Gamma-ray burst theory after Swift

BY TSVI PIRAN1,* AND YI-ZHONG FAN1,2

1The Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel
2Purple Mountain Observatory, Chinese Academy of Science, Nanjing 210008, China

Afterglow observations in the pre-Swift era confirmed to a large extend the relativistic blast wave model for gamma-ray bursts (GRBs). Together with the observations of properties of host galaxies and the association with (type Ic) SNe, this has led to the generally accepted collapsar origin of long GRBs. However, most of the afterglow data was collected hours after the burst. The X-ray telescope and the UV/optical telescope onboard Swift are able to slew to the direction of a burst in real time and record the early broadband afterglow light curves. These observations, and in particular the X-ray observations, resulted in many surprises. While we have anticipated a smooth transition from the prompt emission to the afterglow, many observed that early light curves are drastically different. We review here how these observations are changing our understanding of GRBs.

Keywords: gamma-ray bursts; black holes; neutron stars; X-ray

1. Introduction

According to the current internal–external shocks model, the prompt gamma-ray burst (GRB) is produced by internal shocks within a relativistic outflow, while the afterglow is produced by a slowing down of the relativistic ejecta due to its subsequent interaction with the external medium. According to this model (see Piran 1999, 2004; Zhang & Meszáros 2004; Mészáros 2006, for reviews), the blast wave that forms due to this interaction takes the form of the Blandford–McKee self-similar profile (Blandford & McKee 1976). Relativistic electrons that are accelerated in this blast wave and circle around in the magnetic fields that are generated there produce the observed synchrotron emission.

Overall, the basic external shock afterglow is a simple theory characterized by five parameters: $E_0$, the total energy; $n$, the circumburst density; $p$, the power-law index that characterizes the electron distribution, $n(\gamma) \propto \gamma^{-p}$; and two equipartition parameters, $\epsilon_e$ and $\epsilon_B$, that define the energy density of the electrons and the magnetic field behind the shock in terms of the total energy density ($\epsilon_{e,B} = \epsilon_{e,B}^0$). Two additional parameters were introduced later: the jet opening angle, $\theta_j$ (Rhoads 1999; Sari et al. 1999), and the index, $k$, which characterizes the radial profile of the circumburst density: $n \propto r^{-k}$ (Chevalier & Li 2000). The value $k=2$ is expected if a stellar wind took place before the burst.

* Author for correspondence (tsvi@phys.huji.ac.il).

One contribution of 35 to a Discussion Meeting Issue ‘Gamma-ray bursts’.
This simple theory had many successes. Afterglow observations confirmed relativistic motion (Frail et al. 1997; Goodman 1997; Katz & Piran 1997; Waxman et al. 1998; Oren et al. 2004; Taylor et al. 2004). Multi-wavelength afterglow observations were well fitted (Panaitescu & Kumar 2002; Yost et al. 2003) by the predictions of the simple afterglow model (Sari et al. 1998). Jet breaks (Rhoads 1999; Sari et al. 1999) were seen in many bursts. The Frail relation (Frail et al. 2001; Panaitescu & Kumar 2002) that combined the prompt gamma-ray isotropic equivalent energy, $E_{\text{iso}}$, with the inferred opening angle, $\theta$, led to the conclusion that the total emitted energy $E_{\text{iso}}\theta^2/2$ varies from burst to burst much less than does $E_{\text{iso}}$ and that $\theta$ as inferred from jet breaks is indeed the true physical opening angle. The residual variation in $E_{\text{iso}}\theta^2/2$ correlates linearly with $E_{\text{peak}}^{3/2}$ (Ghirlanda et al. 2004). Most of the scatter in $E_{\text{iso}}$ is thus eliminated when both Frail and Ghirlanda relations are invoked. These relations, which link prompt quantity, $E_{\text{iso}}$, with a quantity, $\theta$, that is inferred from the afterglow, gave further supporting evidence for this model. Otherwise, why will we have such a conspiracy? An additional support arose from the fact that the kinetic energy inferred from the afterglow observation 10 h after the burst was comparable to the prompt gamma-ray energy. Again, one would have to come up with a cunning explanation as to why such an equality came about if the two quantities are not simply related. As we see later, this last observation is particularly challenging now with the Swift observations.

The external shock afterglow model has a very clear prediction of a smooth light curve. Angular smoothing plays a crucial role here (Fenimore et al. 1996; Sari & Piran 1997) and this has lead to the realization (Nakar & Piran 2003; Nakar et al. 2003) that neither angular nor radial fluctuations in the external density could give rise to a significant variability on a very short time-scale, $\delta t/t \ll 1$, in the light curve.

The internal–external shocks model deals with the emitting regions, but has strong implication on the central engine that accelerates the relativistic outflow. It requires a variable central engine that is active for the duration of the prompt burst, typically 100 s for a long burst. Accretion theory (Narayan et al. 2001) indicates that a long efficient accretion phase can arise only from a continuously fed disk. The source could be fallback during a supernova explosion. A process will arise naturally in the collapsar model (see Woosley & Bloom (2006) for a recent review). This theoretical conclusion is in agreement with the observations showing that GRBs take place in star-forming regions (see Fruchter et al. 2006) and the association of some long GRBs with type Ic SNe.

The Swift observations of the early X-ray afterglows of GRBs that took place during or immediately after the bursts have revealed many surprises, which have shaken this simple picture. As in other astronomical observations, it seems that real life is much richer and much more complicated. We discuss, in this short review, the implications of these observations to our understanding of GRBs.

1 Within the collapsar model, the central engine, as far as the emission regions are concerned, includes the propagation within the stellar envelope during which variability can be introduced (see Lazzati et al. 2007).
2. Pre-Swift predictions and other early hints

Even before Swift’s observations, it was realized that most probably this picture of a homogenous and uniform blast wave acting on its own might be an oversimplification. Among the suggestions that were made already in the late 1990s include the following.

— Prolonged activity of the central engine as a source of the afterglow was suggested as early as 1997 (Katz & Piran 1997; Katz et al. 1998). Unlike the external shock afterglow, such an emission was expected to be highly irregular. The smooth afterglow light curves detected before Swift ruled in favour of an external shock origin. Still the possibility that the central engine continues to be active at some level was never excluded.

— Refreshed shocks arise when a new shell collides with the blast wave that has been slowed down by the circumburst matter (Rees & Mészáros 1998; Kumar & Piran 2000). This could be a slow shell that was emitted simultaneously with the rest and caught up with the blast wave only after the latter has slowed down hours or even days after the burst. Alternatively, this could arise due to a prolonged activity of the inner engine that continues to eject fresh shells after the prompt emission has ceased.

— Energy injection and re-energetization of the blast wave due to the activity of the central engine (Cohen & Piran 1999; Sari & Mészáros 2000).

— Off-axis emission (sometimes referred to as ‘high-latitude emission’) will be observed as a fast decaying signal if the prompt emission stops abruptly (Fenimore et al. 1996; Kumar & Panaitescu 2000).

— A patchy shell structure would result in fluctuations in the observed light curve (Kumar & Piran 2000; Nakar & Oren 2004).

— A non-uniform jet with different Lorentz factors at different directions (Lipunov et al. 2001; Rossi et al. 2002; Berger et al. 2003; Huang et al. 2004; Peng et al. 2005).

3. The Swift afterglow light curves

A schematic X-ray afterglow light curve (figure 1) has been summarized by Nousek et al. (2006) and Zhang et al. (2006). The interesting features that emerge are (not all bursts show all phases):

— An early steep decline, $F_X \propto t^{-3.5}$, lasting for the first thousand seconds (Tagliaferri et al. 2005).

— A shallow, $F_X \propto t^{-1/2}$, decline of the X-ray afterglow lasting approximately $10^4$ s.

— A ‘normal’ decay phase that may include a jet break.2

— Energetic X-ray flares that arise in various stages of the afterglow. Flares were seen already by BeppoSAX (Piro et al. 1998a, 2005; Galli & Piro 2006) and even by ASCA (Yoshida et al. 1999). However, their real extent was realized only with Swift (Burrows et al. 2005; Falcone et al. 2006; O’Brien et al. 2006).

2 Interestingly, no jet break was seen before approximately $10^4$ s, setting a low limit on the opening angles of GRB jets.
The breaks in several light curves are not accompanied by spectral changes. Moreover, the breaks in the light curve are at times chromatic and breaks take place in the X-rays while the optical light curve decays as a smooth power-law.

(a) The rapid decline of the early X-ray telescope light curve

A large fraction of X-ray afterglows show an initial steep decay phase with $F_p \propto t^{-\alpha}$ with $\alpha \sim 3-5$ (Tagliaferri et al. 2005). A common interpretation of this phase is an off-axis emission coming from $\theta > \Gamma^{-1}$ (Fenimore et al. 1996; Kumar & Panaitescu 2000). If prompt gamma-ray stops abruptly on a fixed radius at time $t_0$, the off-axis emission observed from $\theta > \Gamma^{-1}$ is

$$F_{\gamma x} \propto \left(\frac{(t - t_0)}{\delta t}\right)^{-2-\beta-\delta/2},$$

(3.1)

where $\delta t$ is the angular spreading time of the last pulse; $\beta \sim 1$ is the spectral index of the emission; and $\delta$ is the possible angular dependence of the form, $\propto \theta^{-5}$, of the prompt emission.

If correct, this off-axis interpretation can be used to constrain the size of the emitting region to approximately $10^{14}-10^{15}$ cm (Lazzati & Begelman 2006; Lyutikov 2006) and possibly also the Lorentz factor approximately tens at the end of the prompt emission phase. However, there are several questions that have to be addressed before this interpretation is accepted and they are as follows.

Figure 1. A schematic of the X-ray light curve based on the Swift XRT data (Nousek et al. 2006; Zhang et al. 2006).
— Why does the prompt emission stop abruptly (or at least faster than $t^{-3}$) on a constant $R_0$ surface?

— According to the standard internal–external shocks model, the afterglow should begin before the end of a long burst (Sari 1997). An essential part of the interpretation of this phase as off-axis emission is a very weak initial afterglow. The question why is the initial afterglow so weak is related, of course, to the origin of the shallow decline phase, which we discuss in §3b.

— A more delicate question involves the exact shape of the light curve. A careful examination of the model (Fan & Wei 2005) shows that with several pulses, the overall off-axis emission declines (for $\delta = 0$) as

$$F_X = \sum_i F_{x,i} \left[ \frac{(t-t_{0,i})}{\delta t_i} \right]^{-(2+\beta_i)} ,$$  

where $i$ represents the $i$th pulse. Such a decline is much steeper (figure 2) than $(t/t_{\text{turn}})^{-(2+\beta)}$ as long as $t_{\text{turn}} \gg \max \{\delta t_i\}$, where $t_{\text{turn}}$ is the turning-off time of the internal shocks.

Note that the alternative explanations for the steep decline such as a dying central engine (Fan & Wei 2005) or emission from a hot cocoon (Pe’er et al. 2006) answer some, but not all, questions raised above.

(b) The shallow decline phase

In about half of Swift GRBs, the X-ray light curves display a long shallow decline with $F_{\nu} \propto t^{-1/2}$ (figure 1) lasting from approximately 1000 to 10 000 s. No spectral evolution is seen before and after the ‘shallow-to-normal’ transition that takes place at the end of this phase.
Since at this stage, the X-ray is above the cooling frequency, we have the simple relation (Kumar 2000; Freedman & Waxman 2001; Granot et al. 2006)

\[
\frac{\epsilon_x E_{k,iso}}{tF_x(t)} = 4\pi d_L^2(1+z)^{\beta-\alpha-1},
\]

where \(\epsilon_x\) is the efficiency of converting the (isotropic equivalent) kinetic energy, \(E_{k,iso}\), to X-ray; \(F_x(t)\) is the observed X-ray flux; \(d_L\) is the luminosity distance; and \(\beta\) and \(\alpha\) are the spectral index and the temporal slope in the X-ray range, respectively. Since \(tF_x(t)\) increases with time and as the r.h.s. is a constant, either \(E_{k,iso}\) (energy injection) or the overall efficiency of converting the kinetic energy to X-rays, \(\epsilon_x\), increases with time.

Energy injection (Nousek et al. 2006; Zhang et al. 2006) in the form of a slow material with a Lorentz factor approximately tens (Granot & Kumar 2006), which might be moving on the sides of the faster-moving material (Rees & Mészáros 1998; Peng et al. 2005), is a common interpretation of this phase. However, the required additional energy is much larger, by about a factor of 10, than the kinetic energy inferred at the beginning of the afterglow (and the end of the prompt phase). This has raised concerns that the prompt efficiency (of converting the initial energy to gamma-rays) is unreasonably large (Ioka et al. 2006; Nousek et al. 2006; Zhang et al. 2006). However, a detailed analysis (Fan & Piran 2006a; Granot et al. 2006) has shown that Lloyd-Ronning & Zhang (2004), on whose results this conclusion was based, have underestimated \(E_{k,iso}\) by a large factor. With a correct estimate of \(E_{k,iso}\), the problem is not so severe. A related puzzle is why is the total kinetic energy (during the late ‘regular’ phase) comparable to the prompt gamma-ray energy.

Another alternative is a variation in the efficiency of producing X-rays. The idea is that this efficiency is particularly low initially and it increases to normal values later. One possibility is that the microscopic parameters are varying (Yost et al. 2003; Fan & Piran 2006a; Ioka et al. 2006; Panaitescu et al. 2006). One can construct ad hoc models that assume variability of \(\epsilon_{e,B}\) with the Lorentz factor, \(\Gamma\), that fits the data. As the Lorentz factor of the blast wave varies from approximately 100 to 30 during the first \(10^4\) s, one can expect that \(\epsilon_e\) or \(\epsilon_B\) would vary. This can arise from a slow build up of the magnetic field or inefficient electron acceleration by collisionless shocks with extremely large Lorentz factor. Both the effects would reduce the efficiency of X-ray production. However, it is difficult to accept such models without a clear physical model for the microscopic processes that lead to the variation. Furthermore, at present, there is not even an
explanation to either the time-scale, approximately $10^4$, or the value of the Lorentz factor, $\Gamma \sim 10$, for which these variations stop. A variant on this theme, which does not involve changes of the intrinsic parameters, is the suggestion that the circumburst density in the region close to the burst (Granot et al. 2006) or along the line of sight (Eichler & Granot 2006) is extremely low and the blast wave simply cannot dissipate enough energy in this extremely low-density region (resulting in a very low $\varepsilon_X$). Again, such a model has to be studied carefully before it is accepted.

An intriguing alternative is that another process takes place during this shallow X-ray decline phase and this process cools the X-ray emitting electrons suppressing their synchrotron X-rays emission. One has to speculate further that this extra cooling process gradually becomes ineffective as the afterglow turns to its normal phase. Inverse Compton cooling, which is important if $\varepsilon_e > \varepsilon_B$ (Sari et al. 1996), could have been such a process and the Klein–Nishina cut-off could have been the mechanism that switches off the inverse Compton cooling. While it is clear that one can easily achieve sufficient cooling with inverse Compton, it is not clear whether the Klein–Nishina transition is rapid enough to make the relevant difference (Narayan et al. in preparation). One has to keep searching for another mechanism.

Another drastic solution for the inconsistency between the X-ray and the optical light curves is that they should be attributed to different physical processes. For example, when proposing the late internal shocks, Fan & Wei (2005) speculated that in some GRBs, the X-ray afterglow might be dominated by prolonged activity of the central engine, whereas the IR/optical afterglow might be dominated by the external shock emission. Such a model, seeming to be artificial, has been supported by the multi-wavelength observations of the afterglow of GRB 060218 (Campana et al. 2006; Soderberg et al. 2006b). In this event, it is clear that the irregular optical emission does not arise from a regular blast wave. However, both the weak radio and the bright X-ray emissions decay as a power-law and seem to be normal. The X-ray afterglow, however, cannot result from an external shock owing to its very steep spectrum. Furthermore, the kinetic energy required to power such an external shock would produce far too much radio. Instead, the X-ray emission might be attributed to a prolonged activity of the central engine (Fan et al. 2006; Soderberg et al. 2006a).

(c) The X-ray flares

Energetic X-ray flares have been detected in several pre-Swift GRBs (Piro et al. 1998b; Yoshida et al. 1999; Piro et al. 2005; Galli & Piro 2006) and in about a half of Swift GRBs (Burrows et al. 2005; O’Brien et al. 2006). It was thought earlier that these flares might be related to a delayed onset of the forward shock (Piro et al. 2005; Galli & Piro 2006). However, this is found to be unlikely (Kobayashi & Zhang submitted). Refreshed shocks might be able to account for some events (Granot et al. 2003; Guetta et al. 2006; Panaitescu 2006). However, refreshed shocks generally result in an increase in the kinetic energy of the afterglow after a flare (as seen for example in GRB 030329; Granot et al. 2003), whereas in some of the observed flares (e.g. GRB 050502B; see Falcone et al. (2006) for details), the emission return exactly to the same power-law decline after the flare. Further evidence for a different origin arise from the very steep decline of some X-ray flares,
which cannot be explained within an external shock origin. The multi-flares detected in GRB 050730 (e.g. Starling et al. 2005), the weak optical emission simultaneous with most X-ray (Burrows et al. 2005) and the huge energy released in the giant flare of GRB 050502b provide additional evidence against external shock origin. All of these observations point towards a reactivation of the inner engine (Katz & Piran 1997; Katz et al. 1998) and the production of late internal shocks (Burrows et al. 2005; Fan & Wei 2005; Zhang et al. 2006).

This interpretation, that seems unavoidable at present, seems to have the most important implication on the central engine. The central engine must be active over much longer time-scales. Instead of being active for 100 s, it should be active for $10^4$ s. One natural speculation is that this activity is caused by an intermittent fallback accretion (Perna et al. 2006) or fragmentation during the collapse and the subsequent accretion (King et al. 2005). There are other possibilities like instability in an MHD accretion (Proga & Zhang 2006). There are fewer options for the X-ray flares following short GRBs (e.g. Barthelmy et al. 2005), as, there, one needs a much larger increase in the relevant time-scales. One can expect here accretion of debris ejected during a merger (Faber et al. 2006).

An interesting implication of these flares is that the copious flare photons can be inverse Compton up-scattered by hot electrons within the blast wave (figure 3). Regardless of the question whether the flare photons are produced by the central engine or within the blast wave itself, this can produce a detectable flare in the GeV–TeV (Wang et al. 2006; Galli & Piro in press) or sub-GeV region (Fan & Piran 2006b).

If the UV/X-ray flares are produced within the inner engine, the up-scattered photons are de-collimated and their arrival time is affected by the spherical curvature of the blast wave. The duration of the high-energy emission, thus, can be estimated as $T_{\text{high}} \sim \text{a few} \times t_{\text{flare}} \gg \Delta T$, where $t_{\text{flare}}$ is the time at which the flare takes place and $\Delta T$ is the duration of the far-UV/X-ray flare. The duration increases.
when the anisotropic radiation of the up-scattered photons has been taken into account, because now the strongest emission is from $\theta \sim 1/G$ (Fan & Piran 2006b). As a result, most of the up-scattered photons will arrive after the far-UV/X-ray flare. This lagging behaviour, in principle, is a signature of up-scattering of internal shock radiation in the external blast wave, which may be tested by observations in the foreseeable future (Fan & Piran 2006b). If the flare photons are also generated at the forward shock front, the high-energy emission should be simultaneous with the soft X-ray photons (Galli & Piro in press).

4. The implication of the Swift gamma-ray burst afterglows

Swift’s observations pose new challenges to the theory of GRBs. The early afterglow is much more complicated than the late afterglow. We do not understand at present what the origin of this complicated activity is. Many speculations and many suggestions have been put forward but none is compelling. Among the various puzzles, two questions stand out as most promising to lead to interesting implications: (i) the nature of the slow decline phase and (ii) the origin of the X-ray flares.

The interpretation of the slow decline phase, in particular for the afterglows showing chromatic breaks, is not clear. The phenomenological approach, which assumes that the shock parameters are shock strength dependent, the density is radius dependent and the energy of the fireball increases with time, may be able to reproduce the observations but it lacks, at present, any physical basis. The simplest interpretation is that some cooling mechanism suppresses the emission of X-ray at early times. To verify this suggestion, we have to look, theoretically, for such a mechanism and search observationally for other channels in which this energy might be emitted.

The X-ray flares indicate that the central engine must be active on a longer time-scale. It has been suggested long ago (Katz & Piran 1997; Katz et al. 1998) that persistent afterglow emission could be attributed to either an external forward shock (normal afterglow) or to a prolonged activity of the central engine (central engine afterglow). The agreement of the predictions of the external shock model (Sari et al. 1998b) with most multi-wavelength afterglow observations and in particular the smooth light curves seen in most afterglows discovered before Swift led to the understanding that afterglows are produced by external shocks (Piran 1999, 2004 for reviews). The observed energetic highly variable flares in numerous GRBs require a prolonged activity of the central engine. The inconsistency of the X-ray afterglow flux with the radio afterglow and the very steep X-ray telescope spectra seen in GRB 060218 suggest a central engine afterglow for this burst (Fan et al. 2006).

In contrast to what was expected before Swift, it is evident now that at times the role of the central engine has to be considered when interpreting GRB afterglows. Several features emerge: (i) the way the central engine turns-off influences the rapid decline phase, (ii) the central engine could reactivate after the end of the prompt ray phase, and (iii) the prolonged activity of the central engine may give rise to an observed power-law decaying (X-ray) afterglow in some cases.
This prolonged activity poses another challenge to the theory of inner engine. The activity lasting $10^4$ or even longer is a factor of approximately $10^8$ times the dynamical time-scale of the inner engine. This is not a drastic revision from the factor of $10^6$ based on the prompt emission alone. However, one needs to understand why the early activity during the prompt is different from the later one. This prolonged activity is particularly puzzling for short bursts.

References


