The Astrophysical Journal, 571:856–865, 2002 June 1

ON THE POSSIBLE REASON FOR NONDETECTION OF TeV PROTONS IN SUPERNOVA REMNANTS

M. A. Malkov and P. H. Diamond
Department of Physics, 9500 Gilman Drive, University of California at San Diego, La Jolla, CA 92093-0319; mmalkov@ucsd.edu

and

T. W. Jones
School of Physics and Astronomy, 116 Church Street SE, University of Minnesota, Minneapolis, MN 55455
Received 2001 November 30; accepted 2002 February 7

ABSTRACT

The theory of shock acceleration predicts the maximum particle energy to be limited only by the acceleration time and the size (geometry) of the shock. This led to optimistic estimates for the galactic cosmic-ray energy achievable in supernova remnant (SNR) shocks. The estimates imply that the accelerated particles, while making *no strong impact on the shock structure* (test-particle approach), are nevertheless scattered by the *strong self-generated* Alfvén waves (turbulent boost) needed to accelerate them quickly. We demonstrate that these two assumptions are in conflict when applied to SNRs of the age required for cosmic-ray acceleration to the "knee" energy. We study the *combined* effect of acceleration nonlinearity (shock modification by accelerated particles) and wave generation on the acceleration process. We show that the refraction of self-generated waves resulting from the deceleration of the plasma flow by the pressure of energetic particles causes enhanced losses of these particles. This effect slows down the acceleration and changes the shape of the particle spectrum near the cutoff. The implications for observations of TeV emission from SNRs are also discussed.

Subject headings: acceleration of particles — cosmic rays — shock waves — supernova remnants — turbulence

1. INTRODUCTION

The first-order Fermi or diffusive-shock acceleration (DSA) has long been considered to be responsible for the production of galactic cosmic rays (CRs) in supernova remnants (SNRs), as well as for the radio, X-ray, and γ-ray emission from these and other shock-related objects. The most crucial characteristic of this process that is usually examined in terms of its capability to explain a given observation is the rate at which it operates. Indeed, what is often expected from theory or even inferred from observations is an extended particle energy spectrum, frequently a power law, but more rapidly decaying at the highest energies observed. Often, this decay is referred to as an energy or momentum cutoff and is usually associated with a finite acceleration time or with losses if the loss rate exceeds the acceleration rate. As long as the losses are unimportant, the cutoff $p_{\max}(t)$ advances with time according to the following equation,

$$\frac{dp_{\max}}{dt} = \frac{p_{\max}}{t_{\rm acc}} \ , \tag{1}$$

whereas in the presence of losses, the acceleration rate $p_{\max}/t_{\rm acc}$ can be equated with the loss rate to yield a steady state value of $p_{\max}$. The acceleration timescale is determined by (e.g., Axford 1981)

$$t_{\rm acc} = \frac{3}{u_1 - u_2} \int_{p_{\min}}^{p_{\max}} \left[\frac{\kappa_1(p)}{u_1} + \frac{\kappa_2(p)}{u_2} \right] \frac{dp}{p} \ , \tag{2}$$

with u_1 and u_2 being the upstream and downstream flow speeds in the shock frame and with $\kappa_{1,2}$ being the particle diffusivities in the respective media. One can recognize in the last formula the sum of average residence times of a particle spent upstream and downstream of the shock front before it completes one acceleration cycle, integrated over the entire acceleration history from $p_{\min}$ to $p_{\max}$. Given the flow speeds $u_{1,2}$, which in many cases are known reasonably well, the most sensitive quantity is the particle diffusivity κ. This, in turn, is determined by the rate at which particles are pitch angle scattered by the Alfvén turbulence. If the latter were just background turbulence in the interstellar medium (ISM), the acceleration process would be too slow to produce the galactic CRs in SNRs (e.g., Lagage & Cesarsky 1983). However, it was realized (e.g., Bell 1978; Blandford & Ostriker 1978) that accelerated particles should create the scattering environment by themselves, generating Alfvén waves on the cyclotron resonance $kp\mu/m = \omega_c$, where k is the magnitude of the wavevector (directed along the magnetic field) and p, μ, m, and ω_c are the particle momentum, cosine of its pitch angle, mass, and nonrelativistic (eB/mc) gyrofrequency. Note that the diffusive character of particle transport (and the determination of κ) has been rigorously obtained within a quasi-linear theory, i.e., it is subject to constraints on the turbulence level.

The wave generation, however, proved to be very efficient (see e.g., Völk, Drury, & McKenzie 1984; Ko 1991, and § 2). In particular, using again the quasi-linear approximation, the normalized wave energy density $(\delta B/B_0)^2$ can be related to the partial pressure P_C of CRs that resonantly drive these waves through

$$(\delta B/B_0)^2 \sim M_{\rm A} P_C/\rho u^2 \ , \tag{3}$$

where $M_{\rm A}$ is the Alfvén Mach number and ρu^2 is the shock ram pressure. Since $M_{\rm A}$ is typically a large parameter, $\delta B/B_0$ may become larger than unity, even if the acceleration itself is relatively inefficient, i.e., if $P_C/\rho u^2 \ll 1$. Strictly speaking, this invalidates the quasi-linear approach as a means for describing the generation of strong turbulence at

shocks. The commonly accepted way to circumvent this difficulty is to assume that the turbulence saturates at $\delta B/B_0 \sim 1$, which means that the mean free path of pitch angle scattered particles is of the order of their gyroradius r_g. Then, $\kappa = \kappa_B \equiv cr_g(p)/3$, where the speed of light c is substituted for the CR velocity and κ_B stands for the Bohm diffusion coefficient. This immediately sets the acceleration timescale in equation (2) at the level of the particle gyroperiod $(eB/p)^{-1}$ times $(c/u_1)^2$. In principle, the turbulence level $\delta B/B_0$ significantly exceeding unity is possible in local shock environments (see, e.g., numerical studies by Bennett & Ellison 1995; Bell & Lucek 2000). As a consequence of that, the diffusion coefficient could be even smaller than κ_B, and hence the acceleration rate would be faster than it is commonly believed to be. At the same time, since Alfvénic-type turbulence is usually considered, the respective velocity perturbations must be super-Alfvénic and supersonic, which raises questions about the ability of the turbulence to sustain itself in an extended area without rapid dissipation that will decrease the $\delta B/B_0$ level. Likewise, decreasing of the turbulence level below the Bohm limit, due to the finite extent of the turbulence zone upstream and wave damping, should slow down the acceleration (Lagage & Cesarsky 1983; Achterberg & Blandford 1986).

However, the acceleration rate given by equation (2) with $\kappa = \kappa_B$ was found to be fast enough to explain (at least marginally) the acceleration of CRs in SNRs up to the "knee" energy, $\sim 10^{15}$ eV, over their lifetime. Much further optimism has been caused by the studies of Drury, Aharonian, & Völk (1994) and Naito & Takahara (1994). They analyzed the prospects for the detection of super-TeV emission from nearby SNRs that should be produced by the decays of π^0 mesons born in collisions of shock-accelerated protons with the nuclei of interstellar gas. The expected fluxes were shown to be detectable by imaging Cerenkov telescopes. Note that similar calculations with similar conclusions have been performed by Berezinsky & Ptuskin (1989). Moreover, EGRET (Esposito et al. 1996) detected a lower energy ($\lesssim 1$ GeV) emission coinciding with some galactic SNRs. The spectra also seemed consistent with the DSA predictions. One can even argue that the low-energy EGRET data verified one of the most difficult elements of the entire acceleration mechanism, the so-called injection. In essence, this is a selection process (not completely understood) whereby a small number of thermal particles become subject to further acceleration (see Gieseler, Jones, & Kang 2000; Zank et al. 2001 for the latest development of the injection theory and Malkov & Drury 2001 for a review) and can then be treated by standard means of the DSA theory that was designed to describe particles with velocities much higher than the shock velocity. Therefore, what seemed left for the theory was to continue the EGRET spectrum (that sets the normalization constant, or injection rate) with some standard DSA slope (nearly E^{-2} or somewhat steeper) and to predict the γ-ray flux in the TeV range in which it could be detected by Cerenkov telescopes.

Unfortunately, despite the physical robustness of the arguments given by Berezinsky & Ptuskin (1989), Drury et al. (1994), and Naito & Takahara (1994), no statistically significant signal that could be attributed to any of the EGRET sources was detected. The further complication is that a critical energy band between GeV and TeV energies is currently not covered by available instruments. Therefore, based on these observational results, it was suggested (e.g., Buckley et al. 1998) that there is probably a spectral break or even cutoff somewhere within this band. However, the spectrum above GeV energies remains an enigma. This will be resolved perhaps with the launch of the *Gamma-Ray Large-Area Space Telescope* mission and when the new generation of Cerenkov telescopes with lower energy thresholds begins to operate. However, the discovery of the 100 TeV emission from SNR 1006 (Tanimori et al. 1998), as well as some other remnants not seen by EGRET at lower energies (see, e.g., Atoyan et al. 2000; Aharonian et al. 2001; Allen, Petre, & Gotthelf 2001; Kirk & Dendy 2001 for a complete discussion), although almost universally identified with electrons diffusively accelerated to similar energies, is widely interpreted as a strong support of the mechanism itself. The above suggests, however, that in reality, it might be not as robust as is its simplified test-particle version with enhanced turbulence and particle scattering.

In this paper we attempt to understand what may happen to the spectrum provided that the acceleration is indeed fast enough to access TeV energies over the lifetime of the SNRs in question. Our starting point is that the fast acceleration also means that the pressure of accelerated particles becomes significant in an early stage of supernova evolution so that the shock structure is highly nonlinear. At first glance, this should not slow down acceleration, since according to equation (3), this changes the turbulence *level*, thus improving particle confinement near the shock front and thus making acceleration faster (smaller κ). However, the formation of a long CR precursor (in which the upstream flow is gradually decelerated by the pressure of CRs, P_C) influences the *spectral properties* of the turbulence by affecting the propagation and excitation of the Alfvén waves. This effect is twofold. First, the waves are compressed in the converging plasma flow upstream and are thus blueshifted, eliminating the long waves needed to keep exactly the highest energy particles diffusively bound to the accelerator. Second, and as a result of the first, at the highest energies there remain fewer particles than expected, so that the level of resonant waves is smaller, and hence the acceleration rate is lower. We believe that these effects have been largely overlooked before, which may have caused a substantial overestimation of the particle maximum energy in strongly nonlinear regimes.

2. BASIC EQUATIONS AND APPROXIMATIONS

We use the standard diffusion-convection equation for describing the transport of high-energy particles (CRs) near a CR-modified shock. First, we normalize the distribution function $f(p)$ to $p^2\,dp$:

$$\frac{\partial f}{\partial t} + U\frac{\partial f}{\partial x} - \frac{\partial}{\partial x}\kappa\frac{\partial f}{\partial x} = \frac{1}{3}\frac{\partial U}{\partial x}p\frac{\partial f}{\partial p}\ . \tag{4}$$

Here x is directed along the shock normal, which for simplicity is assumed to be the direction of the ambient magnetic field. The two quantities that control the acceleration process are the flow profile $U(x)$ and the particle diffusivity $\kappa(x, p)$. The first one is coupled to the particle distribution f through the equations of mass and momen-

tum conservation,

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}\rho U = 0 \,, \tag{5}$$

$$\frac{\partial}{\partial t}\rho U + \frac{\partial}{\partial x}\left(\rho U^2 + P_C + P_g\right) = 0 \,, \tag{6}$$

where

$$P_C(x) = \frac{4\pi}{3} mc^2 \int_{p_{\rm inj}}^{\infty} \frac{p^4 dp}{\sqrt{p^2+1}} f(p,\ x) \tag{7}$$

is the pressure of the CR gas, P_g is the thermal gas pressure, and ρ is its density. The lower boundary in momentum space, $p_{\rm inj}$, separates CRs from the thermal plasma that is not directly involved in our kinetic treatment of CRs but rather enters the equations through the magnitude of f at $p = p_{\rm inj}$, which specifies the injection rate of thermal plasma into the acceleration process. The injection rate is assumed to be small and does not affect the gas density conservation given by equation (5) or the gas pressure in equation (6). These effects have been systematically included by Kang & Jones (1990). The particle momentum p is normalized to mc.

Since we are primarily concerned with the wave generation and particle confinement upstream of the discontinuity, we assume that the upstream region is at $x > 0$, so that the velocity profile can be represented in the shock frame as $U(x) = -u(x)$, where the (positive) flow speed $u(x)$ jumps from $u_2 \equiv u(0_-)$ downstream to $u_0 \equiv u(0_+) > u_2$ across the subshock and then gradually increases up to $u_1 \equiv u(+\infty) \geq u_0$ (see Fig. 2*a*).

We can neglect the contribution of the gas pressure to equation (6) in the upstream region ($x > 0$, but not at $x \leq 0$), restricting our consideration to the high Mach number shocks, $M \gg 1$ (see Malkov 1997 for the details of this approximation). Of course, the gas pressure is retained when treating the subshock (discontinuous part of the shock structure). Since the CR pressure P_C and the related energy E_C do not vary on the subshock scale, the jumps of all relevant physical quantities can be obtained from the conventional Rankine-Hugoniot conditions. In particular, for the flow compression at the subshock, we have

$$\frac{u_0}{u_2} = \frac{\gamma+1}{\gamma - 1 + 2M_0^{-2}} \,. \tag{8}$$

Here M_0 is the Mach number in front of the subshock. When the flow compression in the CR precursor can be considered as adiabatic, this can be expressed through the given far upstream Mach number in a standard way, $M_0^2 = M^2/R^{\gamma+1}$, where $R \equiv u_1/u_0$ is the flow precompression in the CR precursor. We also set $\gamma = 5/3$ in what follows.

Equations (4)–(8) self-consistently describe both the shock structure $u(x)$ and the particle spectrum $f(x,\ p)$, given its normalization (injection rate) and the particle diffusivity $\kappa(x,\ p)$. The determination of these two parameters is a serious problem in its own right, which is discussed in the next subsection. Even when they are parameterized rather than self-consistently determined, there are still two different approaches to the system formed by equations (4)–(8). One is the so-called two-fluid model (TFM), which simply takes the energy moment E_C of equation (4) and thus eliminates the particle momentum p, leaving, however, the remaining system unclosed, generally speaking. The TFM has been introduced and partly analyzed by Axford, Leer, & Skadron (1977). A complete graphical classification of its stationary solutions has been given by Drury & Völk (1981), while Axford, Leer, & McKenzie (1982) gave the full analytic solution. Time-dependent solutions have been studied, e.g., by Dorfi (1990), Jones & Kang (1992), and Donohue et al. (1994). The TFM closure problem becomes serious for κ growing substantially with p. There are theoretical indications that the character of the underlying kinetic solution changes if this growth is faster than $p^{1/2}$ and it is no longer quantitatively described correctly by the TFM (e.g., Malkov & Drury 2001). We therefore take a kinetic approach, which is also discussed in the next subsection.

Turning to the determination of the CR diffusion coefficient κ, we note that since the CR precursor scale height is $\sim\kappa(p_{\rm max})/u_1 \sim (c/u_1) r_g(p_{\rm max})$, which is still $c/u_1 \gg 1$ times larger than the longest wave in the spectrum, $\sim r_g(p_{\rm max})$, we can use a wave kinetic equation in the eikonal approximation for describing the evolution of Alfvén waves:

$$\frac{\partial N_k}{\partial t} + \frac{\partial \omega}{\partial k}\frac{\partial N_k}{\partial x} - \frac{\partial \omega}{\partial x}\frac{\partial N_k}{\partial k} = \gamma_k N_k + {\rm St}\{N_k\} \,. \tag{9}$$

Here N_k is the number of wave quanta and ω is the wave frequency $\omega = -ku + kV_{\rm A} \simeq -ku$. The left-hand side has the usual Hamiltonian form that states the conservation of N_k along the lines of constant frequency, $\omega(k,\ x) = {\rm const}$ on the $(k,\ x)$-plane. The first term on the right-hand side describes the wave generation from the cyclotron instability of a slightly anisotropic particle distribution. It can be expressed through its spatial gradient. The resonance condition for the wave-particle interaction also contains the particle pitch angle $\cos^{-1}\mu$ by means of the expression $kp\mu = eB/c$, which generally speaking requires the treatment of particle distribution in two-dimensional momentum space $(p,\ \mu)$. A significant simplification can be achieved by the so-called resonance sharpening procedure (Skilling 1975; Drury, Duffy, & Kirk 1996), whereby a certain "optimal" value of μ is ascribed to all particles, and the resonance condition puts k and p into a one-to-one relation, i.e., $kp = {\rm const}$. The second term on the right-hand side stands for nonlinear wave-particle and wave-wave interactions such as the induced scattering of waves on thermal protons and mode coupling (Sagdeev & Galeev 1969). We suggest a simple model for this nonlinear term in § 3.2.

To conclude this subsection, we emphasize that while equations (4)–(8) already treat the acceleration process and flow structure on equal footing, the fluctuation part given by equation (9) must be included in this treatment, and as we see in the next section, it by no means plays a subdominant role in this triad.

2.1. *The Significance of Acceleration Nonlinearities*

There are two aspects of the acceleration for which nonlinearity is crucial to its outcome. The first aspect is the excitation of scattering waves by accelerated particles and the second one is the back-reaction of these particles on the shock structure. The latter is critical for both the particle injection and the wave excitation; that is, for particle confinement.

Indeed, as we stated, the system of equations (4)–(8) self-consistently describes particle acceleration and the shock structure (nonlinearly modified by the particle pressure) only if the particle scattering law is known (which is con-

tained in the diffusion coefficient κ) and the injection rate from the thermal plasma is also known. Physically, the scattering rate determines the particle maximum momentum $p_{\max}$, as equation (2) indicates. The difficulty, however, is that both the cutoff momentum $p_{\max}$ and the wavenumber cutoff of the scattering turbulence change in time *simultaneously* (one controlling the other) due to the cyclotron resonance condition. However, the speed at which they change has not been calculated self-consistently. The *linear* solution given by equations (1) and (2) is essentially based on the assumption that $p_{\max}$ is growing with the help of *already-existing stationary turbulence*. In reality, the particle energy cutoff and the corresponding cutoff on the wave spectrum, as we mentioned, both advance together, and since waves need to grow from a very small background amplitude at each current cutoff position, an additional slow-down must be introduced into the entire process. A good analogy here is the problem of beam relaxation in plasmas (Ivanov 1978) (in which a front on the particle velocity distribution also propagates on self-generated rather than on preexisting resonant waves). This suggests that the speed of the front in momentum space, as given by equations (1) and (2), should be reduced by a factor of $\sim\ln(W/W_{\rm ISM})$, where $W_{\rm ISM}$ is the background turbulence amplitude and W is the saturated wave amplitude generated by accelerated particles. As we mentioned, the latter may be associated with $W_{\rm B} = B_0^2/8\pi$, so that the acceleration time given by equation (2) may increase by a factor of $\sim$10 (e.g., Achterberg, Blandford, & Reynolds [1994] estimate $W_{\rm ISM}/W_{\rm B} \sim 10^{-5}$). Evidently, the additional logarithmic factor takes care of the time needed for waves to grow before they start to scatter particles with the current momentum p at the Bohm rate.

The above consideration also shows that particles with $p < p_{\max}$ are confined to the shock through fast pitch angle scattering, while particles with $p > p_{\max}$ are only scattered very slowly due to the absence of self-generated waves and leave the accelerator. Mathematically, this means $f(p > p_{\max}) \equiv 0$ or $\kappa(p > p_{\max}) \equiv \infty$. Note that the propagating front solution must produce a different (sharper) cutoff shape at $p = p_{\max}(t)$ than those from approaches based on the preexisting turbulence, i.e., on a prescribed (for all p) $\kappa(p)$, (e.g., Berezhko, Yelshin, & Ksenofontov 1996). Even if the speed and the form of the front at $p_{\max}(t)$ are unknown, the above *Ansatz* allows an analytic solution of the system in equations (4)–(8) (Malkov 1997) for $p < p_{\max}$ in the limit of strong shocks ($M \gg 1$), for high maximum momentum $p_{\max}$ (that may slowly advance in time) and for essentially arbitrary, but in particular Bohm, $\kappa(p)$ dependence for $p < p_{\max}$ (which as we mentioned is often assumed in numerical studies for all p, e.g., Duffy 1992). The analytic solutions are tabulated, e.g., by Malkov & Drury (2001), and extensively used below.

Since waves are generated by accelerated particles upstream in the precursor, the main nonlinear impact on the wave dynamics and thus on $p_{\max}$ must be from the flow precompression. The latter can be characterized by the parameter $R = u_1/u_0$, which is shown in Figure 1 as a function of the injection parameter ν for different maximum momenta $p_{\max}$. The injection parameter ν is related to the normalization of the particle distribution function f in equation (4) as

$$\nu = \frac{4\pi}{3}\frac{mc^2}{\rho_1 u_1^2} p_{\rm inj}^4 f_0(p_{\rm inj}) \ , \qquad (10)$$

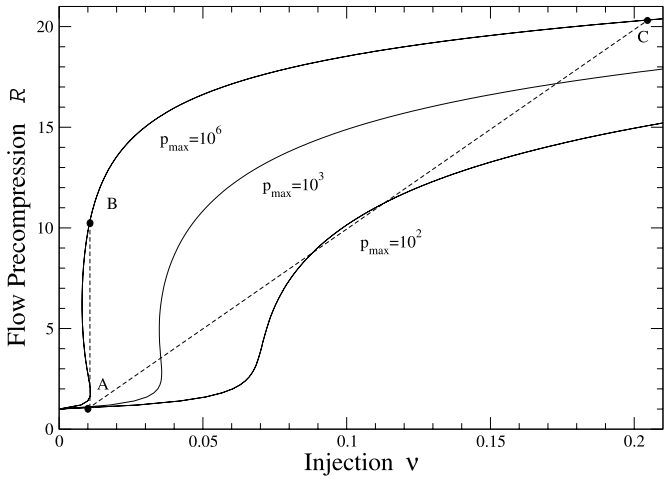

FIG. 1.—Response of the shock structure (bifurcation diagram) to the injection of thermal particles at the rate ν. The strength of the response is characterized by the precompression of the flow in the CR shock precursor $R = u_1/u_0$. The flow Mach number $M = 150$. Different curves correspond to different values of maximum momentum normalized to mc. For each given ν and $p_{\max}$, one (for $p_{\max} < p_{\rm cr} \simeq 500$) or three (for $p_{\max} \geq p_{\rm cr}$) solutions exist. Note that solution multiplicity does not exist for shocks with $M \leq M_{\rm cr} \simeq 70$ (Malkov et al. 2000; Malkov & Drury 2001). Given an initial injection ν and compression R at point A [with $R(A) \approx 1$], the injection and R at point C are calculated as $\nu(C) = R(C)\nu(A)$ (see text for further explanations).

where $f_0(p)$ is the downstream value of f. In this form, the injection rate ν naturally appears as a coefficient in front of the CR pressure in the momentum flux conservation equation (6) when the CR pressure is normalized to the ram pressure $\rho_1 u_1^2$ (see eq. [12] in § 3).

One aspect of the solution shown in Figure 1 that is important here is that for any given injection rate ν, the growing maximum momentum $p_{\max}(t)$ will ultimately exceed a critical value, beyond which the test-particle regime fails to exist. (It is natural to assume that the acceleration starts at this regime, i.e., where $R \approx 1$, e.g., point A on Fig. 1). Formally, the system must then transit to a much higher R that will still be very sensitive to the current values of ν ($\approx$0.01) and $p_{\max}$ ($=10^6$), as can be seen from Figure 1 (point B). Obviously, the further development of the acceleration process will depend on how the parameters ν and $p_{\max}$ react to this strong increase of R. One possibility is to simply assume that a constant fraction of the subshock plasma is injected, so that the injection rate substantially increases because the plasma density at the subshock grows linearly with R. Then, the system must leave the critical region where the $R(\nu)$ dependence is very sharp or even nonunique and proceed to a highly supercritical regime characterized by higher ν (point C). The curve $R(\nu)$ saturates there at the level $\propto M^{3/4}$, which in the most straightforward way can be deduced from the condition of the subshock preservation, $M_0 \gtrsim 1$ (see eq. [8]). A general formula for $R(\nu, M)$ with the $M^{3/4}$ scaling as a limiting case can be found in Malkov (1997). This scenario was realized in many numerical models (e.g., Ellison & Eichler 1985; Kazanas & Ellison 1986; Berezhko et al. 1996), since they normalized the injection parameter to the plasma density at the subshock $\rho_0 = \rho_1 R$, which should clearly lead to the $R \propto M^{3/4}$ scaling. Obviously, the precompression R and thus the acceleration efficiency will then be insensitive to ν (in deep contrast to the case $\nu \simeq$ const, point B) since point C is already on

the saturated part of the $R(\nu)$ curve. Often, this insensitivity is observed in numerical studies with the parameterized injection rate (e.g., Berezhko et al. 1996), so it is tempting to conclude that we do not need to know the injection rate very accurately as soon as it exceeds the critical rate.

However, the injection rate is known to be suppressed by a number of self-regulating mechanisms such as trapping of thermal particles downstream by the injection driven turbulence (Malkov 1998) and insufficient heating of the downstream plasma in strongly modified shocks. These effects seem to more than compensate for the compressive growth of plasma density. Recently, these effects have been systematically included in numerical studies by Gieseler et al. (2000) and Kang, Jones, & Gieseler (2001). They did not confirm the simple $\nu \propto R$ rule. Instead, they indicated that in the course of nonlinear shock modification accompanied by growing R, the injection rate ν remains remarkably constant (Gieseler et al. 2000). Moreover, the preliminary results of a new adaptive mesh refinement (AMR) modification of these schemes, allowing higher $p_{\max}$, indicate that the injection efficiency may even begin to decrease with growing $p_{\max}$ (Kang et al. 2001; Kang, Jones, & Gieseler 2002). These self-regulation mechanisms are applicable to both strictly parallel and oblique shocks, of which the former are clearly an exceptional case. Even slightly oblique shocks have an additional self-regulation of injection via a nonlinearly increasing obliquity. Indeed, since the tangential magnetic field component B_t is amplified at the subshock by a factor of R, the subshock may be strongly oblique even if the shock itself is not. This leads to an exponentially strong suppression of the leakage of downstream thermal particles to upstream (for a Maxwellian downstream distribution), since the intersection point of a field line (which the particles sit on) with the shock front rapidly moves away from these particles. On the other hand, enhanced particle reflection off the oblique subshock should increase injection.

Inspection of Figure 1 shows that if we (conservatively) assume $\nu(R) = \text{const}$ (AB) rather than $\nu \propto R$ (AC), the results will differ dramatically, particularly in terms of the injection rate. Note that particle spectra that correspond to the points B and C also differ very strongly (see Malkov & Drury 2001 for graphical examples). What is important for the subject of the present paper is that in both these cases, as well as for any other point on the part BC of the bifurcation curve, the compression R is very high. It has been pointed out by Malkov, Diamond, & Völk (2000) that this must have a strong impact not only on the injection rate as discussed above, but also on the wave propagation and thus on the particle confinement. This in turn should lead to significant reduction of the maximum momentum achievable by this acceleration mechanism. We quantify these effects in the next section.

3. ANALYSIS

Returning to equations (4) and (9), it is convenient to use the wave energy density I_k normalized to $d\ln k$ and to the energy density of the background magnetic field $B_0^2/8\pi$ instead of N_k,

$$I_k = \frac{k^2 V_{\rm A}}{B_0^2/8\pi} N_k \ , \tag{11}$$

along with the partial pressure of CRs normalized to $d\ln p$ and to the shock ram pressure $\rho_1 u_1^2$:

$$P = \frac{4\pi}{3}\frac{mc^2}{\rho_1 u_1^2}\frac{p^5}{\sqrt{p^2+1}} f(p,\ x) \ . \tag{12}$$

Using these variables, denoting $g = P/p$, and assuming a steady state and $p \gg 1$, equations (4) and (9) can be rewritten as (Bell 1978; Drury et al. 1996)

$$\frac{\partial}{\partial x}\left(ug + \kappa\frac{\partial g}{\partial x}\right) = \frac{1}{3}u_x p\frac{\partial g}{\partial p} \ , \tag{13}$$

$$u\frac{\partial I}{\partial x} + u_x p^3\frac{\partial}{\partial p}\frac{I}{p^2} = \frac{2u_1^2}{V_{\rm A}}\frac{\partial}{\partial x}P - {\rm St}\{I\} \ . \tag{14}$$

Here $u_x \equiv \partial u/\partial x$ and the wave intensity $I \equiv I(p) = I_k$ is now treated as a function of p rather than k, according to the resonance relation $kp = \text{const}$. The CR diffusion coefficient κ can be expressed through the wave intensity by

$$\kappa = \frac{\kappa_{\rm B}}{I} \ , \tag{15}$$

where $\kappa_{\rm B}(p)$ is the Bohm diffusion coefficient. The difference between these equations and those used by, e.g., Bell (1978) and Drury et al. (1996) is due to the terms with $u_x \neq 0$ and the Stoss term on the right-hand side of equation (14). Far away from the subshock, where $u_x \to 0$ and where the wave collision term is also small due to the low particle pressure P, one simply obtains

$$I = \frac{2u_1}{V_{\rm A}}P \ . \tag{16}$$

Note that this shows the limitation of the linear approach in the case of strong shocks, $M_{\rm A} \equiv u_1/V_{\rm A} \gg 1$. The most important change to the acceleration process comes from the terms with $u_x \neq 0$. Indeed, let us first recall how equation (13) can be treated in the linear case $u_x \equiv 0$ for $x > 0$. We integrate both sides between some $x > 0$ and $+\infty$, which yields

$$u_1 g + \frac{\kappa_0 V_{\rm A}}{2u_1}\frac{1}{g}\frac{\partial g}{\partial x} = 0 \ , \tag{17}$$

where we denote $\kappa_0 \equiv \kappa_{\rm B}/p \simeq \text{const}$ for $p \gg 1$. Although this equation has a formal spatial scale $l \sim \kappa_0/u_1 M_{\rm A} g$, its only solution is a power law,

$$g \propto 1/(x + x_0) \ , \tag{18}$$

and thus has no scale [$x_0 = x_0(p)$ is an integration constant]. It simply states the balance between the diffusive flux of particles upstream (second term in eq. [17]) and their advection with thermal plasma downstream (the first term). As we see below, this balance is not possible everywhere upstream, and the physical reason why it appears to be so robust in the case $u_x = 0$ is that flows of particles and waves on the (x, p)-plane (including the diffusive particle transport) are both directed along the x-axis. If, however, the flow modification upstream is significant ($u_x > 0, x > 0$), the situation changes fundamentally. Figure 2 explains how the flows of particles and waves on the (x, p)-plane become misaligned, even though they are both advected with the thermal plasma. In fact, the flows separate from each other, and since neither of them can exist without the other (waves are generated by particles that, in turn, are trapped in the shock precursor by the waves), they both disappear in some part of phase space.

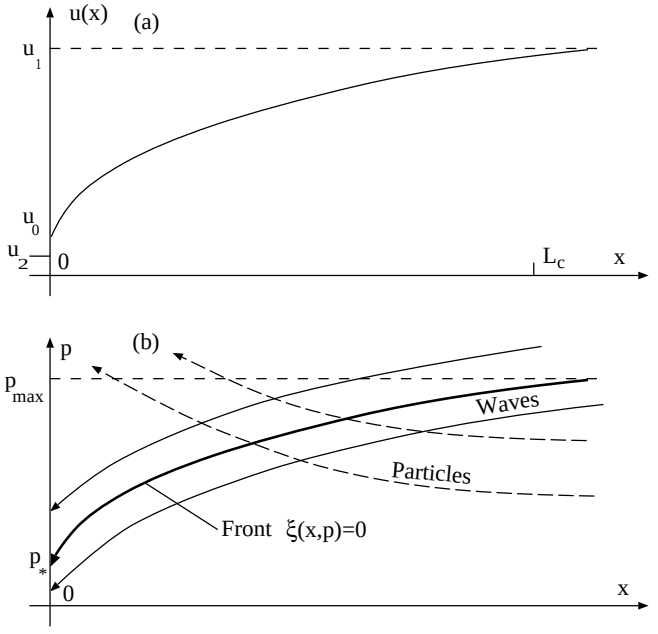

FIG. 2.—(*a*) Flow structure and (*b*) phase plane of particles and resonant waves ($p_* = p_{\max} u_0/u_1$; see text).

To understand how this happens, we rewrite equations (13) and (14) in the following characteristic form (we return to the particle number density f):

$$\left(u\frac{\partial}{\partial x} - \frac{1}{3}u_x p\frac{\partial}{\partial p}\right) f = -\frac{\partial}{\partial x}\kappa\frac{\partial f}{\partial x} \ , \tag{19}$$

$$\left(u\frac{\partial}{\partial x} + u_x p\frac{\partial}{\partial p}\right)\frac{I}{p^2} = \frac{2u_1^2}{V_A p^2}\frac{\partial}{\partial x}P - \frac{1}{p^2}\mathrm{St}\{I\} \ . \tag{20}$$

One sees from the left-hand sides of these equations that particles are transported toward the subshock in x and upward in p along the family of characteristics $up^3 = \text{const}$, whereas waves also move toward the subshock but downward in p along the characteristics $u/p = \text{const}$, so that the longest waves generated by the highest energy particles ($p \simeq p_{\max}$), far upstream where $u \simeq u_1$, are transported with the flow to $p = p_* \equiv p_{\max} u_0/u_1$ when they reach the subshock ($u \equiv u_0$). As long as $u(x)$ does not significantly change ($u_0 \approx u_1$), the waves and particles propagate together (along the x-axis) as, e.g., in the case of unmodified shock or far away from the subshock, where $u_x \to 0$. When the flow compression becomes important ($u_x \neq 0$), their separation leads to the decrease of both the particle and wave energy densities toward the subshock. Note that for strongly nonlinear acceleration regimes, $p_* \ll p_{\max}$. To describe this mathematically, let us assume first that the relation in equation (16) between P and I is still a reasonable approximation, even if u_x is nonzero but small. Then, integrating equation (13) again between some $x > 0$ and $x = \infty$, instead of equation (17), we obtain

$$ug + u_1\frac{L}{g}\frac{\partial g}{\partial x} = -\frac{1}{3}\int_x^\infty u_x p\frac{\partial g}{\partial p}\,dx \ . \tag{21}$$

In contrast to the solution of equation (17), the length scale $L \equiv \kappa_0/2u_1 M_A$ enters the solution of this equation. This is because it has a nonzero right-hand side. In the next subsection we obtain a solution of this equation that rapidly changes on a scale $\sim L$.

3.1. *Internal Asymptotic Solution for g*

First we note that $L \ll L_C$, where $L_C = \kappa(p_{\max})/u_1$ is the total scale height of the CR precursor on which $u(x)$ changes. Next, in addition to x and p, we introduce a fast (internal) variable $\xi(x,\ p)$ as follows,

$$\xi = \frac{x - x_f(p)}{L} \ , \tag{22}$$

where $x = x_f(p)$ is some special curve on the $(x,\ p)$-plane that bounds the solution and is specified below. We rewrite equation (21) for

$$\xi \text{ fixed}, \quad L \to 0 \ . \tag{23}$$

Separating fast variable terms on the right-hand side by replacing

$$\frac{\partial g}{\partial p} \to \frac{\partial g}{\partial p} - L^{-1}\frac{\partial x_f}{\partial p}\frac{\partial g}{\partial \xi} \ ,$$

to the leading order in $L/L_C \to 0$, we obtain

$$u_f g + \frac{u_1}{g}\frac{\partial g}{\partial \xi} = \frac{1}{3}p\frac{du_f}{dp}(G - g) - \frac{1}{3}\int_x^\infty u_x p\frac{\partial G}{\partial p}\,dx \ . \tag{24}$$

Here we denote $u_f(p) \equiv u\big(x_f(p)\big)$ and

$$G(x,\ p) = \lim_{\xi\to\infty} g(\xi,\ x,\ p) \ . \tag{25}$$

The existence of this limit is confirmed on obtaining the solution of equation (24) below. First, we introduce the following notations:

$$w(p) = u_f + \frac{1}{3}p\frac{du_f}{dp} \ ,$$

$$S(x,\ p) = \frac{1}{3}p\frac{du_f}{dp}G - \frac{1}{3}\int_x^\infty u_x p\frac{\partial G}{\partial p}\,dx \ .$$

Equation (24) can then be rewritten as

$$wg + \frac{u_1}{g}\frac{\partial g}{\partial \xi} = S \ , \tag{26}$$

and its solution can thus be written as

$$g(\xi,\ x,\ p) = \frac{S(x,\ p)}{w(p) + e^{-S\xi/u_1}} \ . \tag{27}$$

One sees that the limit in equation (25) indeed exists and is equal to $G = S/w$. Furthermore, equation (27) describes a transition front on the particle distribution between its asymptotic value $g = G$ at $\xi \to \infty$ and $g = 0$ at $\xi \to -\infty$. This front solution results from particle losses caused by the lack of resonant waves toward the subshock, as we argued while discussing equations (19) and (20). Note that according to the ordering in equation (23), we should set $x = x_f(p)$ in $S(x,\ p)$ when solving equation (26) for $g(\xi)$, and we must indeed do it for $\xi \sim 1$ as well as for all negative $\xi < 0$. In the limit $\xi \to \infty$, however, we can use the result from equation (27) for arbitrary $x > x_f(p)$, since it remains valid in this case; however, it merely states that in this region, the complete solution is represented by its "external" part $G(x,\ p)$ (eq. [25]). This, in turn, is yet to be determined. Before we do

this in § 3.3, we should verify the validity of the internal solution and calculate its unknown function $x_f(p)$.

3.2. *Nonlinear Modification of the Internal Solution and Determination of $x_f(p)$*

The way that we resolved equation (14) for I (see eq. [16]) may become inadequate for two reasons. First, the second term on the left-hand side of equation (14) may become comparable to the first one. This problem could be resolved, in principle, by integrating this equation along the characteristic $u/p = \text{const}$ (instead of the x-axis, as we did to obtain eq. [16]), unless the Stoss term on its right-hand side alters the hyperbolic type of this equation. A matter of bigger concern is that the increase of u_x obviously has to do with strong shock modification, so that $P \sim 1$. Clearly, under these circumstances, the balance between the left-hand side of equation (14) and the pressure term on the right-hand side leads to impossibly large I. Evidently, the second term on the right-hand side must come into play before this has happened, so that the steady state will be maintained by the balance between this term and the pressure term, while the left-hand side will become subdominant. Thus, for I we have the following equation:

$$\frac{2u_1^2}{V_A}\frac{\partial}{\partial x}P - \mathrm{St}\{I\} = 0 \; . \tag{28}$$

As is often the case, we can assume that the wave collision term $\mathrm{St}\{I\} \propto I^2$, and in the long-wave limit $k \to 0$, it is also proportional to k^2 (which means to p^{-2}). The pressure gradient can be estimated as P/L_p, where L_p is the scale height of particles of momentum p, which we assume for simplicity to be proportional to p as in the standard Bohm case. Thus, for I we have

$$I^2 \simeq \frac{u_1}{\alpha V_A} pP \; , \tag{29}$$

where α characterizes the strength of nonlinear wave interaction. Using the last estimate, instead of equation (21), we have

$$ug + u_1 \frac{L_{\mathrm{nl}}}{\sqrt{g}}\frac{\partial g}{\partial x} = -\frac{1}{3}\int_x^\infty u_x p \frac{\partial g}{\partial p}\, dx \; , \tag{30}$$

where $L_{\mathrm{nl}} = \kappa_0 \alpha^{1/2}/u_1 M_A^{1/2}$. Introducing the fast variable ξ from equation (22) with L_{nl} instead of L and repeating the derivation in § 3.1, for g we obtain the following equation,

$$wg + \frac{u_1}{\sqrt{g}}\frac{\partial g}{\partial \xi} = S \; ,$$

with the obvious solution

$$g = \frac{S}{w}\tanh^2 \frac{\sqrt{wS}}{2u_1}\xi \; . \tag{31}$$

This solution is valid for $\xi \gtrsim \xi_0 > 0$ ($\xi_0 \sim G^{-1}$), whereas at $-\infty < \xi \lesssim \xi_0$, one should use the linear equation (16) for the wave spectral density and thus the solution in equation (27) instead of equation (31). The uniformly valid solution can also be obtained by using an interpolation between equation (16) and equation (29) for I. We do not need, however, the explicit form of the front transition in the particle distribution in the region $\xi \sim 1$, which means $x \approx x_f(p)$. We merely exploit the fact that this transition is much narrower (its width is $\Delta x \sim L_{\mathrm{nl}}/[G(x_f,\ p)]^{1/2}$) than that of the main part $G(x)$ in the interval $x_f < x < \infty$. The spatial scale of the latter is at least $\sim \kappa_B/u_1$, or even broader if the linear approximation in equation (18) can be used, in which case the length scale is determined by the linear damping of Alfvén waves (Drury et al. 1996).

The only characteristic of the above internal solution that is needed to calculate the external solution $G(x,\ p)$ is the position of the front transition in g on the $(x,\ p)$-plane, i.e., we need to calculate the function $x = x_f(p)$. To do this, we return to equation (20). We solve it by neglecting its left-hand side and finally arrive at the result for g and thus for I in equation (20) that contains the fast variable ξ. Generally, this produces large terms in the next order of approximation coming from the left-hand side. To avoid that, we must choose the position of the transition front [$\xi(x,\ p) = 0$] in such a way that it coincides with one of the characteristics of the operator on the left-hand side of equation (20), i.e.,

$$\left(u\frac{\partial}{\partial x} + u_x p \frac{\partial}{\partial p}\right)\xi(x,\ p) = 0$$

or

$$u_f(p) - p\frac{du_f}{dp} = 0 \; .$$

The choice of the concrete characteristic is based on the existence of the absolute maximum momentum $p_{\max}$, beyond which there are neither particles nor waves. That means

$$u_f(p) \equiv u\big(x_f(p)\big) = u_1 \frac{p}{p_{\max}} \; .$$

This formula shows how the converging flow $u(x)$ transforms the momentum cutoff at the periphery of the shock transition ($u \approx u_1$) to its reduced value $p_* \equiv p_{\max} u_0/u_1$ at $u = u_0$ (Fig. 2). Note that p_* is the momentum beyond which the effect of wave compression on the particle spectrum is significant. Finally, the function $x = x_f(p)$ is defined as

$$x_f(p) = u^{-1}\left(u_1 \frac{p}{p_{\max}}\right) \; .$$

3.3. *External Solution*

While having obtained the form and the position $x = x_f(p)$ of the narrow front in the particle distribution $g(x,\ p)$, we still need to calculate g to the right from the front where it decays with x (see Fig. 3). This would be the external solution $G(x,\ p)$ introduced in the previous subsections. It is clear that

$$\max_x g(x,\ p) \approx G(x_f,\ p) \equiv G_0(p) \; ,$$

so that from equation (21), we have the following equation:

$$u_f(p) G_0(p) = -\frac{1}{3}\int_{x_f(p)}^\infty u_x p \frac{\partial G}{\partial p}\, dx \; . \tag{32}$$

The most important information about $G(x,\ p)$ is contained in $G_0(p)$, for which from the last equation, we obtain

$$\frac{\partial}{\partial p} v(p) G_0(p) + 4\frac{u_1}{p_{\max}} G_0(p) = 0 \; , \tag{33}$$

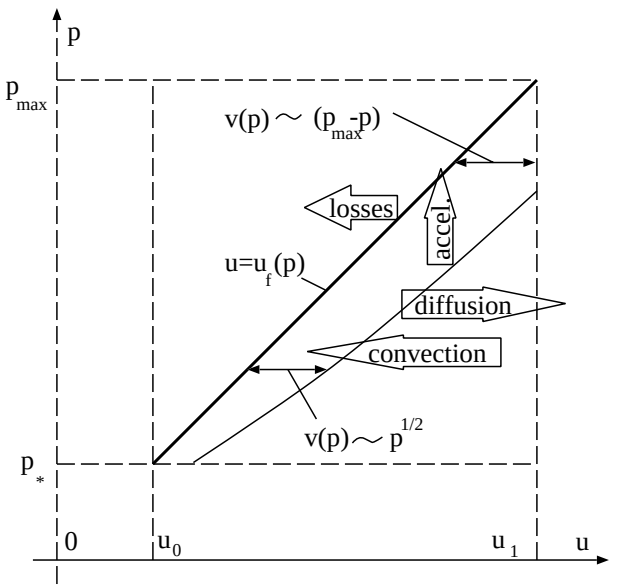

FIG. 3.—Phase plane of accelerated particles in the flow velocity–particle momentum coordinates. The particles are localized between the thick line (above which there are no resonant waves to confine them) and the thin line, where the particle density decays exponentially toward higher u (see text). The relevant transport processes are indicated by arrows.

where we have introduced $v(p)$ by

$$v(p) = \frac{1}{G_0(p)} \int_{u_f(p)}^{u_1} G(x,p)\, du(x) \ . \tag{34}$$

Equation (33) can easily be solved for G_0,

$$G_0(p) = \frac{Cu_1}{v(p)} \exp\left[-4\frac{u_1}{p_{\max}} \int \frac{dp}{v(p)}\right] , \tag{35}$$

where C is a normalization constant that should be determined from matching this solution with that in the region $p < p_*$. However, the function v depends on the solution itself. Fortunately, this quantity can be calculated prior to determining G_0, and therefore, this solution can be written in a closed form. To illustrate this, let us consider a particularly simple case of $p \simeq p_{\max}$ and then turn to the general case afterward. Clearly, $p \simeq p_{\max}$ means $u_f(p) \simeq u_1$. Evidently, we can replace G in equation (34) by G_0, so that for $v(p)$, we have $v(p) \simeq u_1 - u_f(p) = u_1(1 - p/p_{\max})$. Thus, from equation (35) we obtain the following shape of the cutoff near $p_{\max}$:

$$G_0 \simeq C(p_{\max} - p)^3 \ . \tag{36}$$

In the rest of the (x, p)-domain where $x > x_f(p)$ and p is not close to $p_{\max}$, we can assume that the CR diffusion coefficient is close to its Bohm value. Indeed, in contrast to the phase-space region $x \approx x_f(p)$, at each given (x, p) there are waves generated along the entire characteristic of equation (14) passing through that point of the phase space and occupying an extended region of the CR precursor (Fig. 2). We can then use the asymptotic high Mach number solution found in Malkov (1997):

$$g(x,\ p) = g_0(p) \exp\left[-\frac{1+\beta}{\kappa(p)} \int_0^x u\, dx\right] . \tag{37}$$

Here β is numerically small (typically $\simeq \frac{1}{6}$), and this solution without a β term manifests the balance between the diffusion and convection terms on the left-hand side of equation (13), which is a more accurate approximation far upstream where the flow modification (right-hand side) is weak. The flow profile depends on the form of $\kappa(p)$, and for $\kappa = \kappa_B \propto p$ in the internal part of the shock transition, $u(x)$ behaves linearly with x. Adopting this solution to the region $x > x_f$, we can write

$$G(x,\ p) = G_0(p) \exp\left(-\frac{1+\beta}{\kappa_B} \int_{x_f}^{x} u\, dx\right) ,$$

so that for v, we have

$$v(p) = \int_{u_f(p)}^{u_1} du \exp\left[-\frac{1+\beta}{\kappa_B} \int_{u_f}^{u} \frac{u'\, du'}{u_x(u')}\right] .$$

In the Bohm case, we can use the simplified linear approximation for $u(x)$ from Malkov (1997), $u = u_0 + u_1 x/L_C$, where $L_C = \pi\kappa_B(\hat{p})/2\theta u_1$, $\theta \approx 1.09$, and it is implied that the maximum contribution to the particle pressure comes from the momentum $p = \hat{p}$ (specified later). Now we can express v in the form of an error integral:

$$v(p) \simeq \int_{u_f}^{u_1} du \exp\left[-\frac{(1+\beta)L_C}{2u_1\kappa_B(p)}\left(u^2 - u_f^2\right)\right] . \tag{38}$$

The algebra further simplifies in two limiting cases (the second of which has already been mentioned):

$$v(p) = u_1 \begin{cases} \sqrt{\pi\kappa_B/2u_1(1+\beta)L_C} \ , & p \ll p_{\max} \ , \\ 1 - p/p_{\max} \ , & p \simeq p_{\max} \ . \end{cases}$$

This yields the following asymptotic behavior of $G_0(p)$:

$$G_0(p) = C \begin{cases} \sqrt{\hat{p}(1+\beta)/\theta p} \exp\left[-\dfrac{8}{p_{\max}}\sqrt{(1+\beta)p\hat{p}/\theta}\right] , & p \ll p_{\max} \ , \\ (p_{\max} - p)^3 \ , & p \simeq p_{\max} \ . \end{cases} \tag{39}$$

This result was obtained for particles with momenta $p \geq p_* \equiv p_{\max}/R = p_{\max}u_0/u_1$, whereas for $p < p_*$, we can use the spectra tabulated in Malkov & Drury (2001) for different $p_{\max}$, which should now be associated with $p = \hat{p}$. The matching of these two spectra should give the normalization constant C in the solution in equation (39). This is the subject of the next section.

3.4. *Connection with the Main Part of the Spectrum*

A typical solution of the nonlinear acceleration problem with a prescribed maximum momentum $p_{\max}$ calculated using the method of integral equations developed in Malkov (1997) and Malkov & Drury (2001) is shown in Figure 4 with the dash-dotted line. Since the influence of shock modification on the injection rate is not known for high $p_{\max}$ (see, however, Kang et al. 2001, where the values of $p_{\max} \sim 10$ have been reached), we have taken the injection rate $\nu \approx 0.1$, i.e., well inside the interval between the points A and C in Figure 1 (see § 2.1).

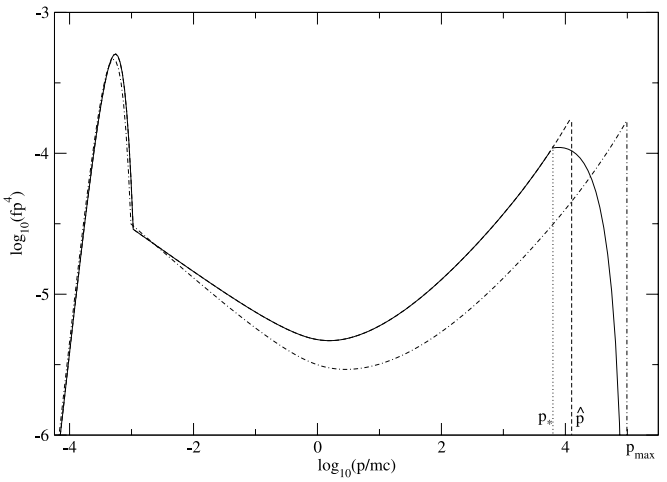

FIG. 4.—Particle spectra at a strong shock obtained from an analytic solution (Malkov 1997; Malkov & Drury 2001) for $M = 150$ (as in Fig. 1) and injection rate $\nu \approx 0.1$. The dash-dotted line shows the solution with the abrupt momentum cutoff at $p = p_{\max} = 10^5$. The spectrum drawn with the solid line demonstrates the effect of wave compression calculated using eqs. (35) and (38). The spectrum that would be obtained using the same technique as for the dash-dotted case but with the maximum momentum at $p = \hat{p}$ (see text) is shown by the dashed line.

To calculate an integral spectrum containing both the part modified by the wave compression in the shock precursor as well as its lower energy downstream part ($p \lesssim p_*$), we proceed as follows (see eq. [37] for the spatial structure of the spectrum). In the momentum range $p < p_* = p_{\max}/R$, we can obviously use the same method of integral equations. However, the role of maximum momentum is now played not by $p_{\max}$ but by $\hat{p}$ (i.e., a dynamical cutoff where the maximum contribution to the CR pressure is coming from). Furthermore, given ν and $\hat{p}$, we calculate the self-consistent flow structure with the precompression R shown in Figure 1 as well as the particle spectrum. The latter is shown in Figure 4 with the dashed line. Note that the spectrum matching point p_* with the cutoff area $p_* < p < p_{\max}$ is now determined, and we are ready to obtain the final spectrum by calculating its cutoff part from equations (35), (38), and (39). The spectrum is drawn with the full line. The momentum $\hat{p}$ can now be obtained as a maximal point of the function pG_0 (particle partial pressure per logarithmic interval) from equation (39) (top line), which yields

$$\hat{p} \approx \frac{p_{\max}}{8}\sqrt{\frac{\theta}{1+\beta}} \approx 0.1 p_{\max} .$$

It should be clear from our treatment that this formula is valid only when the shock is strongly modified, namely when $p_* \equiv p_{\max}/R < 0.1 p_{\max}$ (the case that we are interested in). Since R cannot exceed $M^{3/4}$, this means that the shock must be sufficiently strong, $M^{3/4} > 10$.

4. DISCUSSION

There are at least two reasons to believe that the standard acceleration theory may have estimated the maximum particle energy or the form of the spectrum below it incorrectly. The first reason is simply a possible conflict with the observations of TeV emission from SNRs, as we discussed in the § 1. The second reason is a theoretical one, that arises naturally from considering the nonlinear response of the shock structure to the acceleration that is exemplified in Figure 1. According to this picture, the response is so strong that it is unlikely that the acceleration can proceed at the same rate with no change in physics after such a dramatic shock restructuring (the precompression R may rise by 1–2 orders of magnitude depending on the Mach number). Time-dependent numerical simulations (e.g., Kang et al. 2002) show that the modifications occur very quickly, and compression is increased substantially even before $p_{\max} \sim 1$ [note that this would be consistent with the bifurcation diagram in Fig. 1 for initial $\nu \sim (c/u_1)n_{\rm CR}/n_1 \gtrsim 0.1$, where $n_{\rm CR}/n_1$ is the ratio of CR number density at the shock to that of the background plasma far upstream]. The shock modification, in turn, must follow rather abruptly after the maximum momentum has passed through the critical value. It was argued recently (Malkov et al. 2000) that this should drive crucial acceleration parameters such as the maximum momentum and injection rate back to their critical values, which must limit shock modification and settle it at some marginal level: the so-called self-organized critical (SOC) level (see also Jones 2001; Malkov & Drury 2001; Malkov & Diamond 2001 for more discussions of the critical interrelation between the injection, maximum energy, and shock structure). Mathematically, the SOC state is characterized by the requirement of merging the two critical points on the bifurcation diagram in Figure 1 into one inflection point on the $\nu(R)$ graph. Perhaps the most appealing aspect of this approach is its ability to predict the values of all three order parameters (injection rate, maximum momentum, and compression ratio) given only the control parameter (the Mach number), just from our knowledge of the nonlinear response $R(\nu, p_{\max})$ shown in Figure 1, and no further calculations.

However, the required back-reaction mechanisms on the injection and maximum momentum need to be demonstrated to operate. We have already discussed at a qualitative level how the injection rate is reduced by shock modification. The subject of this paper has been the reduction of particle momenta related to the formation of a spectral break at $p = p_*$, as a result of wave compression in a modified shock precursor. The position of the spectral break is universally related to the degree of system nonlinearity R, since $p_* = p_{\max}/R$. Hence, the problem seems to be converted to the study of nonlinear properties of the acceleration that are formally known from the analytic solution shown in Figure 1. It should be noted, however, that the injection rate ν required for accurate determination of the spectral break p_* through R, in the case of strongly nonlinear acceleration, can currently be obtained only from the SOC *Ansatz*. Another obvious restriction to our mechanism is that the significant reduction of p_* is not to be expected in oblique shocks, where the resonance relation $kp \propto B$ is approximately preserved because of the compression of B simultaneously with k. On the other hand, as we already mentioned in § 2.1, the nonlinear increase of the subshock obliqueness should strongly reduce the proton injection rate, which should ultimately reduce the density of the highest energy particles as well. It can also open the door to the preferential acceleration of electrons at the quasi-perpendicular portions of the subshock, since they may be injected rather efficiently there (see Malkov & Drury 2001, and references therein).

An equally important problem is that strong losses of particles between p_* and $p_{\max}$ must slow down the growth of $p_{\max}(t)$ due to the reduction of resonant waves. As we argued in § 2.1, this may result in an order of magnitude slower acceleration than one would expect from the standard Bohm diffusion paradigm. Consequently, the dynamically and observationally significant spectral break p_* may be at least 2 orders of magnitude below the maximum momentum $p_{\max}$ (again, depending on M) that could be reached in the unimpeded acceleration, which is normally implied in estimates of maximum energy achievable in SNRs over their active lifetime.

In addition to the above mentioned uncertainty in $p_{\max}(t)$, its relation to the position of the spectral break $p_* = p_{\max}/R$ also needs further clarification. Indeed, since R depends on a dynamical cutoff $\hat{p}$, which in general is linked to p_* and $p_{\max}$, the latter relation is still implicit. It can be easily resolved, however, in a supercritical regime [the saturated part of the $R(\nu)$ dependence in Fig. 1, see also Malkov & Drury 2001 for details], which requires[1] $\nu\hat{p}/p_{\rm inj} \gg M^{3/4}$. One then simply has $R \approx M^{3/4}$. As it was argued, however, the injection is unlikely to be high enough to reach this regime. An additional argument against it is that the spectral break becomes unrealistically small in the $M \to \infty$ limit, since $p_* = p_{\max}/M^{3/4}$. In the opposite case, $\nu\hat{p}/p_{\rm inj} \ll M^{3/4}$, the compression ratio saturates with M at $R \approx \nu\hat{p}/p_{\rm inj}$. Note that the injection rate must still be above critical; otherwise $R \approx 1$. Now we need to specify $\hat{p}$. The simple approximation used in the previous section yielded $\hat{p} \approx 0.1 p_{\max}$, so that $p_* \approx 10 p_{\rm inj}/\nu$ (independent of $p_{\max}$), which can be regarded as a lower bound on p_*. Indeed, the above relation between $\hat{p}$ and p_* can be applied only to the outermost part of the shock transition (see §§ 3.3 and 3.4). Downstream, the spectrum cuts off very sharply immediately beyond p_* (§ 3.1). Therefore, the dynamical cutoff $\hat{p} \approx p_*$, and we obtain the following upper bound on p_*, $p_* \approx \left[p_{\rm inj} p_{\max}(t)/\nu\right]^{1/2}$.

It should be clear that unless ν is dramatically reduced as a result of shock modification, even this upper bound places p_* way below $p_{\max}$. This may be the reason for the nondetection of protons at TeV energies in SNRs. Finally, this does not contradict the detection of 10–100 TeV electrons in, e.g., SNR 1006, since they may be accelerated by other mechanisms (e.g., Papadopoulos 1981; Galeev 1984; Bykov & Uvarov 1999; Laming 2001) or may have higher radiation efficiency.

We thank an anonymous referee for helpful suggestions. This study is supported through the University of California, San Diego, by US Department of Energy grant FG 03-88 ER 53275. At the University of Minnesota, this work is supported by NASA through grant NAG 5-8428 and by the University of Minnesota Supercomputing Institute.

[1] This is strictly valid for $\kappa(p) \propto p$.

REFERENCES

Achterberg, A., & Blandford, R. D. 1986, MNRAS, 218, 551
Achterberg, A., Blandford, R. D., & Reynolds, S. P. 1994, A&A, 281, 220
Aharonian, F. A., et al. 2001, A&A, 370, 112
Allen, G. E., Petre, R., & Gotthelf, E. V. 2001, ApJ, 558, 739
Atoyan, A. M., Aharonian, F. A., Tuffs, R. J., & Völk, H. J. 2000, A&A, 355, 211
Axford, W. I. 1981, in IAU Symp. 94, Origin of Cosmic Rays, ed. G. Setti, G. Spada, & A. W. Wolfendale (Boston: Reidel), 339
Axford, W. I., Leer, E., & McKenzie, J. F. 1982, A&A, 111, 317
Axford, W. I., Leer, E., & Skadron, G. 1977, Proc. 15th Int. Cosmic-Ray Conf. (Plovdiv), 11, 132
Bell, A. R. 1978, MNRAS, 182, 147
Bell, A. R., & Lucek, S. G. 2000, Astrophys. Space Sci., 272, 255
Bennett, L., & Ellison, D. C. 1995, J. Geophys. Res., 100, 3439
Berezhko, E. G., Yelshin, V., & Ksenofontov, L. 1996, Soviet Phys.–JETP, 82, 1
Berezinsky, V. S., & Ptuskin, V. S. 1989, ApJ, 340, 351
Blandford, R. D., & Ostriker, J. P. 1978, ApJ, 221, L29
Buckley, J. H., et al. 1998, A&A, 329, 639
Bykov, A. M., & Uvarov, Yu. A. 1999, Soviet Phys.–JETP, 88, 465
Donohue, D. J., Zank, G. P., & Webb, G. M. 1994, ApJ, 424, 263
Dorfi, E. 1990, A&A, 234, 419
Drury, L. O'C., Aharonian, F. A., & Völk, H. J. 1994, A&A, 287, 959
Drury, L. O'C., Duffy, P., & Kirk, J. K. 1996, A&A, 309, 1002
Drury, L. O'C., & Völk, H. J. 1981, ApJ, 248, 344
Duffy, P. 1992, A&A, 262, 281
Ellison, D. C., & Eichler, D. 1985, Phys. Rev. Lett., 55, 2735
Esposito, J. A., Hunter, S. D., Kanbach G., & Sreekumar, P. 1996, ApJ, 461, 820
Galeev, A. A. 1984, Soviet Phys.–JETP, 86, 1655
Gieseler, U. D. J., Jones T. W., & Kang H. 2000, A&A, 364, 911
Ivanov, A. I. 1978, Physics of a Strongly Nonequilibrium Plasma (in Russian; Moscow: Nauka)
Jones, T. W. 2001, in ASP Conf. Ser. 241, The 7th Taipei Astrophysics Workshop on Cosmic Rays in the Universe, ed. C.-M. Ko (San Francisco: ASP), 239
Jones, T. W., & Kang, H. 1992, ApJ, 396, 575
Kang, H., & Jones, T. W. 1990, ApJ, 353, 149
Kang, H., Jones T. W., & Gieseler, U. D. J. 2001, Proc. 27th Int. Cosmic-Ray Conf. (Hamburg), 5, 2088
———. 2002, ApJ, submitted
Kazanas, D., & Ellison, D. C. 1986, ApJ, 304, 178
Kirk, J. G., & Dendy, R. O. 2001, J. Phys. G, 27, 1589
Ko, C.-M. 1991, A&A, 251, 713
Lagage, P. O., & Cesarsky, C. J. 1983, A&A, 125, 249
Laming, J. M. 2001, ApJ, 546, 1149
Malkov, M. A. 1997, ApJ, 485, 638
———. 1998, Phys. Rev. E, 58, 4911
Malkov, M. A., & Diamond, P. H. 2001, Phys. Plasmas, 8, 2401
Malkov, M. A., Diamond, P. H., & Völk, H. J. 2000, ApJ, 533, L171
Malkov, M. A., & Drury L. O'C. 2001, Rep. Prog. Phys., 64, 429
Naito, T., & Takahara, F. 1994, J. Phys. G, 20, 477
Papadopoulos, K. 1981, in Plasma Astrophysics (ESA SP-161; Paris: ESA), 145
Sagdeev, R. Z., & Galeev, A. A. 1969, Nonlinear Plasma Theory (New York: Benjamin)
Skilling, J. 1975, MNRAS, 172, 557
Tanimori, T., et al. 1998, ApJ, 497, L25
Völk, H. J., Drury, L. O'C., & McKenzie, J. M. 1984, A&A, 130, 19
Zank, G. P., Rice, W. K. M., le Roux, J. A., Cairns, I. H., & Webb, G. M. 2001, Phys. Plasmas, 8, 4560