velocity and density change at the shock has little effect on CR at a parallel shock since CR respond to the magnetic field rather than density or velocity and there is no discontinuity in the magnetic field at a parallel shock. However, the magnetic field undergoes an abrupt change in direction at an oblique shock and this has a strong effect on CR trajectories. At an oblique shock the large parameter λ/L_s (>1) invalidates the Chapman-Enskog expansion and the diffusion approximation breaks down. The shock discontinuity injects anisotropies of indefinitely high order and additional terms are needed in the CR distribution function:

$$f(\mathbf{r},\mathbf{p}) = f_0(\mathbf{r},p) + \mathbf{f}_1(\mathbf{r},p).\frac{\mathbf{p}}{p} + \mathbf{f}_2(\mathbf{r},p) : \frac{\mathbf{p}}{p}\frac{\mathbf{p}}{p} + \cdots \tag{9}$$

where $p = |\mathbf{p}|$. Bell et al. [21] solved the CR transport equation by expanding the distribution function in spherical harmonics to orders as high as 15. They considered shock velocities in the range $c/30 \leqslant u_s \leqslant c/3$ as observed in very young SNR. The inclusion of non-diffusive terms ($\mathbf{f}_2$ etc.) makes little difference at strictly parallel shocks but makes a significant difference at oblique shocks over a range of shock obliquities. As expected from the above discussion of perpendicular shocks the extra terms make a very strong difference at quasi-perpendicular shocks with $\theta > 70 - 80^\circ$. The strength of the effect depends on the shock velocity and the ratio of the CR scattering mean free path to the CR Larmor radius. The breakdown of the diffusive approximation shows itself as a deviation of the CR spectrum from p^{-4}.

Two opposing effects lead respectively to either a steepening or a flattening of the CR power-law spectrum:

(1) The first effect is seen at shocks which are nearly perpendicular. It takes effect when the CR motion is similar to that plotted in Fig. 6(b) and results in the CR density profile plotted in Fig. 7(c). CR gyration smooths out the CR density over a Larmor radius. The change in CR density across the shock therefore consists of a ramp extending both upstream and downstream of the shock. The crucial factor is that the CR number density at the shock is smaller than the number density far downstream. Consequently the rate at which CR cross and re-cross the shock is reduced relative to the rate at which CR advect away downstream. Fewer CR make a large number of shock crossings leading to a steepening in the CR energy spectrum.
(2) The second effect applies to oblique shocks which are closer to parallel but still sufficiently oblique that CR notice the change in direction of the magnetic field on crossing the shock. The parallel component of the magnetic field is continuous across the shock, but the perpendicular component is compressed and increases by a factor of four at a strong shock. Hence a CR passing from upstream to downstream encounters an increased magnetic field that acts as a barrier and produces a pile-up of CR at the shock as seen in Fig. 7(b). The spike in the CR density at the shock leads to an increase in the rate at which CR cross and re-cross shock compared with the rate at which they advect away downstream. In contrast to the perpendicular case, on average CR cross the shock a greater number of times gaining more energy and leading to a flattening in the CR energy spectrum. Similar spiked profiles at the shock had been seen previously in simulations by Ellison et al. [40].

The essential quantity determining the flattening or steepening of the CR distribution is the ratio of the CR density at the shock to the CR density far downstream. This controls the relative rates at which CR cross the shock gaining energy and are lost from the system by advection downstream. The momentum spectrum $p^{-\gamma}$ has a spectral index given by

$$\gamma \approx 3 + 3(n_\infty/n_s)/(r-1) \tag{10}$$

where n_s and n_∞ are the CR number densities at the shock and far downstream respectively. If $n_s = n_\infty$ as in the diffusive limit of low-velocity or parallel shocks the spectral index is the standard

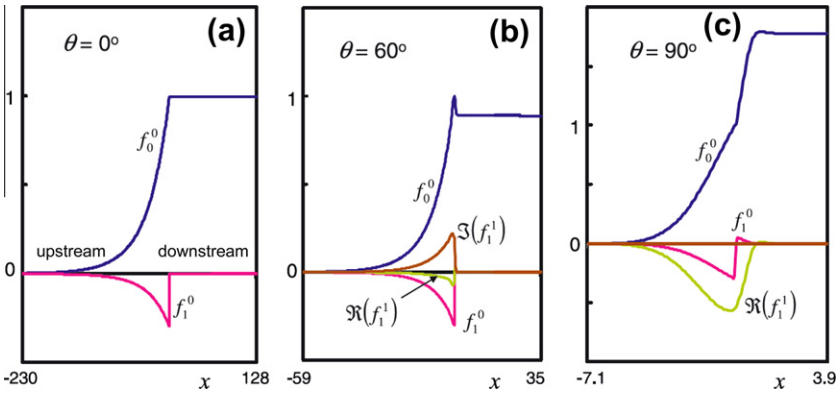

Fig. 7. Spatial profiles of components of the CR distribution function around shocks which are parallel ($\theta = 0^\circ$), oblique ($\theta = 60^\circ$) and perpendicular ($\theta = 90^\circ$). f_0^0 is the isotropic component, f_1^0 is the CR drift along the shock normal, and $\Re\{f_1^1\}$ and $\Im\{f_1^1\}$ are proportional to the CR drifts in the plane of the shock. In the oblique case, the spike in the CR density at the shock produces an excess in the number of CR being accelerated in comparison with the number escaping downstream, thereby flattening the CR spectrum. In the perpendicular case, the smaller number of CR at the shock reduces the acceleration rate and steepens the spectrum. The shock velocity is $c/10$ and the CR scattering frequency is 0.1 times the CR Larmor radius. The spatial distance x from the shock at $x = 0$ is measured in CR Larmor radii. The vertical axis is scaled such that $f_0^0 = 1$ at the shock. Figure reproduced from Bell et al. [21] where more detail can be found.

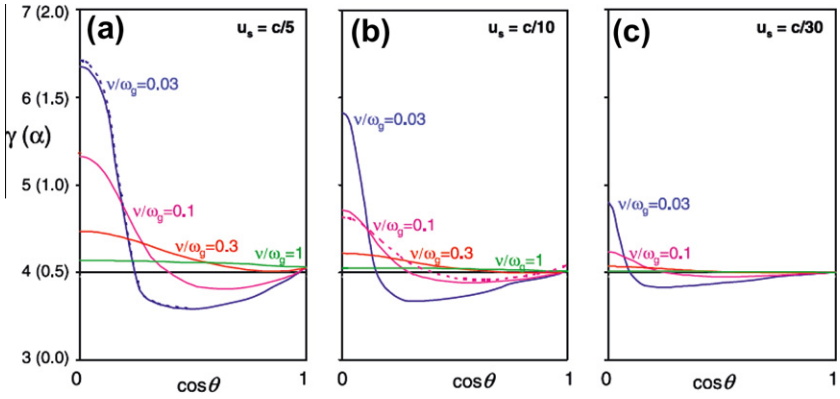

Fig. 8. Plots of CR spectral index $p^{-\gamma}$ and synchrotron spectral index $\nu^{-\alpha}$ for three shock velocities $u_s = c/5,\ c/10$ and $c/30$, four ratios of the CR scattering frequency to the Larmor frequency $\nu/\omega_g = 0.03,\ 0.1,\ 0.3$ and 1.0, and shock obliquities between $\cos\theta = 0$ (perpendicular shock) and $\cos\theta = 1$ (parallel shock). The spectra are steep for quasi-perpendicular shocks and flat for quasi-parallel shocks. The deviations from the diffusive limit ($\gamma = 4, \alpha = 0.5$) are greater at high shock velocity and low collisionality. Figure reproduced from Bell et al. [21] where more detail can be found.

$\gamma = 4$. At perpendicular shocks $n_s < n_\infty$ and the spectrum is steepened. At oblique shocks where $n_s > n_\infty$ the spectrum is flattened.

Numerical results for the spectral index are shown in Fig. 8. They show that even for shock velocities as low as $c/30$ the spectral index can deviate significantly from $\gamma = 4$. This may explain why the Galactic CR spectrum and many radio spectra are steeper than expected for standard CR shock acceleration. The Galactic CR spectrum is remarkably well-fitted by a power law but the spectrum is $E^{-(2.6-2.7)}$ over many orders of magnitude in energy. The steeper spectrum is probably due in part to energy-dependent escape from the Galaxy, but Gaisser et al. [48] and Hillas [56] show that losses can only account for part of the steepening. Electron spectra steeper than E^{-2} are also inferred from the radio spectra of young SNR. The radio synchrotron spectral index of the historical SNR (Tycho, Cas A, Kepler) is $\alpha \sim 0.6 - 0.8$ corresponding to an electron energy spectral $s \sim 2.2 - 2.6$ since $\alpha = (s-1)/2$ where $n(E) \propto E^{-s}$. Radio spectra of recent SN in external galaxies are also steeper, even exceeding $\alpha \sim 1.0,\ s \sim 3.0$. In contrast the synchrotron spectra of SNR in the Sedov phase are usually flatter and approximately consistent with a E^{-2} CR electron energy spectrum. The overall observational picture is that SNR spectra are steep in the first decades following the SN explosion and then flatten with time towards the standard E^{-2} spectrum expected from diffusive acceleration by a strong shock. This trend has been observed during the lifetimes of Cas A and SN1987A. One explanation might be that shocks in young SNR propagate at high velocity into magnetic fields that are predominantly perpendicular rather than parallel either (i) because the SNR expands into a magnetic field frozen into a pre-supernova wind in the form of a Parker spiral or (ii) because in the NRH instability the $\mathbf{j}_{cr} \times \mathbf{B}$ naturally stretches the field in a direction perpendicular to the CR current and the shock normal.

The non-linear processes described in Section 5 might provide an alternative explanation for steep SNR radio spectra where the emitting electrons have relatively low energy, but they would have difficulty accounting for the steepness of the Galactic CR spectrum because of the straightness and lack of concavity in the spectrum over many orders of magnitude in energy.

Deviations from a p^{-4} spectrum in Sedov-phase SNR are unlikely to be caused by shock obliquity since the shock velocities are far below $c/30$. Some older SNR have relatively flat spectra and this may be the result of additional second order Fermi acceleration [80].

8. Other limits to the diffusion model

In the previous section it was shown that the CR spectrum is expected to deviate from the standard p^{-4} spectrum at non-parallel high velocity shocks. Here we outline other ways in which the standard diffusive model might break down.

The diffusive model breaks down if constrictions in magnetic field lines form a 'magnetic bottle'. CR drifting along a field line and encountering a region of high field can be reflected back along the field line. The high field represents a blockage in accordance with conservation of the first adiabatic moment for charged particles. A CR particle can be caught in a magnetic bottle and carried through the shock with a low probability of return from the shock, thus terminating the acceleration of that CR particle. This could reduce the fraction of CR reaching high energies and steepen the spectrum.

Sub-diffusion and super-diffusion [37] may also cause deviations from standard diffusive behaviour. This occurs when CR propagate along wandering field lines. CR execute random walks along the field lines which in turn execute a random walk in space. The combination of the two random walks invalidates the treatment of CR transport as diffusion in a direction parallel to the shock normal. Wandering field lines can trap CR in the equivalent of a magnetic bottle if a field line crosses the shock in more than one place. Kirk et al. [63] showed that sub- and super-diffusion can change the CR spectrum.

At relativistic shocks the diffusion approximation irretrievably breaks down and a treatment is needed that allows for strong anisotropy (e.g. [64,12,1,41]). The Alfven Mach number of the shock and the orientation of the magnetic field are particularly crucial in relativistic shocks [98]. Relativistic shocks are beyond the scope of the present review.

9. Future prospects

Progress in observation and theory over the past decade offers support for the general view that SNR are effective accelerators of cosmic rays to PeV energies. The detection of TeV gamma-rays proves conclusively that particles of some variety are accelerated to at least TeV energies in SNR, and well-resolved X-ray observations prove that TeV electrons and amplified magnetic field are to be found at the outer shocks of SNR. However, showing that SNR shocks accelerate some particles to TeV energies is not the same as proving beyond reasonable doubt that SNR are responsible for all, or nearly all, Galactic cosmic rays at energies up to the knee. There are many outstanding questions that go beyond mere detail in the overall picture.

A particularly outstanding need is for solid observational evidence that protons as well as electrons are accelerated to TeV energies and that the proton spectrum extends to the knee. Gamma-ray

telescopes promise to answer the question by measuring the full gamma-ray spectrum from MeV to 100 TeV. A positive answer to this question, if indeed it is positive, will provide a foundational result for the theory of CR origins. If the sensitivity of Cherenkov techniques can be increased in the 100 TeV range we can hope to discern which SNR at which stage of evolution produce the Galactic cosmic ray population up to the knee.

If PeV cosmic rays are accelerated by relatively young SNR we need to confront the difficult question of how CR escape their parent SNR into the interstellar medium [29,36]. SNR in the pre-Sedov phase have the potential to accelerate CR to high energies because of their high shock velocity and their consequent ability to generate large magnetic field, but CR can only contribute to the Galactic population if they escape the SNR at the time of acceleration. Very few CR escape upstream if the maximum CR energy is limited by timescales as discussed in Section 3. In some detailed analyses of the turnover of the CR spectrum [26,87,29] the acceleration of the highest energy CR is instead terminated by their escape upstream in which case they contribute to the Galactic population without significant adiabatic loss. The fate of the highest energy CR is difficult to determine theoretically and here again observations will be determinative, especially if escaping CR can be detected as they encounter dense clouds outside the SNR.

The shape of the high energy Galactic spectrum in the TeV to PeV range may prove to be governed as much by escape as acceleration [36]. As explained above, the maximum CR energy may drop during the Sedov phase in which case SNR may feed successively lower energy CR into the interstellar medium as the expansion velocity decreases. If it transpires that the high energy part of the Galactic spectrum is set by CR escape and evolution in the maximum CR energy and the low energy part of the spectrum is set by the acceleration process, then we would need to explain why the Galactic CR spectrum follows a single power law from GeV to PeV energies. These issues are far from resolution at the present time.

Our understanding of the Galactic CR spectrum below 1 PeV is incomplete, but the gaps in our understanding of the origin of CR above 1 PeV are considerably greater. The extragalactic origin and the composition of the very highest energy CR is a matter of intense current investigation, but even in the intermediate 1–100 PeV energy range CR origins are very uncertain. Until the possibility of magnetic field amplification was appreciated it was difficult enough theoretically to explain how SNR could accelerate CR to a few PeV. Field amplification probably accounts for acceleration to the knee, but it is difficult to see how the historical SNR, or SNR in the Sedov phase, can accelerate protons beyond the knee. Berezhko and Völk [23] explain the Galactic component above the knee predominantly in terms of progressively higher Z ions with a Galactic origin that have knee energies proportional to Z, while Hillas [56] argues that some CR above the knee must be Galactic protons. If indeed the Galactic proton spectrum extends beyond 1 PeV then this poses a serious challenge to existing theory. One possibility is that the protons are accelerated in the early years following a supernova explosion as the outer SNR shock expands at high velocity into a dense circumstellar medium [20,113,96]. The high circumstellar density and the high shock velocity might allow very strong field amplification [47,31] that more than compensates for the small radius of the shock and boosts the maximum CR energy above 1 PeV. With sufficient circumstellar mass resulting from a dense pre-supernova wind, the energy processed through the shock may be sufficient to produce large numbers of high energy CR. CR acceleration during the very early stages of SNR evolution is an important topic for future study which is theoretically demanding and for which observational data are presently limited.

Cosmic ray research can be presented as a work in progress. Remarkable progress has been made in the past decade and we can hope that future observations, particularly with gamma-ray telescopes, accompanied by supporting theory and simulation will spearhead further remarkable decades of progress.

Acknowledgements

I thank Brian Reville and Klara Schure, my colleagues at Oxford, for insightful discussions on cosmic ray and related physics. The research leading to this review has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement No. 247039 and from Grant No. ST/H001948/1 made by the UK Science Technology and Facilities Council.

References

[1] A. Achterberg et al., MNRAS 328 (2001) 393.
[2] F.A. Aharonian, Very High Energy Cosmic Gamma Radiation, World Scientific Publishing, Singapore, 2004.
[3] F.A. Aharonian, A&A 464 (2007) 235.
[4] Aharonian et al., A&A 531 (2011) C1.
[5] G.E. Allen, J.C. Houck, S.J. Sturner, ApJ 683 (2008) 773.
[6] E. Amato, P. Blasi, MNRAS 371 (2006) 1251.
[7] Auger Collaboration, Science 318 (2007) 938
[8] Auger Collaboration, Astropart. Phys. 34 (2010) 314.
[9] W.I. Axford, E. Leer, G. Skadron, in: Proceedings of the 15th International Cosmic Ray Conference, vol. 11, 1977, p. 132.
[10] W. Baade, F. Zwicky, Proc. Nat. Acad. Sci. 20 (1934) 259.
[11] M.G. Baring, D.C. Ellison, F.C. Jones, ApJ Supp. 90 (1994) 547.
[12] J. Bednarz, M. Ostrowski, Phys. Rev. Lett. 80 (1998) 3911.
[13] M.C. Begelman, J.G. Kirk, ApJ 353 (1990) 66.
[14] A.R. Bell, MNRAS 182 (1978) 147.
[15] A.R. Bell, MNRAS 182 (1978) 443.
[16] A.R. Bell, MNRAS 225 (1987) 615.
[17] A.R. Bell, MNRAS 353 (2004) 550.
[18] A.R. Bell, MNRAS 358 (2005) 181.
[19] A.R. Bell, PPCF 51 (2009) 124004.
[20] A.R. Bell, S.G. Lucek, MNRAS 321 (2001) 433.
[21] A.R. Bell, K.M. Schure, B. Reville, MNRAS 418 (2011) 1208.
[22] E.G. Berezhko, L.T. Ksenofontov, H.J. Völk, A&A 412 (2003) L11.
[23] E.G. Berezhko, H.J. Völk, ApJL 661 (2007) L175.
[24] R.D. Blandford, D. Eichler, Phys. Rep. 154 (1987) 1.
[25] R.D. Blandford, J.P. Ostriker, ApJ 221 (1978) L29.
[26] P. Blasi, S. Gabici, G. Vannoni, MNRAS 361 (2005) 907.
[27] A.M. Bykov, S.M. Osipov, D.C. Ellison, MNRAS 410 (2011) 39.
[28] A.M. Bykov et al., ApJL 735 (2011) 40.
[29] D. Caprioli, P. Blasi, E. Amato, MNRAS 396 (2009) 2065.
[30] G. Cassam-Chenai et al., ApJ 680 (2008) 1180.
[31] R.A. Chevalier, C. Fransson, ApJ 651 (2006) 381.
[32] L.O'C. Drury, Rep. Prog. Phys. 46 (1983) 973.
[33] L.O'C. Drury, F.A. Aharonian, H.J. Völk, A&A 287 (1994) 959.
[34] L.O'C. Drury, S.A.E.G. Falle, MNRAS 222 (1986) 353.
[35] L.O'C. Drury, H.J. Völk, ApJ 248 (1981) 344.
[36] L.O'C. Drury, MNRAS 415 (2011) 1807.
[37] P. Duffy et al., A&A 302 (1995) L21.
[38] D. Eichler, ApJ 229 (1979) 419.
[39] D.C. Ellison, M.G. Baring, F.C. Jones, ApJ 453 (1995) 873.
[40] D.C. Ellison, M.G. Baring, F.C. Jones, ApJ 473 (1996) 1029.
[41] D.C. Ellison, G.P. Double, Astropart. Phys. 22 (2004) 323.
[42] D.C. Ellison, D. Eichler, ApJ 286 (1984) 691.
[43] D.C. Ellison, E. Möbius, G. Paschmann, ApJ 352 (1990) 376.
[44] K.A. Eriksen et al., ApJL 728 (2011) L28.
[45] S.A.E.G. Falle, J.R. Giddings, MNRAS 225 (1987) 399.
[46] E. Fermi, Phys. Rev. 75 (1949) 1169.
[47] C. Fransson, C.-I. Bjornsson, ApJ 509 (1998) 861.
[48] T.K. Gaisser, R.J. Protheroe, T. Stanev, ApJ 492 (1998) 219.
[49] V.L. Galinsky, V.I. Shevchenko, ApJ 734 (2011) 106.
[50] L. Gargate, A. Spitkovsky, ApJ 744 (2012) 67.
[51] J. Giacalone, J.R. Jokipii, ApJL 633 (2007) 41.
[52] F. Giordano et al., ApJ 744 (2012) L2.
[53] E.A. Helder et al., Science 325 (2009) 719.
[54] E.A. Helder et al., Space Sci. Rev. (in press).
[55] A.M. Hillas, ARA&A 22 (1984) 425.
[56] A.M. Hillas, J. Phys. G. Nuclear Phys. 31 (2005) 95.
[57] A.M. Hillas, J. Phys. Conf. Ser. 47 (2006) 168.
[58] P.D. Hudson, MNRAS 131 (1965) 23.
[59] J.P. Hughes, C.E. Rakowski, A. Decourchelle, ApJL 543 (2000) L61.
[60] J.R. Jokipii, ApJ 255 (1982) 716.
[61] J.R. Jokipii, ApJ 313 (1987) 842.
[62] F.C. Jones, D.C. Ellison, Space Sci. Rev. 58 (1991) 259.
[63] J.G. Kirk, P. Duffy, Y.A. Gallant, A&A 314 (1996) 1010.

[64] J.G. Kirk, P. Schneider, ApJ 315 (1987) 425.
[65] G.F. Krymsky, Sov. Phys. Dokl. 23 (1977) 327.
[66] R. Kulsrud, W.P. Pearce, ApJ 156 (1969) 445.
[67] O. Lagage, C.J. Cesarsky, A&A 118 (1983) 223.
[68] O. Lagage, C.J. Cesarsky, ApJ 125 (1983) 249.
[69] M. Lemoine, G. Pelletier, MNRAS 402 (2010) 321.
[70] K.S. Long et al., ApJ 586 (2003) 1162.
[71] M.S. Longair, High Energy Astrophysics, Publ CUP, 2011.
[72] S.G. Lucek, A.R. Bell, MNRAS 314 (2000) 65.
[73] M.A. Malkov, P.H. Diamond, ApJ 692 (2009) 1571.
[74] M.A. Malkov, L.O'C. Drury, Rep. Prog. Phys. 64 (2001) 429.
[75] M.A. Malkov, R.Z. Sagdeev, P.H. Diamond, APJL 748 (2012) L32.
[76] A. Meli, P.L. Biermann, A&A 454 (2006) 687.
[77] M. Nagano, A.A. Watson, Rep. Mod. Phys. 72 (2000) 689.
[78] J. Niemiec, M. Pohl, T. Stroman, K. Nishikawa, ApJ 684 (2008) 1174.
[79] Ohira et al., ApJ 698 (2009) 445.
[80] M. Ostrowski, A&A 345 (1999) 256.
[81] D.J. Patnaude, R.A. Fesen, ApJ 697 (2009) 535.
[82] G. Pelletier, M. Lemoine, A. Marcowith, MNRAS 393 (2009) 587.
[83] G. Pelletier, M. Lemoine, in: G. Alecian, K. Belkacem, R. Samadi, D. Valls-Gabaud (Eds.), SF2A-2011: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, 2011, p. 39.
[84] C.E. Rakowski, ApJL 735 (2011) L21.
[85] J.C. Raymond, ApJL 731 (2011) L14.
[86] B. Reville et al., MNRAS 386 (2008) 509.
[87] B. Reville, J.G. Kirk, P. Duffy, ApJ 694 (2009) 951.
[88] B. Reville, A.R. Bell, MNRAS 419 (2012) 2433.
[89] S.P. Reynolds, D.C. Ellison, ApJ 399 (1992) L75.
[90] M. Riquelme, A. Spitkovsky, ApJ 694 (2009) 626.
[91] M. Riquelme, A. Spitkovsky, ApJ 717 (2010) 1054.
[92] M. Riquelme, A. Spitkovsky, ApJ 733 (2011) 63.
[93] R. Rothenflug et al., A&A 425 (2004) 121.
[94] R. Schlickeiser, ApJ 336 (1989) 243.
[95] K.M. Schure et al., ApSS 43 (2009) 433.
[96] K.M. Schure et al., MNRAS 406 (2010) 2633.
[97] K.M. Schure, A.R. Bell, MNRAS 418 (2011) 782.
[98] L. Sironi, A. Spitkovsky, ApJ 726 (2011) 75.
[99] Skilling, MNRAS 172 (1975) 55.
[100] Skilling, MNRAS 173 (1975) 245.
[101] Skilling, MNRAS 173 (1975) 255.
[102] A. Spitkovsky, in: T. Bulik et al. (Ed.), AIP Conference Proceedings, vol. 801, 2005, p. 345.
[103] A. Spitkovsky, ApJL 673 (2008) L39.
[104] A. Spitkovsky, ApJ 726 (2011) 75.
[105] M.D. Stage, G.E. Allen, J.C. Houck, J.E. Davis, Nature Physics 2 (2006) 614.
[106] Y. Uchiyama et al., Nature 449 (2007) 576.
[107] J. Vink, ApJL 648 (2006) L33.
[108] J. Vink, in: A. Wilson (Ed.), Proceedings of the The X-ray Universe 2005 (ESA SP-604), 2006b, p. 319.
[109] J. Vink, J.M. Laming, ApJ 584 (2003) 758.
[110] J. Vink, AIP Proc. 1085 (2008) 169.
[111] H.J. Völk, E.G. Berezhko, L.T. Ksenofontov, A&A 409 (2003) 563.
[112] H.J. Völk, E.G. Berezhko, L.T. Ksenofontov, A&A 433 (2004) 229.
[113] H.J. Völk, P. Biermann, ApJL 333 (1988) L65.
[114] H.J. Völk, L.O'C. Drury, J.F. McKenzie, A&A 130 (1984) 19.
[115] J.S. Warren et al., ApJ 634 (2005) 376.
[116] E.S. Weibel, Phys. Rev. Lett. 2 (1959) 83.
[117] D.G. Wentzel, ARA&A 12 (1974) 71.
[118] G.P. Zank, G. Li, O. Verkhoglyadova, Space Sci. Rev. 130 (2007) 255.
[119] V.N. Zirakashvili, V.S. Ptuskin, ApJ 678 (2008) 939.
[120] V.N. Zirakashvili, V.S. Ptuskin, H.J. Völk, ApJ 678 (2008) 255.