9 Prompt Emission Physics

The prompt γ-ray emission was the earliest detected signal from GRBs. However, after five decades, its origin is still subject to intense debate. The uncertainties lie in several *open questions* related to GRB jets, as discussed in the beginning of Chapter 7. In particular, for GRB prompt emission, theorists have been struggling to give correct answers to the following three questions:

- *What?* What is the composition of a GRB jet? Is it predominantly composed of matter (baryons and leptons) or a Poynting flux?
- *Where?* Various arguments (§9.2) suggest that GRB prompt emission should be *internal*, i.e. between (and including) the photosphere radius ($R_{ph} \sim 10^{11}$–10^{12} cm) and the deceleration radius ($R_{dec} \sim 10^{16}$–10^{17} cm). This range is 5–6 orders of magnitude in distance from the GRB central engine. Models invoking different jet compositions have different emission radii.
- *How?* How is GRB emission produced? This includes how the kinetic or Poynting flux energies get dissipated and converted to the internal energy (e.g. via shocks or magnetic reconnection), how the particles are accelerated (first-order or second-order Fermi acceleration) in shocks, reconnection sites, or turbulent regions in the outflow, and how the photons are radiated (synchrotron, SSC, or Comptonized thermal photons).

This chapter is dedicated to the prompt emission physics. It starts with a brief introduction (§9.1) to the problem, summarizing the data as well as the challenges one faces in interpreting the data. Next (§9.2), the general arguments for an internal origin of GRB prompt emission and a list of observational constraints on R_{GRB} are presented. Various GRB prompt emission models are then discussed in detail, including the non-dissipative photosphere emission model (§9.3), the dissipative photosphere models (§9.4), the internal shock model (§9.6), optically thin magnetic dissipation models in general (§9.7), and the ICMART model in particular (§9.8), as well as several other prompt emission models (§9.9). The role of electron–positron pairs is discussed generally in §9.5. A critical comparison among various theoretical models on their ability to account for the key observational facts is presented in §9.10.

Unlike previous chapters, this chapter contains many topics that are still subject to heavy debate as of the writing of this book. Even though the physics of various models is presented objectively, the commentaries on various models are inevitably subjective. The composition of GRB jets plays an important role in developing various models. The author's opinion is that there is a distribution of jet composition among GRBs, ranging from matter-dominated fireballs in some cases, to Poynting-flux-dominated outflows in

some others, and to hybrid jets in most cases. The dominant Band-function component observed in GRBs may have different origins in different bursts. The narrow Band-function component observed in a fraction of GRBs (e.g. GRB 090902B) is of a photosphere origin. However, the dominant Band-function component as observed in most GRBs is likely of a synchrotron radiation origin, which is from an optically thin site at a large distance from the central engine. The presentation of this chapter attempts to make an argument for such a picture, but is inevitably influenced by such an opinion. It is worth mentioning that there are strong opinions in the community that essentially all GRBs can be interpreted within one particular model within the matter-dominated fireball context, e.g. the dissipative photosphere model (for some authors) or the internal shock model (for some other authors).

9.1 What Do We Interpret? Why So Difficult?

The rich observational properties of GRB prompt emission have been summarized in §2.1. Here we highlight the key observational facts and explain why they pose great challenges to theorists.

- GRB lightcurves (§2.1.2) are irregular, and do not show characteristic time scales. The power density spectra (PDS) of GRBs suggest a self-similar behavior, and the lightcurves show possible superpositions of fast and slow temporal components.
- The dominant spectral component (§2.1.3) is a Band-function component with a typical low-energy photon index $\alpha \sim -1$. The peak energy E_p is typically around several hundred keV, but it can be as low as several keV and as high as ~ 15 MeV. Evidence of superposition of a quasi-thermal (blackbody) component on the non-thermal Band component has been observed in at least some GRBs.
- Combining spectral and lightcurve information, one has two well-defined observational features: one is the so-called "spectral lag", i.e. lightcurves in the softer bands are typically broader and lag behind the lightcurves in the harder bands; the other is the E_p evolution patterns, i.e. E_p shows either "hard-to-soft evolution" or "intensity tracking" with respect to the *broad pulses* in the lightcurves.
- In broader energy bands, prompt emission in optical, X-rays, and high-energy (GeV) γ-rays has been observed in at least some GRBs, which are temporarily (roughly) correlated with the sub-MeV emission (the GRB itself) at least in some GRBs. An early X-ray steep decay phase seems to signify the end of the prompt emission phase.
- X-ray flares are commonly observed in a good fraction of GRBs, and are the continuation of prompt emission in the softer energy bands. The temporal and spectral properties of X-ray flares seem to be generally consistent with those of prompt emission γ-rays.
- There are several interesting (even though not very tight) correlations. In general, intrinsically more luminous and more energetic GRBs tend to be harder.
- Bright GRBs seem to be strongly polarized in γ-rays (a point that is subject to confirmation by future more sensitive γ-ray polarimeters).

- At the time writing (mid 2018), no high-energy neutrinos had been detected temporally or spatially coincident with any GRB.

The difficulty in understanding prompt emission mainly lies in that no known theoretical models can *straightforwardly* interpret all the observational data collected so far. The most challenging problem is the low-energy photon index α, which has a typical value of -1 for long GRBs, and has a wide distribution (Fig. 9.1). Regardless of the energy dissipation mechanism and emission site, leading radiation mechanisms include synchrotron radiation and thermal Comptonization. The straightforward predictions from these models all deviate from the typical $\alpha \sim -1$ value:

- Synchrotron radiation is usually in the fast cooling regime during prompt emission (Ghisellini et al., 2000; Kumar and McMahon, 2008; Zhang and Yan, 2011). According to the standard fast cooling spectrum (Sari et al., 1998), the F_ν spectral index is $-1/2$, which corresponds to $\alpha = -3/2$. This is too soft compared with the observation.
- If one introduces a slow heating model so that electrons have a steady distribution without significant cooling, then the F_ν spectral index below the minimum synchrotron frequency is $1/3$, which corresponds to $-2/3$. It is interesting to note that the observed typical value -1 is enclosed between $(-3/2, -2/3)$, suggesting that synchrotron radiation may be a relevant mechanism. However, within the synchrotron model, no burst is
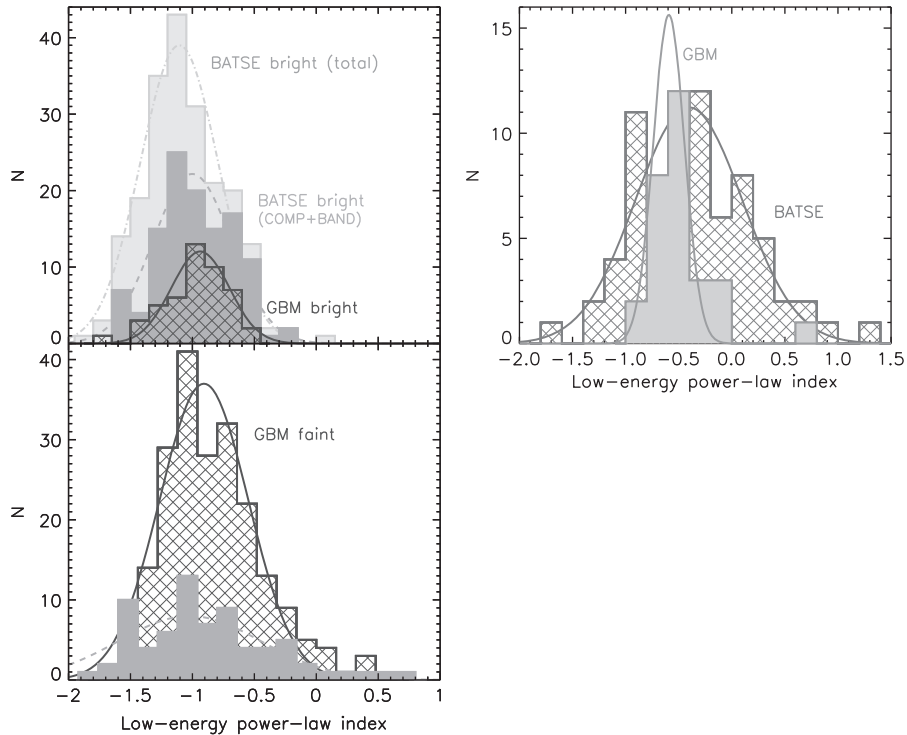

Figure 9.1 The distribution of the low-energy photon index α of Fermi/GBM and BATSE long (*left*) and short (*right*) GRBs. From Nava et al. (2011a).

expected to have an α larger than $-2/3$. This limit is called the synchrotron *line of death* (Preece et al., 1998). Observationally, a fraction of GRBs have an α harder than $-2/3$ (Fig. 9.1 lower panel), suggesting that the synchrotron radiation mechanism alone may not interpret all observations.

- The thermal Comptonization model that is relevant to photosphere emission, on the other hand, predicts a spectrum that is too hard to interpret the data. The standard Rayleigh–Jeans blackbody spectrum has $\alpha = +1$, which is very different from the observed typical value -1. Considering the relativistic equal-arrival-time surface effect and superposition of emission from a continuous wind, the spectrum can be softened to $\alpha \sim +0.4$ (Beloborodov, 2010; Deng and Zhang, 2014b), but it is still too hard for interpreting the data.

- A superposition model invoking both the thermal and the non-thermal components has been proposed in the pre-*Fermi* era, both theoretically (Mészáros and Rees, 2000b) and observationally (Ghirlanda et al., 2003; Ryde, 2005). In the *Fermi* era, at least some GRBs are found to have these superposed components (Guiriec et al., 2010, 2011; Axelsson et al., 2012; Guiriec et al., 2013). These observations suggest that more complicated models invoking multiple observational sites and radiation mechanisms may be needed to interpret GRB spectra.

Besides the α problem, more puzzles arise when the temporal and spectral properties are considered jointly. In particular, the two types of E_p evolution patterns (hard-to-soft evolution and intensity tracking) cannot easily be interpreted with any simple version of the models. Making things more complicated, both patterns sometimes co-exist in different episodes of the same burst (Lu et al., 2012). All these suggest very complicated physical processes at play in producing GRB prompt emission.

9.2 Constraints on R_{GRB}

Before discussing explicit models of GRB prompt emission, it is illustrative to discuss several general observational constraints on the distance of GRB prompt emission from the central engine, R_{GRB}.

9.2.1 Internal vs. External

In the early years, both the external forward shock (Rees and Mészáros, 1992; Mészáros and Rees, 1993b) and internal shocks (Rees and Mészáros, 1994) were proposed as the site of GRB prompt emission. After the discovery of the afterglow, the external shock was identified as the afterglow emission site. The location of the prompt emission was still not settled.

One main observational constraint is the rapid variability observed in GRB lightcurves. While internal shocks can naturally give rise to rapid variabilities (Kobayashi et al., 1997; Maxham and Zhang, 2009), within the external shock framework, rapid variability can arise only if the ambient density is clumpy (Dermer and Mitman, 1999). Figure 9.2 shows
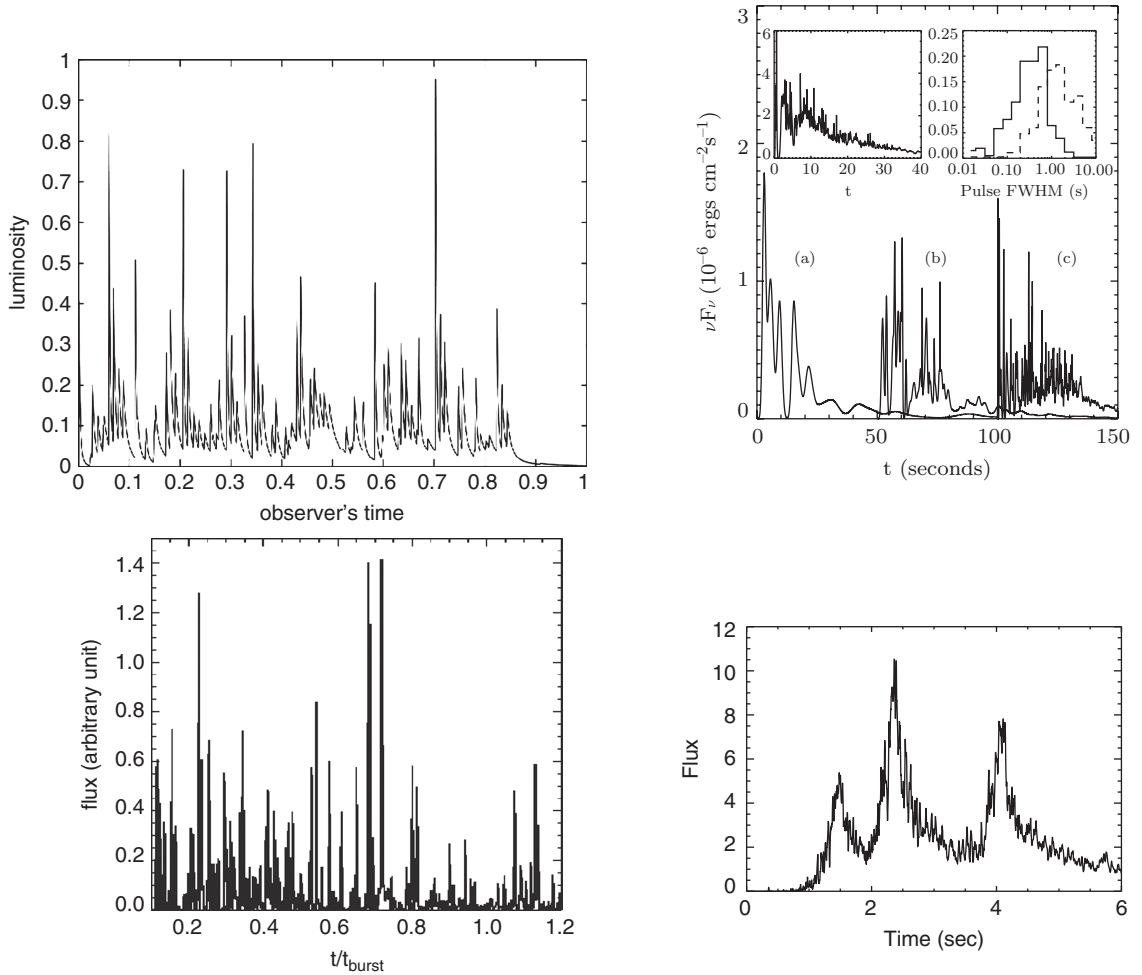
Simulated GRB lightcurves in different theoretical models. From upper left to lower right: internal shock model (reproduced from Figure 2a in Kobayashi et al. (1997) with permission. ©AAS.); external shock model with density clumps (reproduced from Figure 2 in Dermer and Mitman (1999) with permission. ©AAS.); turbulence model (from Narayan and Kumar (2009)); and ICMART model (from Zhang and Zhang (2014)).

the simulated GRB lightcurves in several different models, including the internal shock model (upper left, Kobayashi et al. 1997) and the clumpy external shock model (upper right, Dermer and Mitman 1999).

Sari and Piran (1997) pointed out that, in order to produce rapid variable lightcurves as observed in most GRBs, the external shock model has to invoke clumps with small sizes, so that the covering factor of these clumps within the entire external shock blastwave is small. This gives a very low efficiency for converting the kinetic energy of the blastwave to prompt γ-ray emission. The low efficiency is inconsistent with the afterglow modeling that infers an amount of afterglow kinetic energy comparable with the observed γ-ray energy (e.g. Panaitescu and Kumar, 2001, 2002). Dermer and Mitman (1999) argued that

a reasonable efficiency may be achieved within this model if the number of small clouds is large and if each cloud has a large column density. However, this requires contrived conditions.

The definite clue came from the *Swift* early X-ray afterglow observations. A steep decay phase commonly observed in GRBs is consistent with the tail of prompt emission (Tagliaferri et al., 2005; Barthelmy et al., 2005b), which later breaks to a shallow decay phase consistent with an external shock origin. This suggests that the afterglow emission is detached from the prompt emission region. Since an external shock origin of the afterglow is well established, this points directly to an internal origin of GRB prompt emission for most GRBs (Zhang et al., 2006).[1]

In general, for the majority of GRBs, the GRB prompt emission should be "internal", i.e.

$$R_{\mathrm{ph}} \leq R_{\mathrm{GRB}} < R_{\mathrm{dec}}. \tag{9.1}$$

9.2.2 Angular Spreading Time

One important time scale for spherical relativistic jets is the angular spreading time scale. Let us consider a spherical relativistic outflow (or a conical jet with $\theta_j \gg 1/\Gamma$) with constant Lorentz factor and *uniform* and *isotropic* emissivity in the comoving frame. Since the emission is beamed within the $1/\Gamma \ll 1$ cone, when the emission from the jet ceases suddenly everywhere for some reason (e.g. end of shock crossing), the bright emission fades within a time scale

$$t_{\mathrm{ang}} = \frac{R_{\mathrm{GRB}}}{c} \left[1 - \cos\left(\frac{1}{\Gamma}\right) \right] \simeq \frac{R_{\mathrm{GRB}}}{2\Gamma^2 c}. \tag{9.2}$$

This is the *angular spreading time scale*, which is the *smallest* time scale an observer would observe for a spherical jet.[2]

For an observed variability time scale δt, a comoving-frame uniform, isotropic jet with spherical/conical geometry should satisfy

$$\delta t \geq t_{\mathrm{ang}}, \tag{9.3}$$

where the $>$ sign applies when the variability is defined by the central engine activity rather than the angular spreading scale. This gives

$$R_{\mathrm{GRB}} \leq 2\Gamma^2 c \delta t \simeq R_{\mathrm{IS}}, \tag{9.4}$$

where R_{IS} is the internal shock radius. So for a uniform, isotropic, and spherical/conical jet, the GRB emission radius is allowed to be smaller than or at most equal to the internal shock radius R_{IS} defined by δt. The case of $R_{\mathrm{GRB}} < R_{\mathrm{IS}}$ is relevant to the photosphere model (see §7.3.2). Within this scenario, the variability in the observed lightcurves is

---

[1] Some bursts show a smooth, FRED-like, single-pulse lightcurve. It is still possible that these GRBs do not have an internal dissipation phase, with the emission produced from the external shock during the onset of the afterglow (Huang et al., 2018).

[2] Notice that the emission can last longer than t_{ang} all the way until the high-latitude angle reaches θ_j, but the flux drops steeply with the curvature effect relation (e.g. $\alpha = 2 + \beta$ or steeper), as discussed in §3.4.4. For γ-ray lightcurves that are usually displayed with linear scales in flux, t_{ang} is the typical time scale for defining the pulse width.

completely defined by the variability at the central engine. The ups and downs in the lightcurve correspond to the increases and decreases of the jet power as launched from the central engine.

The emission radius R_{GRB} may be greater than R_{IS} defined by δt, only if the uniform, isotropic, or spherical/conical assumption is broken. One may consider two examples. The first scenario is to invoke small blobs in the external medium so that the emission region (external forward shock) is no longer uniform. This scenario (Dermer and Mitman, 1999) has $R_{\mathrm{GRB}} = R_{\mathrm{FS}} \gg 2\Gamma^2 c\delta t$. However, this external shock model is disfavored in interpreting prompt emission for various reasons (Sari and Piran, 1997; Zhang et al., 2006).

The second scenario is to invoke an internal, non-uniform emission scenario. This is realized in magnetic dissipation models that invoke a (moderately) high-σ outflow. Forced reconnections due to internal collisions (Zhang and Yan, 2011) or current instabilities (Lyutikov and Blandford, 2003) would induce local Lorentz-boosted regions or "mini-jets" within the bulk jet. Such a "jets-in-the-jet" scenario would break the uniformity assumption, giving rise to smaller δt values than t_{ang} defined by R_{GRB}, i.e. $R_{\mathrm{GRB}} \gg 2\Gamma^2 c\delta t$. The γ-ray efficiency and the lightcurve shape of this model depend on the so-called covering factor, i.e. the number of mini-jets per unit time per unit volume. For example, the relativistic turbulence models studied by Narayan and Kumar (2009) and Lazar et al. (2009) produce very spiky lightcurves that seem not consistent with the observational data (lower left panel of Fig. 9.2). This is because the mini-jet covering factor is too small, i.e. at any instant, at most one mini-jet enters the observer's field of view. These models would also suffer from the efficiency difficulty faced by the clumpy external shock model. The ICMART model (Zhang and Yan, 2011) invokes a runaway generation of mini-jets, so that at any instant there could be many mini-jets contributing simultaneously to the observed flux. The superposed emission from all the mini-jets would give rise to the superposed slow and fast variability components (Zhang and Zhang 2014, lower right panel of Fig. 9.2), showing more consistency with the observational data. Within this scenario, the angular spreading time scale of the slow component (δt_{slow}) defines the emission radius, i.e. $R_{\mathrm{GRB}} \simeq 2\Gamma^2 c\delta t_{\mathrm{slow}}$.

In summary, the angular spreading time scale carries important information about the GRB emission site, and may be used to infer R_{GRB}. However, the dependence of R_{GRB} on δt is model dependent. In particular, one has

- $R_{\mathrm{GRB}} < 2\Gamma^2 c\delta t$: photosphere models;
- $R_{\mathrm{GRB}} \sim 2\Gamma^2 c\delta t$: internal shock models;
- $R_{\mathrm{GRB}} > 2\Gamma^2 c\delta t$: jets-in-the-jet models and clumpy external shock model.

As a result, one needs additional criteria to constrain R_{GRB}, as discussed below.

9.2.3 X-ray Steep Decay

When the prompt emission phase is observed in the X-ray band, as has been done in some GRBs that are long enough for *Swift*/XRT to catch the end of the prompt emission, the decay phase of the last episode of prompt emission can be followed to a much deeper level, and usually a steep decay phase is observed (Tagliaferri et al., 2005; Barthelmy et al.,

2005b). If the decay phase is controlled by the high-latitude curvature effect (§3.4.4), then the observed duration of the decay time depends on the emission radius R_{GRB} and the jet opening angle (Zhang et al., 2006):

$$t_{\mathrm{tail}} \leq (1+z) \left(\frac{R_{\mathrm{GRB}}}{c} \right) (1 - \cos\theta_j) \simeq (330\ \mathrm{s}) \left(\frac{1+z}{2} \right) R_{\mathrm{GRB},15} \theta_{j,-1}^2, \qquad (9.5)$$

where the $<$ sign embraces the possibility that the end of the curvature effect phase is not observed due to the emergence of the afterglow (shallow decay segment) component. One can see that in order to account for the typically observed steep decay phase that lasts for several hundred seconds, for a typical jet opening angle $\theta_j \sim 0.1$ (Frail et al., 2001), R_{GRB} should be large, i.e. $\geq 10^{15}$ cm. This is much larger than the photosphere radius and the typical internal shock radius. Detailed analyses by several authors reached this conclusion based on similar arguments (Lyutikov, 2006; Lazzati and Begelman, 2006; Kumar et al., 2007; Hascoët et al., 2012a).

This argument is not relevant if the steep decay is not controlled by the curvature effect, but rather reflects the natural dying-off of the central engine (e.g. Fan and Wei, 2005).

9.2.4 Prompt Optical Emission

For GRBs whose prompt optical emission has been detected to track γ-rays (e.g. GRB 041219A, Vestrand et al. 2005; GRB 050820A, Vestrand et al. 2006; GRB 051111, Yost et al. 2007; GRB 061121, Page et al. 2007; GRB 080319B, Racusin et al. 2008; and GRB 110205A, Zheng et al. 2012), the condition must be such that optical emission is not suppressed by synchrotron self-absorption (SSA).

Assuming that GRB prompt emission is dominated by synchrotron radiation, and that optical and γ-ray emission originate from the same location R_{GRB}, one can then use the SSA condition to derive a lower limit on R_{GRB} (Shen and Zhang, 2009). This method does not apply if γ-rays and optical emission originate from two different zones.

Based on the synchrotron self-absorption theory discussed in §5.1.6, in principle there are four different cases for prompt optical emission: two cases for $\nu_m < \nu_a$ (so that $F_\nu \propto \nu^{5/2}$ below ν_a) with ν_{opt} either below or above ν_a, and two cases for $\nu_a < \nu_m$ (so that $F_\nu \propto \nu^2$ below ν_a), again with ν_{opt} either below or above ν_a. These four cases are shown in Fig. 9.3 upper panel. For all four cases, the observed optical flux $F_{\nu,\mathrm{opt}}$ can be used to constrain R_{GRB} (with dependence on Γ and B'). The constrained ranges of R_{GRB} for five GRBs are presented in the lower panel of Fig. 9.3 (Shen and Zhang, 2009). Generally R_{GRB} is constrained to be above 10^{14} cm.

The optical emission of the naked-eye GRB 080319B is about 2 orders of magnitude higher than the extrapolated flux from γ-rays (lower right panel of Fig. 2.19). If one still assumes that the optical emission and γ-ray emission are from the same site (as expected in the SSC model, Kumar and Panaitescu 2008; Racusin et al. 2008), then R_{GRB} is constrained to be larger than 10^{16} cm. Alternatively, optical emission of this burst may come from a different zone from the γ-ray emission (e.g. Zou et al., 2009; Fan et al., 2009).

GRB 110205A was detected by *Swift* XRT and UVOT when the γ-ray emission was still present. The UV/optical flux was found to track the γ-ray emission during the prompt phase. The SSA constraint gives a measurement of $R_{\mathrm{GRB}} \sim 3 \times 10^{13}$ cm, which is
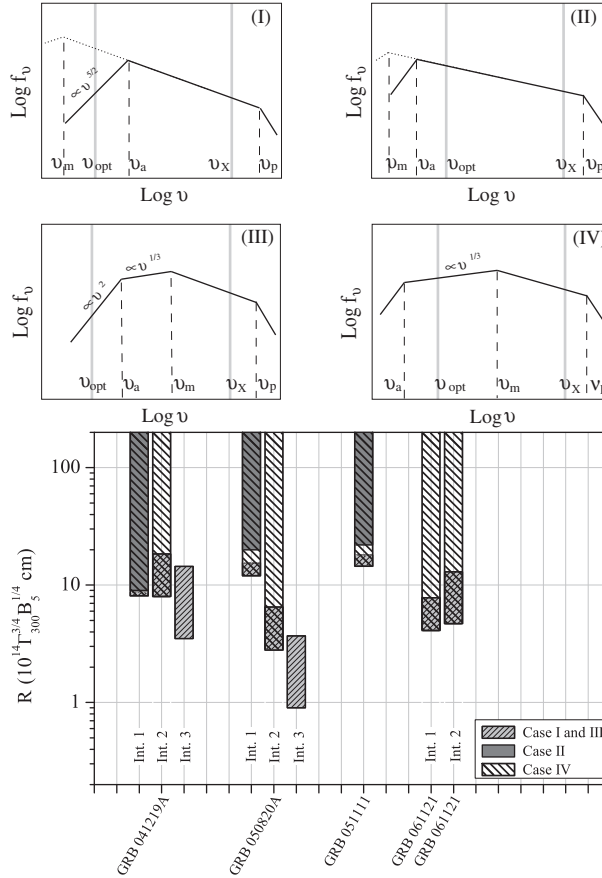
Figure 9.3 *Upper:* Four possible broad-band synchrotron spectra of GRB prompt emission. *Lower:* The derived ranges of R_{GRB} for five GRBs based on the synchrotron self-absorption constraints. From Shen and Zhang (2009).

consistent with the internal shock radius. The broad-band spectral energy distribution is also consistent with the standard internal shock fast cooling synchrotron model (Zheng et al., 2012).

9.2.5 Prompt GeV Emission

As discussed in §7.1.3, the highest energy photon carries the information of both Γ and R_{GRB}, and hence can place a constraint on R_{GRB} in the two-dimensional Γ–R_{GRB} plane (Gupta and Zhang, 2008). An example was presented for GRB 080916C (Fig. 7.2). This method makes the assumption that GeV emission and sub-MeV emission are from the same location. Under this assumption, the constrained R_{GRB} is usually large, $R_{\mathrm{GRB}} > 10^{15}$ cm for GRB 080916C (Zhang and Pe'er, 2009).

9.2.6 Sub-MeV Data

Even though the sub-MeV data do not directly constrain R_{GRB}, some observational data (lightcurves and spectral indices) can provide useful hints on the possible location of R_{GRB}.

The GRB lightcurves typically have some broad pulses (the slow variability component) with a typical variability time scale of seconds. Even though rapidly variable spikes (the fast variability component) are superposed on the broad pulses, E_p seems to evolve with respect to the profile of the broad pulses, either in the form of hard-to-soft evolution or tracking (e.g. Lu et al., 2012). This hints that the broad pulses may be the fundamental radiation unit of GRBs. If so, the emission radius would be

$$R_{\mathrm{GRB}} \sim (1+z)^{-1} \Gamma^2 c \delta t_{\mathrm{slow}} \simeq 3 \times 10^{14} \, \mathrm{cm} (1+z)^{-1} \Gamma_2^2 (\delta t_{\mathrm{slow}}/1\,\mathrm{s}). \qquad (9.6)$$

Within this scenario, the second-long slow component is produced from one single emission region as it streams outwards so that different observational time is related to the different emission time of this single emission unit at different emission radii.

A related argument is the requirement from spectral modeling. Uhm and Zhang (2014b) found that synchrotron radiation of electrons in a decaying magnetic field can give rise to a Band-like spectrum with $\alpha \sim -1$. In order to achieve this, a relatively large $R_{\mathrm{GRB}} \sim 10^{15}$ cm is needed, so that the magnetic field strength in the emission region is not strong, and the electrons are not in the "deep" fast cooling regime, i.e. t_c'/t_{dyn}' is less than but not much less than unity. Such a set-up can also interpret the spectral lags and the E_p evolution patterns observed in some GRBs (Uhm and Zhang, 2016b).

9.2.7 Summary

Independent pieces of evidence from different approaches seem to point towards a consistent picture, i.e. the GRB emission radius $R_{\mathrm{GRB}} > 10^{13}$ cm, with a typical radius $\sim 10^{15}$ cm at least for some GRBs. This suggests that at least for these bursts, some emission components (likely the dominant emission component) should originate from such a large radius where Thomson optical depth is much smaller than unity. Synchrotron radiation is likely the emission mechanism at such a radius. On the other hand, GRBs are different. The constraints derived from some GRBs may not apply to all the GRBs. It is possible and even likely that the observed GRB emission may originate from more than one location. Emission from the photosphere is known to shape the GRB spectra, sometimes even appearing as the dominant component in some GRBs. In the following sections, we will discuss various emission components in turn.

9.3 Non-Dissipative Photosphere

9.3.1 Non-Dissipative and Dissipative Photospheres

Copious photons are generated from the GRB central engine via various mechanisms, e.g. nuclear interactions in the central engine (e.g. an accretion torus or a new-born, hot, millisecond magnetar) or in the outflow due to the unequal velocities of different nuclear components (e.g. protons and neutrons), bremsstrahlung, double Compton scattering, or synchrotron. Depending on the Thomson scattering optical depth τ_{es} where thermal photons are generated, the emergent spectra from the photosphere ($\tau_{\mathrm{es}} = 1$) could differ. There are three regimes (Beloborodov, 2013):

- Planck regime ($\tau_{es} \gtrsim 10^5$):

  If most photons are generated at $\tau_{es} \gtrsim 10^5$ and no signficant dissipation happens at smaller τ_{es} regions, the photosphere emergent spectrum would be Planckian.

  There are two conditions to form a Planck spectrum:

  (1) The photons are rapidly produced, providing the photon number needed for a Planck distribution, i.e.

  $$\dot{n}_\gamma t_{dyn} \geq n_{\gamma,bb},\tag{9.7}$$

  where t_{dyn} is the dynamical time scale of the system, $\dot{n}_\gamma$ is the photon generation rate density (e.g. Eq. (5.160) for bremsstrahlung and Eq. (5.146) for double Compton scattering), and

  $$n_{\gamma,bb} \sim \frac{aT^4}{2.7kT} \simeq \frac{0.2}{\lambdabar^3}\Theta^3\tag{9.8}$$

  is the rough photon number density of a blackbody emission source, where $a = \pi^2 k^4/15c^3\hbar^3$ is the Stefan–Boltzmann energy density constant (Rybicki and Lightman, 1979), $\lambdabar = \hbar/m_e c$, and $\Theta = kT/m_e c^2$.

  (2) Thomson scattering optical depth is much greater than unity ($\tau_{es} \gg 1$) so that the generated photons can undergo rapid thermalization with electrons (or electron–positron pairs in a pair-rich fluid).

  Let us consider (9.7) in the comoving frame of the GRB ejecta, so that $t'_{dyn} \sim r/c\Gamma$. Beloborodov (2013) showed that, for GRB parameters, double Compton scattering is a more efficient mechanism for producing photons at high optical depths. Applying Eq. (5.146) with Θ replaced by Θ', and noting $\tau_{es} = n'_e\sigma(r/\Gamma)$, the condition (9.7) can be reduced to

  $$\chi\tau_{es}\Theta'^2 \gtrsim 1,\tag{9.9}$$

  where χ is the parameter defined in §5.2.6. Considering typical parameters, the radius above which the Planck condition is not satisfied (R_P) is of the order 10^{10} cm, which is greater than the coasting radius $R_c = 10^9$ cm $\Gamma_2 R_{0,7}$. One can therefore estimate $\Theta' = (\Theta_0/\Gamma)(R_P/R_c)^{-1/2} \simeq 0.01 L_{w,52}^{1/2}\Gamma_2^{1/2}R_{P,10}^{-1/2}$. Noting $\chi \sim 0.1$ in Eq. (5.146), one can derive $\tau_{es} \gtrsim 10^5$ as the Planck condition (Beloborodov, 2013). As a result, the second condition ($\tau_{es} \gg 1$) is naturally satisfied.

- Wien regime ($10^5 \gtrsim \tau_{es} \gtrsim 10^2$):

  If dissipation happens at τ_{es} below 10^5, the photon generation rate is too slow to maintain a Planck distribution (Eq. (9.7) not satisfied). The photon number is then essentially conserved. If there is no energy dissipation process to heat electrons at $\tau_{es} < 10^5$, electrons would adiabatically cool as the fireball expands and maintain the same temperature as photons. The emergent spectrum would remain a blackbody. However, it is possible or even likely that certain energy dissipation processes (e.g. internal shocks, nuclear collisions between protons and neutrons, and magnetic reconnection) may happen between $\tau_{es} \sim 10^5$ (at Planck radius R_P) and $\tau_{es} \sim 1$ (at photosphere radius R_{ph}). If this is

the case, the electron temperature would be higher than the radiation temperature.[3] The emergent spectrum would deviate from the blackbody form.

The outcome depends on the "Compton y" parameter (§5.2.5). For $y \gg 1$, one is in the saturated Comptonization regime. Multiple scattering forces photons to be in thermal equilibrium with a temperature equal to the electron temperature, i.e. $T_\gamma = T_e = T$. However, since photon number is conserved, the spectral shape becomes Wien (Eq. (5.139)), with $F_\nu \propto \nu^3$ in the low-frequency regime.

From Eq. (5.136), and noting $kT \gg \epsilon$ (electron temperature much greater than seed photon temperature due to dissipation) and $\tau_{es} \gg 1$, one can write

$$y = 4\Theta N_{\mathrm{es}} = 4\Theta \tau_{\mathrm{es}}. \tag{9.10}$$

Notice that at $\tau_{\mathrm{es}} \gg 1$, one is supposed to adopt the number of scattering events $N_{\mathrm{es}} \sim \tau_{\mathrm{es}}^2$ rather than $N_{\mathrm{es}} \sim \tau_{\mathrm{es}}$, if the emission region is static. However, for a relativistically expanding shell, τ_{es} drops in the expansion time scale, so that N_{es} is essentially τ_{es} (Beloborodov, 2011).

For $y \gtrsim 1$, or

$$\tau_{\mathrm{es}} \gg (4\Theta)^{-1} \sim 10^2, \tag{9.11}$$

Comptonization is in the saturated regime. The emergent spectrum is of the Wien shape. So the optical depth range $10^5 \gtrsim \tau_{\mathrm{es}} \gtrsim 10^2$ is called the Wien regime.

- Comptonization regime ($10^2 \gtrsim \tau_{\mathrm{es}} \gtrsim 1$):

 If significant dissipation occurs at $10^2 \gtrsim \tau_{\mathrm{es}} \gtrsim 1$, one would have $y \lesssim 1$. Comptonization causes a power-law tail to develop above the Comptonized thermal peak (see §5.2.5 for details). The low-energy photon index below the thermal peak remains Rayleigh–Jeans (if no significant dissipation occurs in the Wien zone). The non-thermal nature of the spectrum in the high-energy regime is the rationale for interpreting the typical Band spectrum of GRBs as the emission from a dissipative photosphere.

Lacking knowledge of the dissipation status of the GRB outflow, we can discuss two types of photospheres. In the following discussion in this section, we discuss the case where no significant dissipation occurs above the Planck zone, so that the photosphere is *non-dissipative*. The case of a *dissipative photosphere* is discussed in the next section.

9.3.2 The Case of a Fireball

The non-dissipative photosphere emission of a pure fireball can be precisely predicted from the theory (Mészáros and Rees, 2000b). We discuss two approaches to the problem: a bottom-up approach to characterize photosphere emission properties based on the central engine properties (R_0 and η, Mészáros and Rees 2000b), and a top-down approach to diagnose central engine properties based on the observed properties of the photosphere emission (Pe'er et al., 2007).

---

[3] Rapid cooling and heating processes would form a quasi-thermal (rather than a power-law) distribution of electrons.

Bottom-Up Approach

The bottom-up approach of photosphere emission has been discussed in §7.3.2. Here we repeat some key equations.

The initial temperature of the fireball at the central engine is (Eqs. (7.44) and (7.45)):

$$T_0 \simeq \left(\frac{L_w}{4\pi R_0^2 g_0 \sigma_B}\right)^{1/4} \simeq (1.5 \times 10^{10} \text{ K}) L_{w,52}^{1/4} R_{0,7}^{-1/2},$$

$$kT_0 \simeq 1.3 \text{ MeV} L_{w,52}^{1/4} R_{0,7}^{-1/2}.$$

If this blackbody emission component is directly observed, its observed temperature is

$$T_{0,\text{obs}} \simeq \left(\frac{L_w}{4\pi R_0^2 g_0 \sigma_B}\right)^{1/4} (1 + z_g)^{-1}(1 + z)^{-1}$$

$$\simeq (1.5 \times 10^{10} \text{ K}) L_{w,52}^{1/4} R_{0,7}^{-1/2}(1 + z_g)^{-1}(1 + z)^{-1}, \tag{9.12}$$

or

$$kT_{0,\text{obs}} \simeq 1.3 \text{ MeV} L_{w,52}^{1/4} R_{0,7}^{-1/2}(1 + z_g)^{-1}(1 + z)^{-1}, \tag{9.13}$$

where

$$z_g = \frac{1}{\sqrt{1 - \frac{r_s}{R_0}}} - 1 \tag{9.14}$$

is the gravitational redshift, and $r_s = 2GM_\bullet/c^2$ is the Schwarzschild radius of the central engine with mass $M_\bullet$.

The corresponding peak energy is

$$E_{p,0} \simeq 2.8kT_{0,\text{obs}} \simeq 3.6 \text{ MeV} L_{w,52}^{1/4} R_{0,7}^{-1/2}(1 + z_g)^{-1}(1 + z)^{-1}. \tag{9.15}$$

Notice that, for a standard blackbody, the νF_ν peak is at $3.92kT_{0,\text{obs}}$. However, for a relativistic outflow, the shape of the blackbody is somewhat modified, which may be approximated in the form (Li and Sari, 2008)

$$F_\nu \propto \frac{\nu^2}{c^2} kT \int_{h\nu/kT}^{\infty} \frac{dx}{e^x - 1}. \tag{9.16}$$

The νF_ν peak of this modified spectrum is at $\sim 2.82kT$ (Zhang et al., 2012a).

Based on the derivations in §7.3.3 (Mészáros and Rees, 2000b), the photosphere temperature is

$$\frac{T_{\text{ph}}}{T_0} = \begin{cases} \left(\frac{R_{\text{ph}}}{R_c}\right)^{-2/3} = \left(\frac{\eta}{\eta_*}\right)^{8/3}, & \eta < \eta_*, R_{\text{ph}} > R_c, \\ 1, & \eta > \eta_*, R_{\text{ph}} < R_c, \end{cases} \tag{9.17}$$

and the photosphere luminosity is

$$\frac{L_{\text{ph}}}{L_w} = \begin{cases} \left(\frac{R_{\text{ph}}}{R_c}\right)^{-2/3} = \left(\frac{\eta}{\eta_*}\right)^{8/3}, & \eta < \eta_*, R_{\text{ph}} > R_c, \\ 1, & \eta > \eta_*, R_{\text{ph}} < R_c, \end{cases} \tag{9.18}$$

where

$$\eta_* = \left(\frac{L_w\sigma_{\mathrm{T}}\mathcal{Y}}{8\pi m_p c^3 R_0}\right)^{1/4} \simeq 8.7 \times 10^2 \left(\frac{L_{w,52}\mathcal{Y}}{R_{0,7}}\right)^{1/4} \tag{9.19}$$

(Eq. (7.64)), and $\mathcal{Y}$ is pair multiplicity.

It is informative to write down the explicit expressions of the photosphere temperature (energy): If the photosphere is in the acceleration phase ($R_{\mathrm{ph}} < R_c$ and $\eta > \eta_*$), one has

$$kT_{\mathrm{ph}} = kT_{0,\mathrm{obs}} \simeq 1.3\ \mathrm{MeV}\ L_{w,52}^{1/4}R_{0,7}^{-1/2}(1+z_g)^{-1}(1+z)^{-1}. \tag{9.20}$$

If the photosphere is in the coasting phase ($R_{\mathrm{ph}} > R_c$ and $\eta < \eta_*$), adiabatic cooling gives a lower photosphere temperature:

$$kT_{\mathrm{ph}} \simeq 4.1\ \mathrm{keV}\ L_{w,52}^{-5/12}R_{0,7}^{1/6}\eta_2^{8/3}(1+z_g)^{-1}(1+z)^{-1}. \tag{9.21}$$

One can see that in this regime the dependence on η, which is the Lorentz factor of the fireball Γ, is very sensitive ($\propto \eta^{8/3}$).

It is also worth commenting that the photosphere model has been argued to be favorable in interpreting the $E_p \propto E_{\gamma,\mathrm{iso}}^{1/2}$ (Amati et al., 2002) and $E_p - L_{\gamma,\mathrm{iso}}^{1/2}$ (Yonetoku et al., 2004) and some other relations (e.g. Thompson, 2006; Thompson et al., 2007; Fan et al., 2012). The rationale is the $kT_{\mathrm{ph}} \propto L^{1/4}$ relation in Eq. (9.20), which is close to the observations. However, Eq. (9.21), which is more relevant for typical η values, predicts an opposite trend, $kT_{\mathrm{ph}} \propto L^{-5/12}$, which deviates badly from the observed relations. Furthermore, the three-parameter relation $kT_{\mathrm{ph}} \propto L^{-5/12}\eta^{8/3}$ is very different from the observed three-parameter relation $E_p \propto L^{0.55}\Gamma_0^{-0.5}$ (Liang et al., 2015). This suggests that the peak energy E_p of most GRBs is likely not defined by the non-dissipative photosphere temperature of the fireball.

Observed Spectral Shape

Even though the spectrum is a blackbody in the comoving frame, for a relativistic outflow, the observed spectrum is distorted (Goodman, 1986; Li and Sari, 2008; Deng and Zhang, 2014b). There are several factors that distort the observed spectrum. First, the spatial "shape" of the photosphere is not a conical jet. This is because photosphere is defined by the "last scattering surface" with $\tau_{\mathrm{es}} = 1$, while the scattering probability is angle and distance dependent (Pe'er, 2008; Beloborodov, 2011; Deng and Zhang, 2014b). Second, if one considers a continuous wind with a certain thickness, the observed flux at any observational time comes from emission elements associated with different emission times and locations, i.e. the observed emission comes from an "equal-arrival-time-volume" in the wind. Third, for a dynamical ejecta, the outer boundary of the ejecta is time dependent. Finally, the variation of the wind luminosity (caused by the central engine variability) further complicates the received spectrum.

An analytical treatment of the process is complicated but possible (Deng and Zhang, 2014b). For a constant luminosity wind, the specific flux at time t for a layer ejected at time $\hat{t}$ may be expressed as

$$\hat{F}_\nu(\nu, t, \hat{t}) = \frac{\dot{N}_0(\hat{t})}{4\pi D_L^2} \int \int \hat{P}(r, \Omega) P(\nu, T) h\nu$$

$$\times \delta\left(t - \hat{t} - \left(\frac{ru}{\beta c} - t_0\right)\right) d\Omega dr. \tag{9.22}$$

Here,

$$P(\nu, T) = \frac{n_\gamma(\nu, T)}{\int_0^\infty n_\gamma(\nu, T) d\nu} = \frac{n_\gamma(\nu, T)(hc)^3}{16\pi\zeta(3)(kT)^3} \tag{9.23}$$

is the probability function of a photon with frequency ν in a Planck distribution with a temperature T at the coordinate (r, θ) as observed by an observer located at $\theta = 0$, and

$$n_\gamma(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{1}{\exp(h\nu/kT) - 1} \tag{9.24}$$

is the specific photon number density at frequency ν for an observed temperature T. The mathematical relation

$$\int_0^\infty \frac{x^2 dx}{e^x - 1} = 2\zeta(3) = 2 \times 1.202\ldots \tag{9.25}$$

has been applied when calculating the integration $\int_0^\infty n_\gamma(\nu, T) d\nu$. The factor $P(r, \Omega)$ is the probability density function of last scattering at location r and solid angle (or direction) Ω, which is defined by

$$P(r, \Omega) = \frac{\sigma_T n \mathcal{D}^2 e^{-\tau(r,\mu,r_{out})}}{4\pi A}, \tag{9.26}$$

where the normalization factor is

$$A = \int \int P(r, \Omega) dr d\Omega = \int_{r_{min}}^{r_{max}} \int_0^1 \sigma_T n \frac{\mathcal{D}^2}{2} e^{-\tau(r,\mu,r_{out})} dr d\mu, \tag{9.27}$$

$\mathcal{D} = [\Gamma(1 - \beta\mu)]^{-1}$ is the Doppler factor, and $\tau(r, \mu, r_{out}) = \int_r^{r_{out}}(1 - \beta\mu)\sigma_T n dr/\mu$ is the scattering optical depth at the point (r, μ). It is $\hat{t}$ dependent, so it is expressed as $\hat{P}(r, \Omega)$ in Eq. (9.22).

The modifications to the thermal spectrum by the above effects include a slight broadening of the spectral peak and softening of the Rayleigh–Jeans slope below the peak from $F_\nu \propto \nu^2$ to $F_\nu \propto \nu^{1.5}$. The upper left panel of Fig. 9.4 shows an example of the calculated instantaneous spectra of a continuous wind at different epochs, with two dashed reference lines $F_\nu \propto \nu^2$ and $F_\nu \propto \nu^{1.5}$ marked (Deng and Zhang, 2014b).

If the photosphere emission ceases suddenly, the high-latitude emission will develop a flat $F_\nu \propto \nu^0$ spectrum (Pe'er and Ryde, 2011; Deng and Zhang, 2014b), see the upper right panel of Fig. 9.4. This corresponds to the typical low-energy Band spectrum index $\alpha = -1$. However, whenever there is emission along the line of sight, the on-beam emission will dominate the shape of the spectrum. The high-latitude emission fades rapidly with a decay index -2 (Pe'er and Ryde, 2011). However, the fiducial zero time has to be set to $t_N \sim R_{ph}/2\Gamma^2 c \sim 4 \times 10^{-5}$ s before the decay starts in order to have an immediate $F_\nu \propto t^{-2}$ decay in the log-log lightcurve. For a long-lasting wind emission whose duration $\Delta T \gg t_N$,
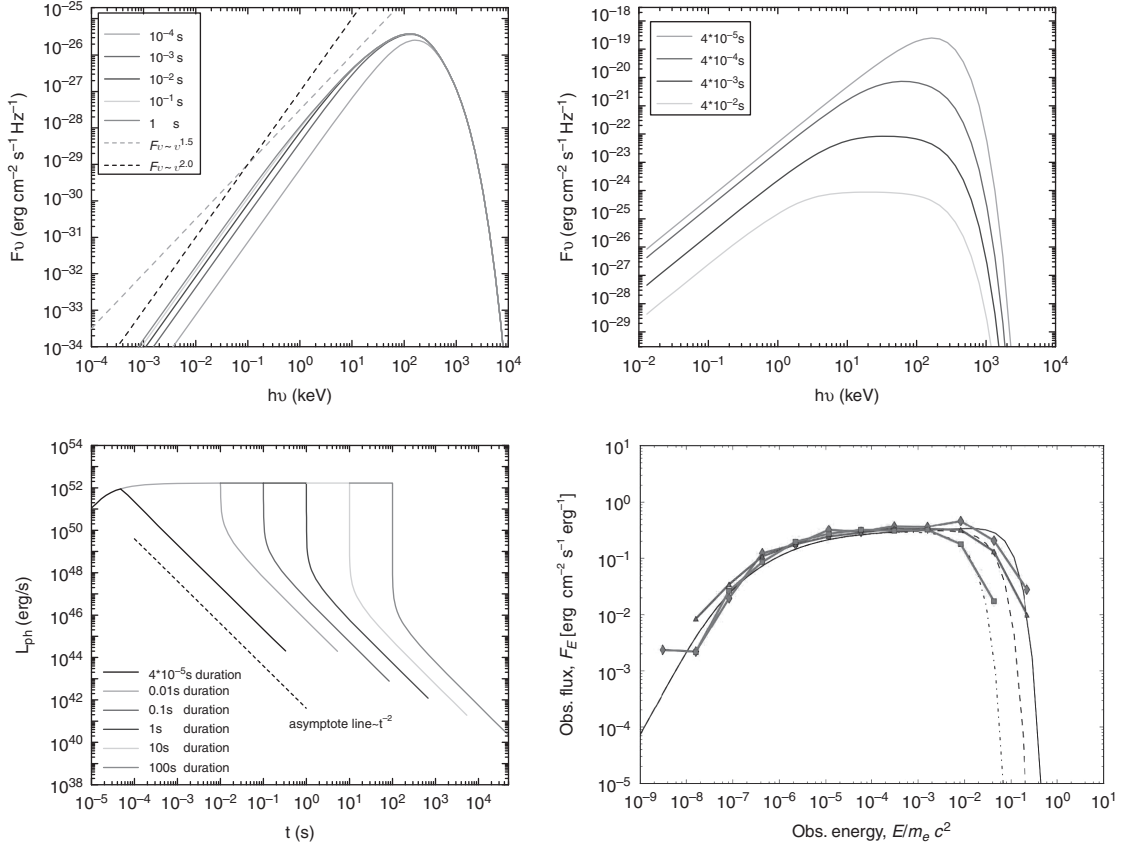
Figure 9.4 The predicted observational spectra of a non-dissipative photosphere. *Upper left*: Instantaneous spectra of a constant luminosity continous wind. *Upper right*: The high-latitude-emission-dominated photosphere emission that shows a flat spectrum. *Lower left*: The lightcurve of photosphere emission for high-latitude emission, showing the abrupt drop followed by a t^{-2} decay. *Lower right*: An example spectrum for a special type of structured jet that reproduces a flat spectrum corresponding to $\alpha \sim -1$. First three panels from Deng and Zhang (2014b), last panel from Lundman et al. (2013). A black and white version of this figure will appear in some formats. For the color version, please refer to the plate section.

the lightcurve shows an initial rapid decay before landing at the t^{-2} decay segment (Fig. 9.4, lower left). Such a feature has not been observed from GRB γ-ray emission data, suggesting that the observed emission is not from the high-latitude emission phase within the photosphere model.

The very hard low-energy photon index $\alpha = +0.5$ ($F_\nu \propto \nu^{1.5}$) is a common feature of photosphere models. As long as the central engine radius R_0 does not change significantly during a burst, it is very hard to soften this spectral index to the typical value -1, even if one considers various superposition effects (collecting photons from different (r, θ) for different emission times $\hat{t}$; Deng and Zhang 2014b). Since dissipation only modifies the spectral shape in the high-energy regime, this low-energy photon index issue is a challenge to all the photosphere models (both non-dissipative and dissipative).

Lundman et al. (2013) considered a special type of structured jet with an essentially constant luminosity per solid angle but a decreasing Γ with polar angle. They showed that $\alpha \sim -1$ can be reproduced within such a model (Fig. 9.4, lower right). This is because smaller Γ's at larger polar angles increase the $1/\Gamma$ cones at the high latitudes, so that high-latitude emission (which predicts an $\alpha = -1$ spectrum) gives the dominant contribution to the observed spectrum.

Top-Down Approach

Observationally a quasi-thermal component is identified in the time-resolved spectra of some GRBs (Ryde, 2005; Guiriec et al., 2011; Axelsson et al., 2012; Guiriec et al., 2013). Within the framework of the fireball model, Pe'er et al. (2007) developed a method to derive central engine parameters, the fireball energy-to-mass ratio η and the central engine radius R_0, based on the observed flux and temperature of the blackbody component.

The method applies to the condition $\eta < \eta_*$, or $R_{\rm ph} > R_c$. This is because, in the case of $\eta > \eta_*$, the observed photosphere temperature is the same as the temperature at the central engine, regardless of the Lorentz factor of the fireball (§7.3.3, Eq. (9.17)). The measured photosphere temperature therefore cannot be used to infer η or Γ.

When $\eta < \eta_*$ is satisfied, the photosphere radius is in the coasting regime, so that $\Gamma = \eta$. In this regime, the observed photosphere temperature decreases with radius as $T_{\rm ph} \propto R_{\rm ph}^{-2/3}$. Setting $\Gamma = \eta$, and

$$L_w = 4\pi D_{\rm L}^2 f_\gamma^{-1} F^{\rm ob}, \tag{9.28}$$

where $F^{\rm ob}$ is the observed total flux of the GRB, and $f_\gamma = L_\gamma/L_w < 1$ is the correction factor from the observed γ-ray luminosity to the true wind luminosity,[4] from Eq. (9.17) one can immediately derive (Pe'er et al., 2007) (Exercise 9.1)

$$\eta = \left[C_1 (1+z)^2 D_{\rm L} \frac{F^{\rm ob}\sigma_{\rm T}}{2m_p c^3 \mathcal{R} f_\gamma} \right]^{1/4} \tag{9.29}$$

and

$$R_0 = \frac{4^{3/2}}{C_1^4 C_2^6} \frac{D_{\rm L}}{(1+z)^2} \left(\frac{f_\gamma F_{\rm bb}^{\rm ob}}{F^{\rm ob}} \right)^{3/2} \mathcal{R}, \tag{9.30}$$

where

$$\mathcal{R} \equiv \left(\frac{F_{\rm BB}^{\rm ob}}{\sigma_B T_{\rm ob}^4} \right)^{1/2} = C_1 \frac{(1+z)^2}{D_{\rm L}} \frac{R_{\rm ph}}{\Gamma} \tag{9.31}$$

is related to the effective radius of the observed photosphere emission $R_{\rm ph}/\Gamma$ (the radius $R_{\rm ph}$ multiplied by the $1/\Gamma$ angle). The numerical factors C_1 and C_2 are the correction factors of order unity that take into account complicated angle- and radius-dependent

---

[4] In the original paper, Pe'er et al. (2007) defined a parameter $\mathcal{Y} = f_\gamma^{-1} > 1$. Since $\mathcal{Y}$ is already adopted to denote pair multiplicity in this chapter, we use f_γ in the notations to be consistent with the hybrid treatment of Gao and Zhang (2015) in §9.3.3.

probability for last scattering (Pe'er, 2008; Beloborodov, 2011; Deng and Zhang, 2014b). In Pe'er et al. (2007), $C_1 = 1.06$ and $C_2 = 1.48$ were used.

9.3.3 The Case of a Hybrid Outflow

If the GRB central engine is strongly magnetized but in the meantime also ejects a hot component with a significant release of thermal photons, then the jet has a hybrid composition. If dissipation is not important near the photosphere (e.g. as expected for a helical magnetic field configuration), the photons released at the photosphere are expected to still be in thermal equilibrium. The observed photosphere spectrum would be in the modified thermal form (upper left panel of Fig. 9.4), but the total flux is suppressed. In the following, we discuss the properties of a non-dissipative photosphere of such a hybrid jet with arbitrary central engine parameters (η, σ_0) following Gao and Zhang (2015). Both the bottom-up and top-down approaches are presented.

Bottom-Up Approach

For the hybrid model, the dynamics are modified. As discussed in §7.5, a new characteristic radius R_{ra} is defined to denote the separation from a rapid acceleration regime to a much slower acceleration regime. Following the dynamics outlined in §7.5, and again based on the definition of the photosphere radius

$$\tau = \int_{R_{ph}}^{\infty} n_e \sigma_T ds = 1, \tag{9.32}$$

where the lab-frame electron number density is now defined as (considering only a fraction $(1 + \sigma_0)$ of L_w luminosity is in the matter form)

$$n_e = \frac{L_w \mathcal{Y}}{4\pi r^2 m_p c^3 \eta (1 + \sigma_0)} \tag{9.33}$$

and

$$ds = \frac{(1 - \beta \cos\theta)}{\cos\theta} dr, \tag{9.34}$$

one can derive the "line-of-sight" photosphere radius

$$R_{ph} = \begin{cases} \left(\dfrac{L_w \mathcal{Y} \sigma_T R_0^2}{8\pi m_p c^3 \eta (1+\sigma_0)} \right)^{1/3}, & R_0 < R_{ph} < R_{ra}, \\[2ex] \left(\dfrac{L_w \mathcal{Y} \sigma_T R_{ra}^{2\delta}}{8\pi m_p c^3 \Gamma_{ra}^2 \eta (1+\sigma_0)} \right)^{1/(2\delta+1)}, & R_{ra} < R_{ph} < R_c, \\[2ex] \dfrac{L_w \mathcal{Y} \sigma_T}{8\pi m_p c^3 \Gamma_c^2 \eta (1+\sigma_0)}, & R_{ph} > R_c, \end{cases} \tag{9.35}$$

where R_0, R_{ra}, and R_c denote the radii of the initial outflow, end of rapid acceleration phase, and coasting, respectively, and δ is the acceleration index during the slow acceleration phase ($\Gamma \propto r^\delta$), with a maximum value of 1/3.

Without magnetic heating, the thermal energy undergoes adiabatic cooling, with $r^2 e^{3/4} \Gamma = $ const (e.g. Piran et al., 1993). Noting $e \propto T'^4$, the comoving photosphere temperature reads

$$T'_{\rm ph} = \begin{cases} T_0 \left(\frac{R_{\rm ph}}{R_0} \right)^{-1}, & R_0 < R_{\rm ph} < R_{\rm ra}, \\[2mm] T_0 \left(\frac{R_{\rm ra}}{R_0} \right)^{-1} \left(\frac{R_{\rm ph}}{R_{\rm ra}} \right)^{-(2+\delta)/3}, & R_{\rm ra} < R_{\rm ph} < R_c, \\[2mm] T_0 \left(\frac{R_{\rm ra}}{R_0} \right)^{-1} \left(\frac{R_c}{R_{\rm ra}} \right)^{-(2+\delta)/3} \left(\frac{R_{\rm ph}}{R_c} \right)^{-2/3}, & R_{\rm ph} > R_c, \end{cases} \tag{9.36}$$

where

$$kT_0 \simeq \left(\frac{L_w}{4\pi R_0^2 g_0 \sigma_B (1+\sigma_0)} \right)^{1/4} \simeq 1.3 \ {\rm MeV} \ L_{w,52}^{1/4} R_{0,7}^{-1/2} (1+\sigma_0)^{-1/4} \tag{9.37}$$

is the temperature at R_0.

The characteristic radii ($R_{\rm ra}$ and R_c) and photosphere properties ($R_{\rm ph}$, $\Gamma_{\rm ph}$, $(1+\sigma_{\rm ph})$, $T_{\rm ob}$, and $F_{\rm BB}$) can be derived in six regimes:

Regime I ($\eta > (1+\sigma_0)^{1/2}$ and $R_{\rm ph} < R_{\rm ra}$):

$$R_{\rm ra} = (1.0 \times 10^{11} \ {\rm cm}) \ R_{0,9} \eta_2,$$
$$R_c = (1.0 \times 10^{17} \ {\rm cm}) \ R_{0,9} \eta_2 (1+\sigma_0)_2^3,$$
$$R_{\rm ph} = (8.34 \times 10^{10} \ {\rm cm}) \ L_{w,52}^{1/3} R_{0,9}^{2/3} \eta_2^{-1/3} (1+\sigma_0)_2^{-1/3},$$
$$\Gamma_{\rm ph} = 83.4 L_{w,52}^{1/3} R_{0,9}^{-1/3} \eta_2^{-1/3} (1+\sigma_0)_2^{-1/3}, \tag{9.38}$$
$$1 + \sigma_{\rm ph} = 100 (1+\sigma_0)_2,$$
$$T_{\rm ob} = 56.1 \ {\rm keV} \ (1+z)^{-1} L_{w,52}^{1/4} R_{0,9}^{-1/2} (1+\sigma_0)_2^{-1/4},$$
$$F_{\rm BB} = (1.07 \times 10^{-7} \ {\rm erg \ s}^{-1} {\rm cm}^{-2}) \ L_{w,52} (1+\sigma_0)_2^{-1} d_{L,28}^{-2}.$$

Regime II ($\eta > (1+\sigma_0)^{1/2}$ and $R_{\rm ra} < R_{\rm ph} < R_c$):[5]

$$R_{\rm ra} = (1.0 \times 10^{11} \ {\rm cm}) \ R_{0,9} \eta_2,$$
$$R_c = (1.0 \times 10^{17} \ {\rm cm}) \ R_{0,9} \eta_2 (1+\sigma_0)_2^3,$$
$$R_{\rm ph} = (7.22 \times 10^{10} \ {\rm cm}) \ L_{w,52}^{3/5} R_{0,9}^{2/5} \eta_2^{-7/5} (1+\sigma_0)_2^{-3/5},$$
$$\Gamma_{\rm ph} = 89.7 L_{w,52}^{1/5} R_{0,9}^{-1/5} \eta_2^{1/5} (1+\sigma_0)_2^{-1/5}, \tag{9.39}$$
$$1 + \sigma_{\rm ph} = 111.5 L_{w,52}^{-1/5} R_{0,9}^{1/5} \eta_2^{4/5} (1+\sigma_0)_2^{6/5},$$
$$T_{\rm ob} = 64.8 \ {\rm keV} \ (1+z)^{-1} L_{w,52}^{-1/60} R_{0,9}^{-7/30} \eta_2^{16/15} (1+\sigma_0)_2^{1/60},$$
$$F_{\rm BB} = (1.24 \times 10^{-7} \ {\rm erg \ s}^{-1} {\rm cm}^{-2}) \ L_{w,52}^{11/15} R_{0,9}^{4/15} \eta_2^{16/15} (1+\sigma_0)_2^{-11/15} d_{L,28}^{-2}.$$

---

[5] A typo in the coefficient of $1 + \sigma_{\rm ph}$ in Gao and Zhang (2015) is corrected.

Regime III ($\eta > (1 + \sigma_0)^{1/2}$ and $R_{\mathrm{ph}} > R_c$):

$$
\begin{aligned}
R_{\mathrm{ra}} &= (1.0 \times 10^{11}\ \mathrm{cm})\, R_{0,9}\eta_2, \\
R_c &= (1.0 \times 10^{17}\ \mathrm{cm})\, R_{0,9}\eta_2(1 + \sigma_0)_2^3, \\
R_{\mathrm{ph}} &= (5.81 \times 10^{12}\ \mathrm{cm})\, L_{w,52}\eta_1^{-3}(1 + \sigma_0)_1^{-3}, \\
\Gamma_{\mathrm{ph}} &= 100\eta_1(1 + \sigma_0)_1, \\
1 + \sigma_{\mathrm{ph}} &\simeq 1, \\
T_{\mathrm{ob}} &= 6.65\ \mathrm{keV}\ (1 + z)^{-1}L_{w,52}^{-5/12}R_{0,9}^{1/6}\eta_1^{8/3}(1 + \sigma_0)_1^{29/12}, \\
F_{\mathrm{BB}} &= (7.15 \times 10^{-8}\ \mathrm{erg\ s^{-1}cm^{-2}})\, L_{w,52}^{1/3}R_{0,9}^{2/3}\eta_1^{8/3}(1 + \sigma_0)_1^{5/3}d_{L,28}^{-2}.
\end{aligned}
\tag{9.40}
$$

Regime IV ($\eta < (1 + \sigma_0)^{1/2}$ and $R_{\mathrm{ph}} < R_{\mathrm{ra}}$):

$$
\begin{aligned}
R_{\mathrm{ra}} &= (2.15 \times 10^{10}\ \mathrm{cm})\, R_{0,9}\eta_2^{1/3}(1 + \sigma_0)_2^{1/3}, \\
R_c &= (2.15 \times 10^{18}\ \mathrm{cm})\, R_{0,9}\eta_2^{7/3}(1 + \sigma_0)_2^{7/3}, \\
R_{\mathrm{ph}} &= (8.34 \times 10^{10}\ \mathrm{cm})\, L_{w,52}^{1/3}R_{0,9}^{2/3}\eta_2^{-1/3}(1 + \sigma_0)_2^{-1/3}, \\
\Gamma_{\mathrm{ph}} &= 83.4L_{w,52}^{1/3}R_{0,9}^{-1/3}\eta_2^{-1/3}(1 + \sigma_0)_2^{-1/3}, \\
1 + \sigma_{\mathrm{ph}} &= 5.56L_{w,52}^{-1/3}R_{0,9}^{1/3}\eta_2^{4/3}(1 + \sigma_0)_2^{4/3}, \\
T_{\mathrm{ob}} &= 56.1\ \mathrm{keV}\ (1 + z)^{-1}L_{w,52}^{1/4}R_{0,9}^{-1/2}(1 + \sigma_0)_2^{-1/4}, \\
F_{\mathrm{BB}} &= (1.07 \times 10^{-7}\ \mathrm{erg\ s^{-1}cm^{-2}})\, L_{w,52}(1 + \sigma_0)_2^{-1}d_{L,28}^{-2}.
\end{aligned}
\tag{9.41}
$$

Regime V ($\eta < (1 + \sigma_0)^{1/2}$ and $R_{\mathrm{ra}} < R_{\mathrm{ph}} < R_c$):[6]

$$
\begin{aligned}
R_{\mathrm{ra}} &= (2.15 \times 10^{10}\ \mathrm{cm})\, R_{0,9}\eta_2^{1/3}(1 + \sigma_0)_2^{1/3}, \\
R_c &= (2.15 \times 10^{18}\ \mathrm{cm})\, R_{0,9}\eta_2^{7/3}(1 + \sigma_0)_2^{7/3}, \\
R_{\mathrm{ph}} &= (2.46 \times 10^{11}\ \mathrm{cm})\, L_{w,52}^{3/5}R_{0,9}^{2/5}\eta_2^{-13/15}(1 + \sigma_0)_2^{-13/15}, \\
\Gamma_{\mathrm{ph}} &= 48.5L_{w,52}^{1/5}R_{0,9}^{-1/5}\eta_2^{-1/15}(1 + \sigma_0)_2^{-1/15}, \\
1 + \sigma_{\mathrm{ph}} &= 206L_{w,52}^{-1/5}R_{0,9}^{1/5}\eta_2^{16/15}(1 + \sigma_0)_2^{16/15}, \\
T_{\mathrm{ob}} &= 19.0\ \mathrm{keV}\ (1 + z)^{-1}L_{w,52}^{-1/60}R_{0,9}^{-7/30}\eta_2^{8/15}(1 + \sigma_0)_2^{17/60}, \\
F_{\mathrm{BB}} &= (3.63 \times 10^{-8}\ \mathrm{erg\ s^{-1}cm^{-2}})\, L_{w,52}^{11/15}R_{0,9}^{4/15}\eta_2^{8/15}(1 + \sigma_0)_2^{-7/15}d_{L,28}^{-2}.
\end{aligned}
\tag{9.42}
$$

Regime VI ($\eta < (1 + \sigma_0)^{1/2}$ and $R_{\mathrm{ph}} > R_c$):

$$
\begin{aligned}
R_{\mathrm{ra}} &= (2.15 \times 10^{10}\ \mathrm{cm})\, R_{0,9}\eta_2^{1/3}(1 + \sigma_0)_2^{1/3}, \\
R_c &= (2.15 \times 10^{18}\ \mathrm{cm})\, R_{0,9}\eta_2^{7/3}(1 + \sigma_0)_2^{7/3}, \\
R_{\mathrm{ph}} &= (5.81 \times 10^{12}\ \mathrm{cm})\, L_{w,52}\eta_1^{-3}(1 + \sigma_0)_1^{-3}, \\
\Gamma_{\mathrm{ph}} &= 100\eta_1(1 + \sigma_0)_1, \\
1 + \sigma_{\mathrm{ph}} &\simeq 1,
\end{aligned}
\tag{9.43}
$$

---

[6] A typo in the coefficient of $1 + \sigma_{\mathrm{ph}}$ in Gao and Zhang (2015) is corrected.
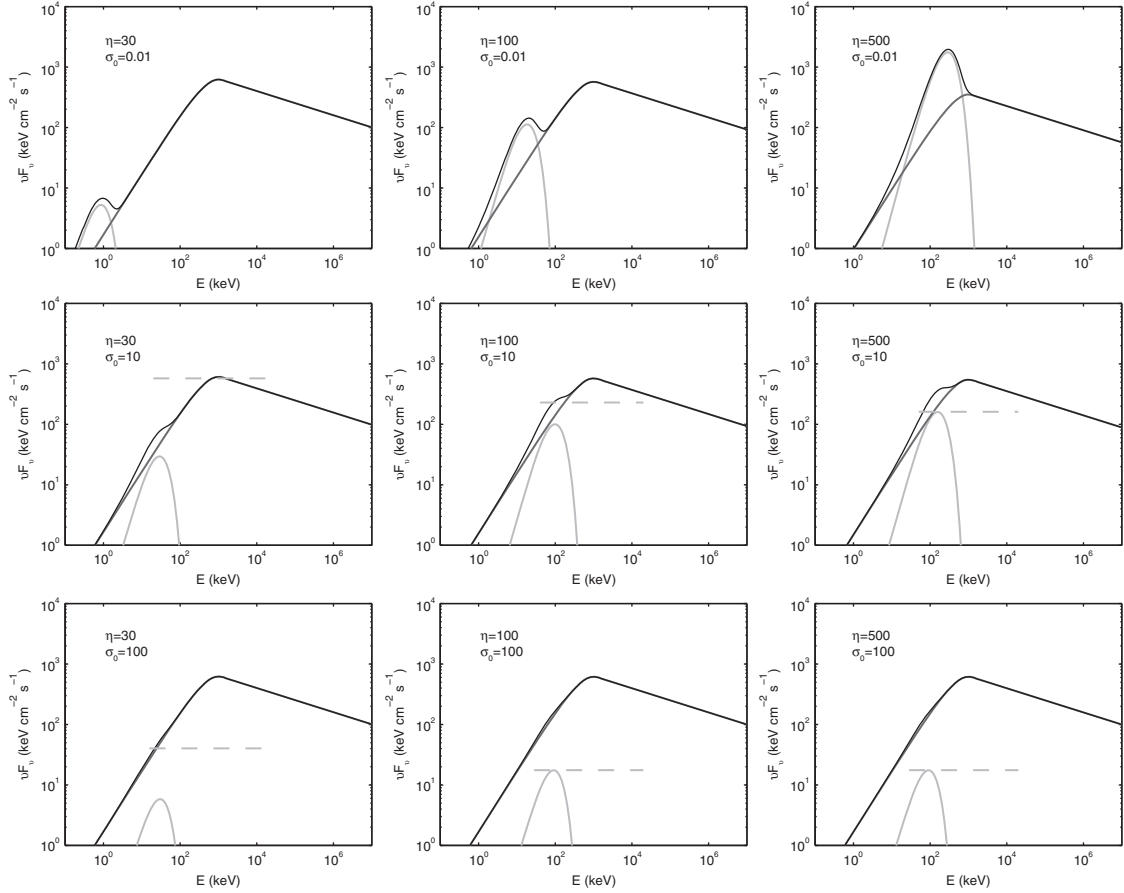
Figure 9.5 The predicted thermal plus non-thermal spectra for different input parameter pairs (η, σ_0) at the central engine. The parameters $L_w = 10^{52}$ erg s^{-1}, $R_0 = 10^9$ cm, and $z = 1$ are adopted. From Gao and Zhang (2015).

$$T_{\rm ob} = 6.65 \text{ keV } (1+z)^{-1} L_{w,52}^{-5/12} R_{0,9}^{1/6} \eta_1^{8/3} (1+\sigma_0)_1^{29/12},$$

$$F_{\rm BB} = (7.15 \times 10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}) L_{w,52}^{1/3} R_{0,9}^{2/3} \eta_1^{8/3} (1+\sigma_0)_1^{5/3} d_{L,28}^{-2}.$$

Assuming that a certain fraction (say 50%) of the wind energy is finally dissipated in an optically thin region and gives rise to a Band-function spectral component (probably via synchrotron radiation; Uhm and Zhang 2014b), one may simulate the superposition between the photosphere component and the Band component (Fig. 9.5). One can see that the relative importance of the thermal and non-thermal components depends on the central engine parameters η and σ_0. The photosphere component is progressively suppressed when σ_0 increases. The suppression of the photosphere emission by a Poynting flux has been discussed in Daigne and Mochkovitch (2002) and Zhang and Mészáros (2002a). The lack of or the weak photosphere emission detected in some GRBs in the *Fermi* era may be attributed to a Poynting-flux-dominated outflow (Zhang and Pe'er, 2009; Gao and Zhang, 2015).

Top-Down Approach

Similar to the pure fireball model, from the observed properties of photosphere temperature and flux, one may infer the physical properties of a hybrid jet. Besides R_0 and η, there is a third parameter σ_0 that characterizes the central engine. From the two observational parameters (temperature and luminosity) of the photosphere emission, one cannot uniquely determine all three parameters. On the other hand, if one assumes the fireball model, the derived η and R_0 values from the GRB data sometimes give curious, unrealistic conclusions, suggesting that σ_0 should be included in the modeling (Iyyani et al., 2013; Gao and Zhang, 2015). With all three parameters in the problem, it is more natural to fix R_0 to a certain value, and attribute the variation of the photosphere properties to the variations of both η and σ_0. In the following, we present a top-down approach to infer the (η, σ_0) pair from the observational properties of the photosphere emission by assuming a fixed R_0 value (Gao and Zhang, 2015).

One still considers the six regimes discussed above from the bottom-up approach. Regimes I and IV are similar to the $\eta > \eta_*$ regime for fireballs, in which the photosphere properties do not depend on Γ and η. So we cannot fully solve the photosphere parameters. Solutions can be found for the other four regimes, with Regimes III and VI having the same solutions. Given the measured temperature $T_{\rm ob}$, blackbody flux $F_{\rm BB}$, the thermal-to-non-thermal flux ratio $f_{\rm th}$, the γ-ray luminosity to wind luminosity ratio f_γ (which is $\mathcal{Y}^{-1}$ in the notation of Pe'er et al. 2007), and redshift z, one can derive various parameters in different regimes:

Regime II:

$$1 + \sigma_0 = 25.5(1+z)^{4/3} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{4/3} \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-1/3}$$
$$\times R_{0,9}^{2/3} f_{\rm th,-1}^{-1} f_\gamma^{-1} d_{L,28}^{-2/3},$$

$$\eta = 74.8(1+z)^{11/12} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{11/12} \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{1/48} R_{0,9}^{5/24} d_{L,28}^{1/24},$$

$$R_{\rm ph} = (1.78 \times 10^{10} \text{ cm})(1+z)^{-25/12} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{-25/12}$$
$$\times \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{37/48} R_{0,9}^{-7/24} d_{L,28}^{37/24} R_{0,9}^{-7/24} d_{L,28}^{13/24}, \quad\quad (9.44)$$

$$\Gamma_{\rm ph} = 46.4(1+z)^{-1/12} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{-1/12} \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{13/48},$$

$$1 + \sigma_{\rm ph} = 41.2(1+z)^{7/3} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{7/3} \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-7/12}$$
$$\times R_{0,9}^{7/6} f_{\rm th,-1}^{-1} f_\gamma^{-1} d_{L,28}^{-7/6},$$

$$1 + \sigma_{r_{15}} = 1.08(1+z)^{59/36} \left(\frac{T_{\rm ob}}{50 \text{ keV}} \right)^{59/36} \left(\frac{F_{\rm BB}}{10^{-8} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-47/144}$$
$$\times R_{0,9}^{77/72} f_{\rm th,-1}^{-1} f_\gamma^{-1} d_{L,28}^{-47/72}.$$

Regimes III and VI:

$$1 + \sigma_0 = 5.99(1+z)^{4/3} \left(\frac{T_{\rm ob}}{30 \text{ keV}} \right)^{4/3} \left(\frac{F_{\rm BB}}{10^{-7} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-1/3}$$
$$\times R_{0.9}^{2/3} f_{\rm th,-1}^{-1} f_\gamma^{-1} d_{L,28}^{-2/3},$$

$$\eta = 20.3(1+z)^{-5/6} \left(\frac{T_{\rm ob}}{30 \text{ keV}} \right)^{-5/6} \left(\frac{F_{\rm BB}}{10^{-7} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{11/24}$$
$$\times R_{0.9}^{-2/3} f_{\rm th,-1}^{3/4} f_\gamma^{3/4} d_{L,28}^{11/12},$$

$$R_{\rm ph} = (4.09 \times 10^{11} \text{ cm})(1+z)^{-3/2} \left(\frac{T_{\rm ob}}{30 \text{ keV}} \right)^{-3/2} \left(\frac{F_{\rm BB}}{10^{-7} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{5/8} \quad (9.45)$$
$$\times f_{\rm th,-1}^{-1/4} f_\gamma^{-1/4} d_{L,28}^{5/4},$$

$$\Gamma_{\rm ph} = 121.3(1+z)^{1/2} \left(\frac{T_{\rm ob}}{30 \text{ keV}} \right)^{1/2} \left(\frac{F_{\rm BB}}{10^{-7} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{1/8}$$
$$\times f_{\rm th,-1}^{-1/4} f_\gamma^{-1/4} d_{L,28}^{1/4}.$$

Regime V:

$$1 + \sigma_0 = 6.43(1+z)^{4/3} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{4/3} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-1/3}$$
$$\times R_{0.9}^{2/3} f_{\rm th,-1}^{-1} f_\gamma^{-1} d_{L,28}^{-2/3},$$

$$\eta = 105.0(1+z)^{7/6} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{7/6} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{5/24}$$
$$\times R_{0.9}^{1/12} f_{\rm th,-1}^{1/2} f_\gamma^{1/2} d_{L,28}^{5/12},$$

$$R_{\rm ph} = (4.62 \times 10^{10} \text{ cm})(1+z)^{-13/6} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{-13/6} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{17/24}$$
$$\times R_{0.9}^{-1/4} f_{\rm th,-1}^{-1/6} f_\gamma^{-1/6} d_{L,28}^{17/12},$$

$$\Gamma_{\rm ph} = 15.3(1+z)^{-1/6} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{-1/6} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{5/24} \quad (9.46)$$
$$\times R_{0.9}^{-1/4} f_{\rm th,-1}^{-1/6} f_\gamma^{-1/6} d_{L,28}^{5/12},$$

$$1 + \sigma_{\rm ph} = 44.2(1+z)^{8/3} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{8/3} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-1/3}$$
$$\times R_{0.9} f_{\rm th,-1}^{-1/3} f_\gamma^{-1/3} d_{L,28}^{-2/3},$$

$$1 + \sigma_{r_{15}} = 1.59(1+z)^{35/18} \left(\frac{T_{\rm ob}}{10 \text{ keV}} \right)^{35/18} \left(\frac{F_{\rm BB}}{10^{-9} \text{ erg s}^{-1}\text{cm}^{-2}} \right)^{-7/72}$$
$$\times R_{0.9}^{11/12} f_{\rm th,-1}^{-7/18} f_\gamma^{-7/18} d_{L,28}^{-7/36}.$$

When applying such a technique, R_0 is set to a characteristic value (e.g. 10^9 cm). The results depend on the R_0 value assumed. Notice that the σ value at $r = 10^{15}$ cm is also presented. This radius is the typical radius where GRB non-thermal emission is emitted based on the $\delta t_{\rm slow} \sim 1$ s (§9.2.6). The σ value at this radius gives a hint as to whether
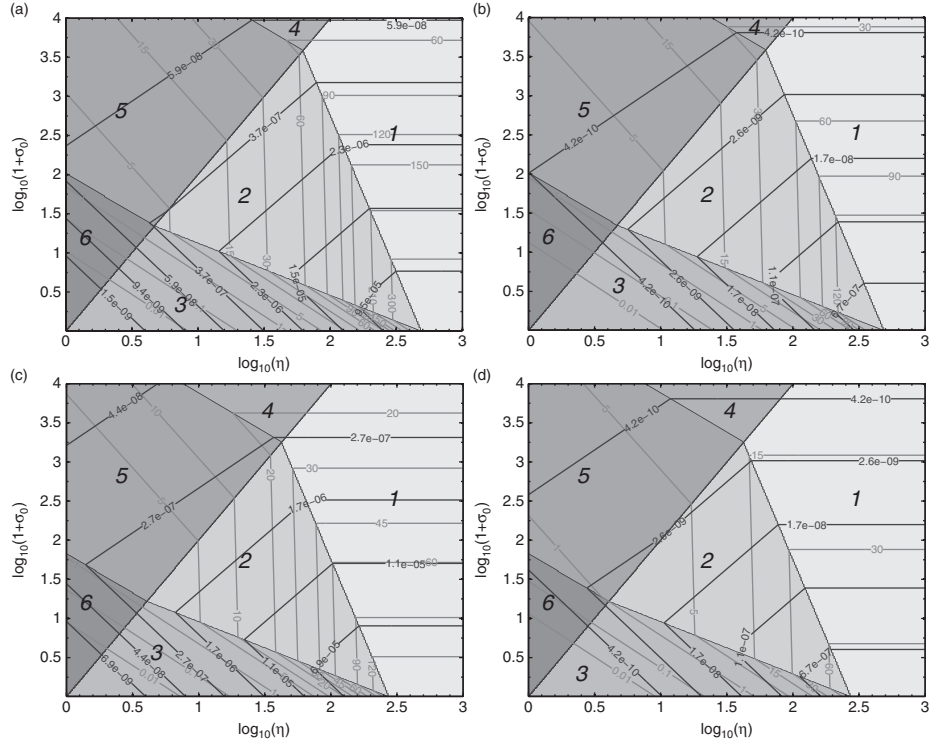
Figure 9.6 The temperature (in units of keV) and flux (in units of erg cm^{-2} s^{-1}) in the $(\eta, (1 + \sigma_0))$ domain. Parameters: $L_w = 10^{52}$ erg s^{-1}, $R_0 = 10^8$ cm (*upper panels*) or $R_0 = 10^9$ cm (*lower panels*), and $z = 0.1$ (*left panels*) or $z = 1$ (*right panels*). By measuring T_{ob} and F_{BB}, one can find a solution of $(\eta, (1 + \sigma_0))$ by assuming a R_0 value. From Gao and Zhang (2015). A black and white version of this figure will appear in some formats. For the color version, please refer to the plate section.

significant energy dissipation is through internal shocks ($\sigma_{r_{15}} < 1$) or magnetic dissipation (such as ICMART events) ($\sigma_{r_{15}} > 1$).

A convenient way to apply these results is to use the F_{BB} and T_{ob} contours in the (η, σ_0) domain to identify the (η, σ_0) values of the central engine for a particular set of data. Figure 9.6 gives an example of the contours of temperature (T_{ob}, in units of keV) and flux (F_{BB}, in units of erg cm^{-2} s^{-1}) in the $(\eta, (1 + \sigma_0))$ domain.

9.4 Dissipative Photosphere

If significant energy dissipation occurs below the photosphere, sub-photosphere electrons may have a different temperature from the seed photons. Comptonization would distort the seed photon spectrum (see §9.3.1 for a discussion of different optical depth regimes). In the literature, several possible mechanisms have been proposed for the sub-photosphere

energy dissipation (e.g. Rees and Mészáros, 2005), including small-radii internal shocks, nuclear and Coulomb collisions in a neutron-rich outflow, and magnetic dissipation. Within the collapsar scenario for long GRBs, significant energy dissipation happens when the jet breaks out from the progenitor star (Lazzati et al., 2009, 2013).

Dissipative photosphere models have been extensively studied by many authors over the years to interpret the main emission component (the Band component) of GRBs (Thompson, 1994; Ghisellini and Celotti, 1999; Rees and Mészáros, 2005; Pe'er et al., 2006; Thompson, 2006; Thompson et al., 2007; Giannios, 2006, 2008). A wave of studies on dissipative photospheres was triggered in the *Fermi* era (e.g. Lazzati and Begelman, 2010; Beloborodov, 2010, 2011, 2013; Ioka, 2010; Pe'er and Ryde, 2011; Mizuta et al., 2011; Vurm et al., 2011; Toma et al., 2011; Mészáros and Rees, 2011; Murase et al., 2012; Veres and Mészáros, 2012; Veres et al., 2012; Lazzati et al., 2013; Lundman et al., 2013; Asano and Mészáros, 2013; Bégué and Pe'er, 2015), when broad-band GRB prompt spectra of at least some GRBs (e.g. GRB 090902B) were identified to have a photosphere origin (Ryde et al., 2010; Zhang et al., 2011; Pe'er et al., 2012).

9.4.1 General Considerations

One major goal of the dissipative photosphere models is to interpret the main spectral component of GRBs, i.e. the Band-function component, as emission from the photosphere, with the GRB spectral peak energy E_p defined by the (upscattered) plasma temperature at the photosphere radius R_{ph}. A successful dissipative photosphere model should simultaneously interpret the value of E_p and the spectral index below it (the α parameter of the Band function).

As discussed in §9.3.1, there are three optical depth regimes from which different spectral shapes are formed. In the Planck regime ($\tau_{es} \gtrsim 10^5$), dissipation processes give no distortion to the spectra, since photons and leptons stay in thermal equilibrium due to the adequate photon generation rate. In the Wien regime ($10^5 \gtrsim \tau_{es} \gtrsim 10^2$), not enough photons are generated to maintain a Planck distribution. Since $y \gg 1$ in this regime, if leptons have a higher temperature than the seed photons ($T_e > T_{ph,0}$), photons will be upscattered to the lepton temperature ($T_{ph} \sim T_e$), but possess a low-energy spectral index harder than the blackbody Rayleigh–Jeans form ($F_\nu \propto \nu^2$, $\alpha = +1$), i.e. of the Wien form ($F_\nu \propto \nu^3$, $\alpha = +2$). Finally, in the Comptonization regime ($10^2 \gtrsim \tau_{es} \gtrsim 1$), the Comptonizaton parameter y is small enough that a non-thermal tail will be developed. The low-energy spectral index would stay between Rayleigh–Jeans and Wien (close to Rayleigh–Jeans). Since the Band function is characterized by a high-energy power-law segment, the relevant regime for dissipative photospheres is the Comptonization regime.

In the Comptonization regime, usually one has $y \ll 1$. The emergent spectrum essentially retains the energy E_p of the seed photons, but develops a power-law tail above E_p. The photon spectral index (convention $N(\nu) \propto \nu^\beta$) is defined by $\beta = \ln(\tau_{es}/A_f)/\ln(A_f)$, where

$$A_f = \nu_s/\nu_i \equiv \max(1 + 4kT/m_e c^2, 1 + 4\gamma_e^2/3) \qquad (9.47)$$

is the ratio between the upscattered photon frequency ν_s and the incident seed photon frequency ν_i.
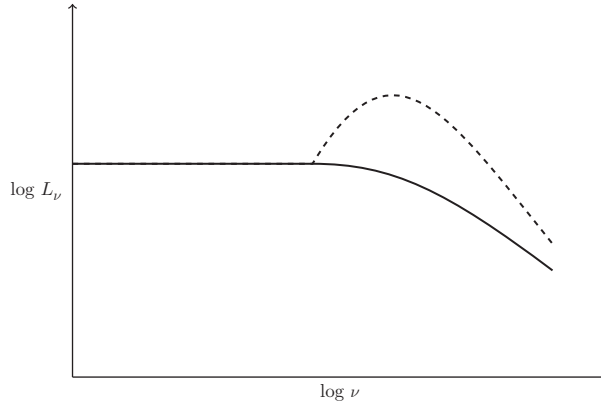
Figure 9.7   An indicative spectrum due to Comptonization with a moderate y. From Kumar and Zhang (2015).

The emergent spectrum for the intermediate regime between Comptonization and Wien regimes is more complicated. One has $y \sim 1$ and $\tau_{es} > 1$ in this regime. The spectrum has to be calculated numerically by solving the Kompaneets equation. Nonetheless, a qualitative picture is available. The peak of the emergent spectrum is larger than the thermal peak of the incident blackbody spectrum, but is still much smaller than the electron temperature, i.e. $kT_{ph} \lesssim E_p \ll kT_e$. A Comptonized power-law tail would be developed, with a photon index $\beta = -1 - 4/(3y)$. Immediately below E_p, the spectrum would have a sharp rise due to accumulation of photons in the frequency space as their energy approaches the electron temperature. Far below the peak, a flat spectrum (corresponding to $\alpha = -1$) would be developed (Ghisellini, 2013). An indicative Comptonized spectrum is presented in Fig. 9.7. The bump feature is not observed in GRB prompt emission spectra.

In general, the dissipative photosphere models can reproduce the high-energy photon index of GRBs (e.g. Lazzati and Begelman, 2010; Beloborodov, 2010; Santana et al., 2016). The predicted low-energy spectral index is very hard, which matches a small fraction of GRBs (e.g. GRB 090902B, Ryde et al. 2010; Pe'er 2012). However, the simplest versions of the dissipative photosphere models have difficulty reproducing the typical low-energy photon index -1 of GRBs (Deng and Zhang, 2014b), unless a special geometry is assumed (Lundman et al., 2013).

9.4.2 Dissipation through Internal Shocks

The condition to have a dissipative photosphere through internal shocks is $R_{IS} \leq R_{ph}$, which in the $\eta < \eta_*$ regime (§7.3.3) relevant for typical GRB parameters can be written as

$$\eta^2 c \delta t \leq \frac{\sigma_T L_w \mathcal{Y}}{8\pi \eta^3 m_p c^3}, \tag{9.48}$$

or

$$\eta \leq \left(\frac{\sigma_T L_w \mathcal{Y}}{8\pi m_p c^4 \delta t} \right)^{1/5} \sim 180 L_{w,52}^{1/5} \mathcal{Y}^{1/5} (\delta t_{-3})^{-1/5}. \tag{9.49}$$

Heating of electrons in internal shocks may be achieved by direct Fermi acceleration of electrons in the shocked region or through hadronic cascade processes (e.g Murase et al., 2012).

9.4.3 Collisional Dissipative Photosphere

If the fireball is neutron rich, as expected in various progenitor models (Beloborodov, 2003b), there is an additional characteristic radius, R_{pn}, below which neutrons and protons are in equilibrium due to nuclear elastic collisions. The calculation of R_{pn} is similar to that of R_{ph}, except that the Thomson cross section σ_T is replaced by the nuclear collision cross section σ_{pn}. According to §7.3.5, there are two regimes for proton–neutron interactions (Mészáros and Rees, 2000a): If $\eta < \eta_{pn}$ is satisfied, one has $R_{pn} > R_c$. Both protons and neutrons reach the coasting phase before decoupling. The final proton and neutron shells essentially travel with the same velocity, so that nuclear collisional interaction is not important. If, however, $\eta > \eta_{pn}$ is satisfied, one has $R_{pn} < R_c$. Neutrons decouple from protons before reaching the coasting velocity. No longer attached to protons, neutrons travel with a smaller Lorentz factor (that at the decoupling time) while protons continue to accelerate. A relative velocity develops between protons and neutrons, leading to significant *inelastic* nuclear collisions with pion production (§6.2.3). This process would give rise to multi-GeV neutrino emission from GRBs (Bahcall and Mészáros, 2000) but also photosphere heating as discussed below.

Since $\sigma_T/\sigma_{pn,0} \sim 22$, $R_{pn} < R_{ph}$ is always satisfied. If $\eta > \eta_{pn} \sim 400$ is satisfied, inelastic nuclear collisions below the photosphere naturally give rise to a dissipative photosphere (Beloborodov, 2010). According to this model, pion decay gives rise to a pair-dominated fireball ($\pi^\pm \to \mu^\pm \nu_\mu(\bar{\nu}_\mu) \to e^\pm \nu_e(\bar{\nu}_e)\nu_\mu\bar{\nu}_\mu$), with a typical Lorentz factor $\gamma_e \sim m_\pi/m_e \sim 300$ ($m_\pi c^2 \simeq 140$ MeV) in the rest frame of the outflow. Coulomb collisions between heated protons and leptons keep the lepton temperature T_e higher than the photon temperature T_{ph}, so that Comptonization gives rise to a non-thermal tail of the spectrum.

The process can give rise to a Band-like spectrum (upper left panel of Fig. 9.8), but the low-energy spectral index is too hard: $\alpha = +0.4$ (Beloborodov, 2010). Such a hard spectrum is a consequence of the thermal nature of the spectrum. It is ubiquitous for photosphere models (Deng and Zhang, 2014b), unless a special structured jet is invoked (Lundman et al., 2013). Such a model is an excellent candidate to interpret GRBs with a very hard α and a narrow spectral component (e.g. GRB 090902B). The α values of most GRBs are softer than this value (peaking at ~ -1). Since pairs at the dissipative photosphere also radiate synchrotron radiation, Vurm et al. (2011) invoked a synchrotron component to make a bump-like feature below E_p in order to soften the observed spectrum (upper right panel of Fig. 9.8). The required ϵ_B is low ($\sim 10^{-3}$–10^{-2}). Since the number of photons is large at low energies, such a bump feature plus a hard spectrum could be readily identified from the data, but so far has not been observed. Some GRBs show a featureless Band-function spectrum with $\alpha \sim -1$ in time-resolved spectra (e.g. Zhang et al., 2016b). These bursts are difficult to interpret with this collisional dissipative photosphere model.
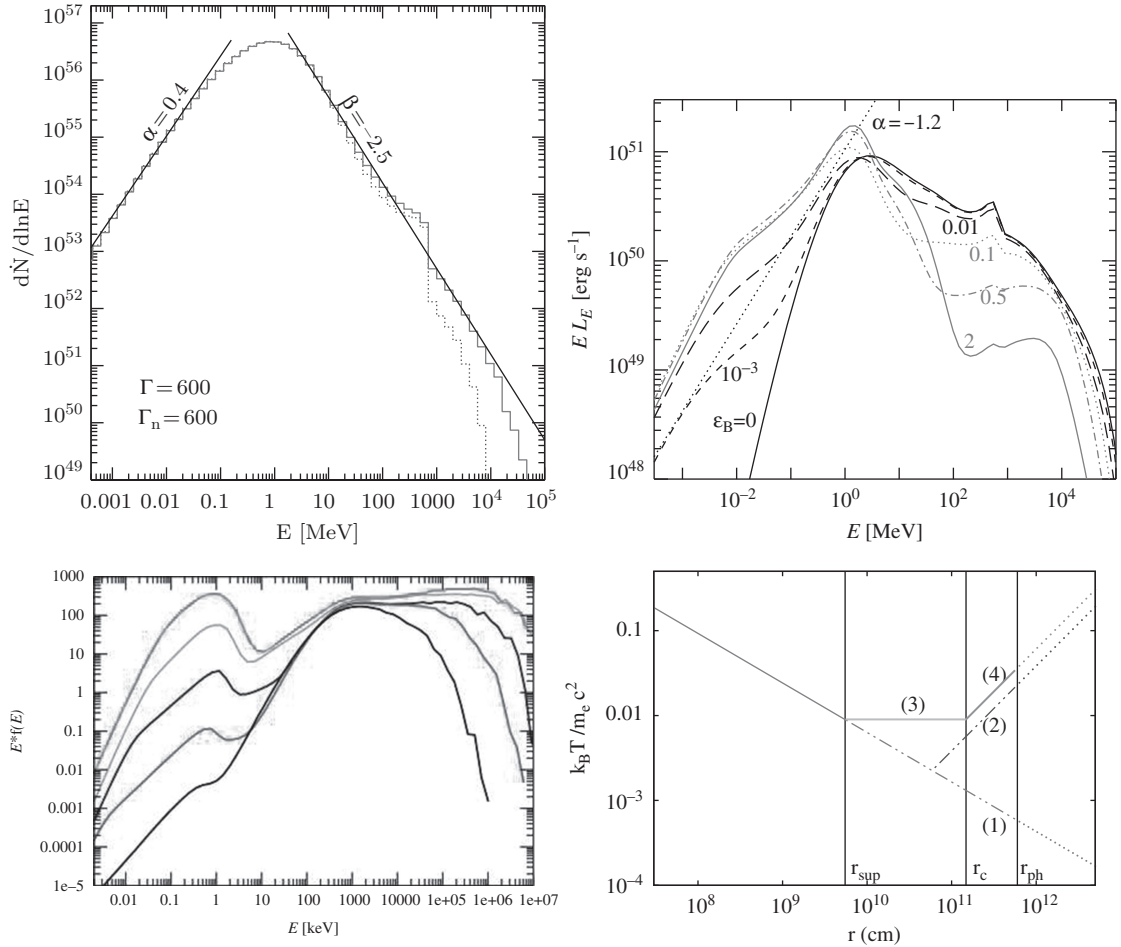
Figure 9.8   Predicted GRB spectra in various dissipative photosphere models. *Upper left*: The quasi-thermal spectrum predicted in the collisional heating photosphere model. From Beloborodov (2010). *Upper right*: The spectrum of the collisional heating model with synchrotron radiation included (different lines correspond to different values of σ (or ϵ_B)). Reproduced from Figure 5 in Vurm et al. (2011) with permission. ©AAS. *Lower left*: The predicted spectrum of the magnetic dissipative photosphere model assuming thermalization. From Giannios (2008). *Lower right*: The predicted comoving electron temperature evolution as a function of radius, which shows a significant deviation from thermalization with photons at large radii. Reproduced from Figure 1 in Bégué and Pe'er (2015) with permission. ©AAS.

9.4.4 Magnetically Dissipative Photosphere

Another possibility of having significant sub-photosphere dissipation is through magnetic dissipation. This is possible if the magnetic configuration is striped-wind-like (§7.4.1), plausible for a magnetar central engine. A detailed magnetic dissipative photosphere model has been proposed by several authors (Drenkhahn and Spruit, 2002; Drenkhahn, 2002; Giannios and Spruit, 2007; Giannios, 2008).

The basic physical picture is the following: a Poynting-flux-dominated outflow with a striped-wind configuration is launched from the central engine. Beyond the sonic point, the outflow still undergoes slow acceleration with

$$\Gamma = \Gamma_\infty \left(\frac{r}{r_s} \right)^{1/3} \simeq 150 r_{11}^{1/3} (\epsilon\Omega)_3^{1/3} \sigma_{0,3}^{1/3} = 150 r_{11}^{1/3} (\epsilon\Omega)_3^{1/3} \sigma_{ra,2}^{1/2} \qquad (9.50)$$

due to the continuous magnetic reconnection within the wind, where Ω is the angular velocity of the central engine rotator and $\epsilon = v_{rec}/v_A \sim 0.1$ defines how fast reconnection proceeds, i.e. the ratio between the reconnection velocity v_{rec} and the Alfvén velocity v_A, which is about 0.1 for high-σ flows (e.g. Lyubarsky, 2005). The parameter

$$\Gamma_\infty = \sigma_0 = \sigma_{ra}^{3/2} \qquad (9.51)$$

is the final Lorentz factor of the outflow, and σ_0 and $\sigma_{ra} = \sigma_0^{2/3}$ are the magnetization parameters at the central engine and at the sonic point (end of rapid acceleration phase), respectively.[7]

One major assumption of these models is complete thermalization of photons and electrons at sub-photosphere magnetic reconnection sites. Under this assumption, the emerged photospheric emission can be calculated (e.g. Giannios, 2008) (lower left panel of Fig. 9.8), which may have a variety of the low-energy photon index α values below the thermal peak. This model predicts a dominant prompt optical emission component, which is consistent with the case of the naked-eye GRB 080319B (Racusin et al., 2008). However, most GRBs have an optical emission not exceeding the extrapolation of the γ-ray flux to the optical band (Shen and Zhang, 2009). This model over-predicts optical flux in these GRBs.

There are some issues for this type of dissipative photosphere model in accounting for the main (Band) component of GRB emission. First, from the theoretical point of view, the total energy released at the photosphere would be at most $L_w/(1 + \sigma_{ph})$. In order to maintain the $\Gamma \propto r^{1/3}$ dynamics, one needs to have $\sigma > 1$. On the other hand, in order to have a bright photosphere emission component, one needs to have $\sigma_{ph} \sim 1$, otherwise the majority of the energy is still contained in the Poynting flux. This raises a "fine-tuning" problem to the magnetic photosphere model: σ_0 and σ_{ra} need to be such that at R_{ph}, σ drops to close to unity. Second, a closer scrutiny of energy dissipation and thermalization processes suggests that photon generation mechanisms (e.g. double Comptonization and bremsstrahlung) cannot generate enough photons to allow thermal equilibrium (Vurm et al., 2013; Beloborodov, 2013; Bégué and Pe'er, 2015). The electron temperature continues to increase with radius (lower right panel of Fig. 9.8). As a result, photons are up-scattered to much higher energies above at least 8 MeV (Bégué and Pe'er, 2015), a feature usually not observed in GRBs.

---

[7] Notice that in the literature discussing magnetically dissipative photosphere models (Drenkhahn and Spruit, 2002; Drenkhahn, 2002; Giannios and Spruit, 2007; Giannios, 2008), σ_0 is defined at the sonic point, which is σ_{ra} in our notation (see Gao and Zhang (2015) and §7.4.2 for a more detailed discussion of the dynamics of magnetic acceleration).

9.4.5 Summary

The dissipative photosphere models are widely discussed in the literature. This is because the model has some features that are attractive in interpreting GRB prompt emission of at least some GRBs. It is likely that the GRBs with a hard α (exceeding the synchrotron line of death, i.e. $\alpha > -2/3$) and a "narrow" Band-function spectrum are powered by (dissipative) photosphere emission. One clear example is GRB 090902B (Abdo et al., 2009b; Ryde et al., 2010; Zhang et al., 2011; Pe'er et al., 2012). This burst has an average value of $\alpha \sim +0.11$ for the time-resolved spectra, which is much harder than the typical value $\alpha < -0.7$ for most GRBs.

An interesting question is whether the dissipative photosphere can be extended to explain the dominant Band-function component in the spectra of all GRBs. This is a topic subject to intense debate in the community. Below, we list the arguments for and against this suggestion, and the author's opinion is that it is difficult for the dissipative photosphere models to explain the main Band-function component of all GRBs.

The arguments in favor of the dissipative photosphere models in interpreting the dominant GRB Band-function component (e.g. Beloborodov, 2013) include the following:

- The observed E_p in GRBs are typically hundreds of keV. This is close to the temperature of the GRB central engine (Eq. (7.44)). If the photosphere radius is above the coasting radius, E_p also depends on unknown parameters such as η, but in any case, E_p in the photosphere model relies on fewer free parameters than the synchrotron model. This argument requires that the dissipation is not in the saturated Comptonization regime, so that E_p is still defined by the seed photon temperature. The observed $E_p \sim 15$ MeV for GRB 110721A is higher than the maximum temperature allowed by the non-saturated dissipative photosphere, i.e. it is beyond the "death line" of the photosphere model (Zhang et al., 2012a). In order to explain such a high E_p, either synchrotron radiation is needed (Veres and Mészáros, 2012), or one needs to introduce a saturated Comptonization with electron temperature much higher than the seed photon temperature.

- It has been argued (e.g. Beloborodov, 2013; Axelsson and Borgonovo, 2015; Yu et al., 2016) that the shape of the Band function is too narrow for synchrotron radiation, but the thermal peak is narrow enough to fit the Band spectrum. Indeed, the mathematical shape of the Band function is narrower than the synchrotron function. Such an argument applies perfectly to GRB 090902B-like GRBs that show a narrow Band component with a hard α. Its applicability to GRBs in general, however, needs further scrutiny. This is because in high-energy astrophysics data analysis, due to the limited photon numbers, the data tend to adapt to a spectral model once it is defined. One should not compare one model with another model (e.g. believe that the Band function is the true representation of the data and compare the synchrotron function with the Band function). Rather, one should apply each model to directly confront the data. Indeed, Zhang et al. (2016b) showed that both the synchrotron model (Uhm and Zhang, 2014b) and the Band-function empirical model can fit the time-resolved spectra of the bright GBM burst GRB 130606B (whose α values are typically < -0.7) equally well.

- Afterglow modeling reveals an afterglow kinetic energy comparable to the emitted γ-ray energy, suggesting that GRB prompt emission has a relatively large radiative efficiency (e.g. Panaitescu and Kumar, 2001; Zhang et al., 2007a). Dissipative photosphere emission has a high radiative efficiency and naturally interprets the observations (Lazzati et al., 2009, 2013).
- Within the collapsar model of long GRBs, significant energy dissipation is expected as the jet breaks out from the stellar envelope, so that a fireball is "re-born" (Ghisellini et al., 2007; Lazzati et al., 2013).

The arguments against dissipative photosphere emission as the dominant mechanism in producing a Band-function emission component in the GRBs that do not show a narrow spectral component include the following:

- The observed typical low-energy photon index α of the Band function (usually < -0.7) is much too soft for the photosphere models (which typically predict $\alpha \sim +0.4$). The proposals to soften α have not been completely satisfactory: the structured jet model (Lundman et al., 2013) needs to invoke a specific structure (constant L but decreasing Γ) in order to enhance the high-latitude emission; invoking synchrotron emission to produce a soft bump in the low-energy regime (Vurm et al., 2011) gives rise to a spectral feature not commonly observed in GRBs.
- It is not easy to satisfy the R_{GRB} constraints provided by X-ray, optical, and GeV emission, as discussed in §9.2. The only way to avoid this is to argue that the sub-MeV emission is detached from all the emission in other wavelengths.
- It is difficult to interpret the observed spectral lags and the E_p evolution patterns with respect to the broad pulses (not rapid spikes) in the GRB lightcurves (Deng and Zhang, 2014b; Uhm and Zhang, 2016b).
- Theoretically, the physical conditions that demand sub-photospheric energy dissipation and photon generation are contrived (Vurm et al., 2013; Asano and Mészáros, 2013; Kumar and Zhang, 2015). The magnetic dissipative photosphere models predict a high E_p ($\sim$8 MeV) feature (Bégué and Pe'er, 2015), which is disfavored by the data.

9.5 The Role of Electron–Positron Pairs

In GRB prompt emission models, especially at the small radii where compactness is high, pair production and annihilation usually play important roles in defining GRB dynamics and radiation properties.

9.5.1 Pair Freeze-Out

Within the standard fireball picture, the initial fireball is composed of photons, electron–positron pairs, a small contamination of baryons, and their associated electrons to ensure neutrality.

As the fireball expands, e^+e^- pairs gradually cool and eventually annihilate to produce γ-rays. The annihilation radius $R_{\rm ann}$ is typically below the photosphere radius $R_{\rm ph}$, which we prove in the following (Kumar and Zhang, 2015).

Let us first write down the general photon Thomson scattering optical depth in the comoving frame:

$$\tau_{\rm es} \sim \sigma_{\rm T}\Delta'(n'_{\pm} + n'_e), \tag{9.52}$$

where Δ' is the comoving shell width, and $n'_{\pm}$ and n'_e are the comoving number density of pairs and baryon-associated electrons, respectively.

In the initial fireball, due to the huge two-photon pair production optical depth one has $n'_{\pm} \gg n'_e$, so that n'_e is negligible.

As the fireball expands, the comoving temperature drops, i.e. $T' \propto R^{-1}$. Since at the central engine $kT_0 \sim 2$ MeV $\sim 4m_ec^2$, soon the comoving temperature falls into the non-relativistic regime for electrons. The electrons move with a speed $\beta_e = v_e/c \ll 1$. The pair annihilation cross section in this regime is (Eq. (5.177))

$$\sigma^{e^{\pm}\to 2\gamma} \simeq \frac{3}{8}\sigma_{\rm T}\beta_e^{-1}, \tag{9.53}$$

which increases linearly as β_e decreases. Noticing $n_+ = n_- = n_{\pm}/2$, the comoving time scale for a positron to annihilate with an electron is

$$t'_{e^{\pm}\to 2\gamma} \simeq \frac{2}{\sigma^{e^{\pm}\to 2\gamma}n'_{\pm}\beta_e c} \simeq \frac{2}{(3/8)\sigma_{\rm T}n'_{\pm}c}. \tag{9.54}$$

In the comoving frame, the pair number density can be estimated as

$$n'_{\pm} = \frac{2(2\pi km_eT')^{3/2}}{h^3}\exp\left(-\frac{m_ec^2}{kT'}\right). \tag{9.55}$$

Letting $t'_{e^{\pm}\to 2\gamma} = t'_{\rm dyn} = r/c\Gamma$, one gets the condition that most pairs annihilate. In other words, pairs "freeze out". This usually happens during the fireball acceleration phase (Eq. (9.60)) below, so that $r/\Gamma \sim R_0$ is satisfied. Therefore the freeze-out condition becomes

$$\sigma_{\rm T}n'_{\pm}R_0 \sim \frac{16}{3}. \tag{9.56}$$

Applying Eq. (9.55), one gets the condition

$$T'^{3/2}_{\rm ann}\exp\left(-\frac{5.9 \times 10^9\ {\rm K}}{T'_{\rm ann}}\right) \simeq 165R^{-1}_{0,7}, \tag{9.57}$$

or

$$kT'_{\rm ann} \simeq 21.3\ {\rm keV}. \tag{9.58}$$

So the Lorentz factor at the annihilation radius is

$$\Gamma_{\rm ann} \sim T(R_0)/T'_{\rm ann} \sim 61L^{1/4}_{w,52}R^{-1/2}_{0,7}. \tag{9.59}$$

This is lower than the typical Lorentz factor of GRBs. Therefore, the assumption that the annihilation radius is below the coasting radius is justified, and the annihilation radius can be written as

$$R_{\text{ann}} \sim \Gamma_{\text{ann}} R_0 \sim (6.1 \times 10^8 \text{ cm}) L_{w,52}^{1/4} R_{0,7}^{1/2}. \tag{9.60}$$

Above R_{ann}, one has $n'_{\pm} \ll n'_e$, so that essentially only n'_e defines the Thomson optical depth (Eq. (9.52)). This justifies the treatment of the baryonic photosphere in §7.3.3.

9.5.2 Pair Photosphere

Since most of the primordial pairs have annihilated below the photosphere, pairs must be re-generated in order to play an important role. In general, in the widely discussed GRB emission sites (e.g. R_{ph}, R_{IS} (§9.6), R_{ICMART} (§9.8), and R_{dec}), the contribution of the initial fireball pairs may be neglected. However, under certain conditions, various processes, such as two-photon pair production ($\gamma\gamma \to e^{\pm}$), the Bethe–Heitler process ($p\gamma \to pe^+e^-$), and hadronic cascade via $p\gamma$ interactions (Eq. (6.2)), can produce new pairs in the emission site. If the pair multiplicity parameter exceeds unity, i.e. $n'_{\pm} > n'_e$, then pairs play an important role in GRB emission.

Let us consider the situation that copious high-energy photons are produced around the photosphere radius, and photons with $E_{\gamma} > E_{\gamma,c}$ are absorbed and produce pairs with energy $E_{\pm} \sim E_{\gamma}/2$. If the high-energy photons have a power-law distribution then the generated pairs also have a power-law energy distribution. Since pairs have a rest mass of 0.511 MeV, the minimum Lorentz factor of the injected pairs reads

$$\gamma_{\pm,m} = 978(E_{\gamma,\text{cut}}/1 \text{ GeV}). \tag{9.61}$$

These pairs cool rapidly via synchrotron and Compton scattering and would quickly pile up around $\gamma_{\pm} \sim 1$. Due to their huge number (compared to the baryon-associated electrons) they would increase the Thomson optical depth significantly, and would increase the photosphere radius (Kobayashi et al., 2002; Mészáros et al., 2002). Within the framework of the fireball, and assuming a significant fraction of internal energy goes to pairs, one can estimate the radius of the pair photosphere as (e.g. Kobayashi et al., 2002; Mészáros et al., 2002)

$$R_{\text{ph},\pm} \sim \left(\frac{\sigma_T \epsilon_e E_{\text{iso}}}{40\pi mc^2 N\Gamma} \right)^{1/2}$$
$$\simeq (5 \times 10^{14} \text{ cm}) \, \epsilon_{e,-1}^{1/2} E_{52}^{1/2} N_2^{-1/2} \Gamma_2^{-1/2} (\theta_j/0.2)^{-1/2}, \tag{9.62}$$

where $N = 100 N_2$ is the number of pulses in the GRB, Γ is the average for the shells, and ϵ_e is the fraction of the internal energy that is given to pairs. This is much larger than the baryonic photosphere radius R_{ph} (Eqs. (7.65) and (7.68)). As a result, significant pair production would effectively place the photosphere at a larger radius.

9.5.3 Self-Consistent Numerical Modeling of Pair-Rich Fireballs

No simple analytical solution is available for the emission of a system invoking pair production and annihilation. Numerical modeling is needed (e.g. Pilla and Loeb, 1998; Pe'er and Waxman, 2005; Pe'er et al., 2006).

A self-consistent solution of the problem requires solving a set of time-dependent differential equations for the number densities of electrons (n_{e-}), positrons (n_{e+}), protons (n_p), and photons (n_{ph}) as a function of energy (γ for particles and $\epsilon = \hbar\omega$ for photons) and as a function of time (Pe'er and Waxman, 2005):

$$\frac{\partial n_{e-}(\gamma, t)}{\partial t} = Q_e(\gamma, t) + \frac{\partial}{\partial\gamma}\left\{n_{e-}(\gamma, t)[P_{\text{syn}}(\gamma, t) + P_{\text{IC}}(\gamma, t)]\right.$$
$$\left. + H(\gamma, t)\beta\gamma^2\frac{\partial}{\partial\gamma}\left[\frac{n_{e-}(\gamma, t)}{\beta\gamma^2}\right]\right\} + Q_{\pm}^+ - Q_{\pm}^-, \quad (9.63)$$

$$\frac{\partial n_{e+}(\gamma, t)}{\partial t} = \frac{\partial}{\partial\gamma}\left\{n_{e+}(\gamma, t)[P_{\text{syn}}(\gamma, t) + P_{\text{IC}}(\gamma, t)] + Q_{\pi}(\gamma, t)\right.$$
$$\left. + H(\gamma, t)\beta\gamma^2\frac{\partial}{\partial\gamma}\left[\frac{n_{e+}(\gamma, t)}{\beta\gamma^2}\right]\right\} + Q_{\pm}^+ - Q_{\pm}^-, \quad (9.64)$$

$$\frac{\partial n_p(\gamma, t)}{\partial t} = Q_p(\gamma, t) + \frac{\partial}{\partial\gamma}[n_p(\gamma, t)P_{\pi}(\gamma, t)], \quad (9.65)$$

$$\frac{\partial n_{ph}(\epsilon, t)}{\partial t} = R_{\text{syn}}(\epsilon, t) + R_{\text{IC}}(\epsilon, t) - R_{\pm}^+(\epsilon, t) + R_{\pm}^-(\epsilon, t)$$
$$+ R_{\pi}(\epsilon, t) - cn_{ph}(\epsilon, t)\alpha(\epsilon, t). \quad (9.66)$$

Here $Q_e(\gamma, t)$ and $Q_p(\gamma, t)$ are the injection rates of electrons and protons in the system (e.g. from internal shocks or reconnection sites); $Q_{\pi}(\gamma, t)$ is the injection rate of positrons through $p\gamma$ interactions;

$$H(\gamma, t) = \int d\omega \frac{I_{\omega}(t)}{4\pi m_e\omega^2}P(\omega, \gamma) \quad (9.67)$$

denotes heating of electrons and positrons and their diffusion in energy due to synchrotron self-absorption ($I_{\omega}(t) = n(\epsilon, t)\epsilon c\hbar/4\pi$ is the specific intensity, and $P(\omega, \gamma)$ is the total synchrotron emission power per unit angular frequency ω of an electron with Lorentz factor γ); $Q_{\pm}^+$ and $Q_{\pm}^-$ are pair production rate and pair annihilation rate, respectively; $P_{\text{syn}}(\gamma, t)$ and $P_{\text{IC}}(\gamma, t)$ are the emission powers (or energy loss rates) of an electron/positron through synchrotron radiation and inverse Compton scattering, respectively; $P_{\pi}(\gamma, t)$ is the power of protons to transfer energy to pions; $R_{\text{syn}}(\epsilon, t)$ and $R_{\text{IC}}(\epsilon, t)$ are photon generation rates via synchrotron and inverse Compton, respectively; $R_{\pm}^+(\epsilon, t)$ and $R_{\pm}^-(\epsilon, t)$ are the photon loss rates due to pair production and photon generation due to pair annihilation, respectively; $R_{\pi}(\epsilon, t)$ is the photon generation rate due to decay of energetic π^0; and $\alpha(\epsilon, t)$ is the self-absorption coefficient.

Detailed numerical modeling of GRB spectra has been carried out (e.g. Pe'er and Waxman, 2005; Pe'er et al., 2006). The resulting spectra depend on the "compactness" of the fireball (Fig. 9.9). At low compactness the role of pairs is minimized, and emerging spectra are close to the optically thin spectra, as discussed in later sections. However, at high
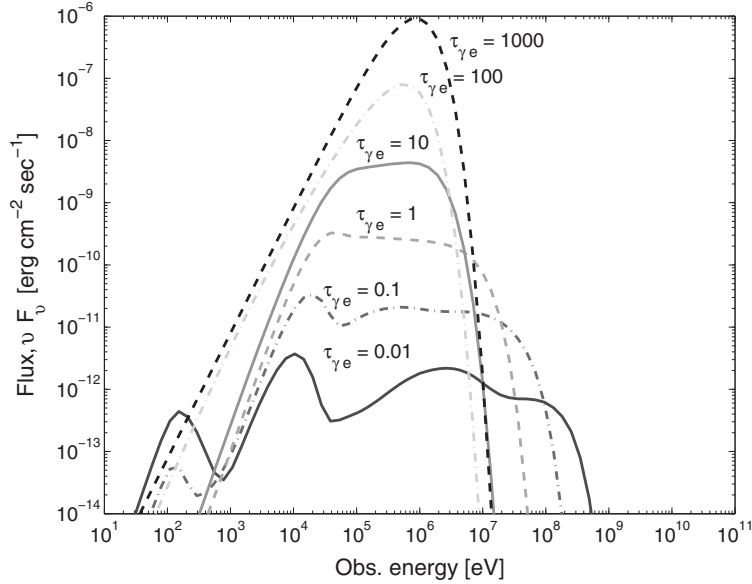
Figure 9.9 The numerical GRB prompt emission spectra for different compactness parameters. Reproduced from Figure 1 in Pe'er et al. (2006) with permission. ©AAS.

compactness, pairs play an essential role in shaping the GRB prompt emission spectrum. Under the high-compactness condition, one can produce a Comptonized thermal spectrum that mimics a Band-function spectrum (Pe'er et al., 2006).

In a pair-dominated fireball, a pair annihilation line is predicted. Since the pair annihilation cross section increases linearly with decreasing $\beta_\pm$, annihilations mostly proceed in the non-relativistic regime, so that the comoving-frame photon energy is at the electron rest mass energy. In the observer frame, one predicts a broad line at the energy (e.g. Murase et al., 2008)

$$E_{\pm,\text{line}} = \frac{\Gamma}{1+z} \cdot 0.511 \text{ MeV}. \tag{9.68}$$

Identifying such a line emission would suggest a pair-dominated emission region and lead to a direct measurement of the bulk Lorentz factor of the emitting region.

9.6 Internal Shocks

9.6.1 General Considerations

After introducing all the photosphere-related models, starting from this section, we discuss the models that invoke synchrotron radiation in optically thin regions. The first model to discuss is the *internal shock model*.

The internal shock model (Rees and Mészáros, 1994; Paczyński and Xu, 1994; Kobayashi et al., 1997; Daigne and Mochkovitch, 1998) was regarded as the standard model for GRB prompt emission before the beginning of the *Fermi* era. The rationale and merits of the model include the following:

- Internal shocks are naturally expected for an erratic central engine, which launches an unsteady central engine wind. This is likely relevant to catastrophic events such as GRBs. The highly variable lightcurves may manifest the erratic central engine activity. Indeed, simulations (Kobayashi et al., 1997; Maxham and Zhang, 2009; Gao and Mészáros, 2015) suggest the time sequence as observed in a GRB lightcurve mostly reflects the time history of the central engine activity.

- Internal shocks are natural sites for dissipating kinetic energy of a baryonic "fireball". The early developments of the GRB fireball models suggested that once a small amount of baryons has been added to the fireball, a significant energy is converted to the kinetic energy of the outflow (Shemi and Piran, 1990). In order to reconvert the kinetic energy back to random energy and particle radiation, shocks are the most natural sites. This was the key ingredient of the fireball shock model proposed by Rees and Mészáros (1992), Mészáros and Rees (1993b), and Rees and Mészáros (1994).

- Internal shocks are natural sites for particle acceleration and non-thermal radiation. It has been well established that non-thermal particles are accelerated due to the first-order Fermi acceleration process in shocks (Spitkovsky, 2008; Sironi and Spitkovsky, 2009a, 2011). The same shocks also amplify magnetic fields through plasma (Weibel, 1959; Medvedev and Loeb, 1999; Nishikawa et al., 2005) or fluid instabilities (Sironi and Goodman, 2007). As a result, synchrotron radiation is naturally produced, which likely contributes to the observed γ-ray emission from GRBs.

9.6.2 Dynamics and Efficiency

The interaction between two matter-dominated shells is similar to jet–medium interaction at the external shock. Upon collision, a pair of shocks propagate into the two shells. A four-region structure is formed (§4.3), with the four regions separated by the forward shock (FS), contact discontinuity (CD), and the reverse shock (RS). The relative strength of the two shocks depends on the Lorentz factor and mass of each shell. In particular, the ratios γ_4/γ_1 and L_4/L_1 ($L = \gamma \dot{M} c^2$, and the subscripts 4 and 1 denote Regions IV and I of the four-region structure, which correspond to the unshocked fast and slow shells, respectively) define which shock is relativistic (e.g. Yu and Dai, 2009). In general, the RS is stronger and more likely to be relativistic. Similar to the external FS and RS, the internal FS and RS both accelerate particles. The accelerated electrons radiate synchrotron photons and power the observed γ-rays. Accelerated protons are also expected, which would power neutrino emission from GRBs (§12.2 for a more detailed discussion).

In most internal shock modeling, the details of FS/RS dynamics are ignored. One may more generally consider the dissipation efficiency of internal shocks. Consider a trailing, faster shell with mass m_2 and Lorentz factor Γ_2 catching up with a leading, slower shell with mass m_1 and Lorentz factor Γ_1. The highest efficiency for a collision is achieved by

assuming that the collision is fully inelastic, i.e. the two shells merge and stick together with a Lorentz factor Γ_m after the collision. Energy and momentum conservation can be written as

$$\Gamma_1 m_1 + \Gamma_2 m_2 = \Gamma_m(m_1 + m_2 + \hat{\gamma}U/c^2), \tag{9.69}$$

$$\Gamma_1 \beta_1 m_1 + \Gamma_2 \beta_2 m_2 = \Gamma_m \beta_m(m_1 + m_2 + \hat{\gamma}U/c^2), \tag{9.70}$$

where U is the internal energy of the merged shell. Solving for Γ_m, one gets (Kobayashi et al., 1997)

$$\Gamma_m = \left(\frac{\Gamma_1 m_1 + \Gamma_2 m_2}{m_1/\Gamma_1 + m_2/\Gamma_2} \right)^{1/2}. \tag{9.71}$$

The *efficiency* of internal shock dissipation is therefore (Exercise 9.2)

$$\eta_{\mathrm{IS}} = \frac{\Gamma_m \hat{\gamma} U}{\Gamma_1 m_1 c^2 + \Gamma_2 m_2 c^2} = 1 - \frac{m_1 + m_2}{\sqrt{m_1^2 + m_2^2 + m_1 m_2 \left(\frac{\Gamma_2}{\Gamma_1} + \frac{\Gamma_1}{\Gamma_2} \right)}}. \tag{9.72}$$

The efficiency depends on the mass ratio m_2/m_1 and the relative Lorentz factor Γ_2/Γ_1. Fixing Γ_2/Γ_1, an optimal efficiency is achieved for equal mass shells, i.e. $m_2/m_1 = 1$. Fixing m_2/m_1, a larger relative Lorentz factor gives a higher efficiency. For reasonable parameters, e.g. Γ_2/Γ_1 not too large, the efficiency η_{IS} is typically 1–10%.

The observed γ-ray radiative efficiency is

$$\eta_{\gamma,\mathrm{IS}} = \eta_{\mathrm{IS}} \cdot \epsilon_e \cdot \min[1, (\gamma_c/\gamma_m)^{2-p}], \tag{9.73}$$

which is even smaller by a factor of at least $\epsilon_e \sim 0.1$. So internal shocks are inefficient in producing γ-rays, although more efficient than the external shock to give the same variability as observed.

9.6.3 Lightcurves

The internal shock model can give a straightforward interpretation of the GRB lightcurves, in particular the observed variability. The internal shock radius is (§3.5.2)

$$R_{\mathrm{IS}} \simeq 2\Gamma^2 c\delta t = (6 \times 10^{12}\ \mathrm{cm})\,\Gamma_2^2 \delta t_{-2}. \tag{9.74}$$

The observer time when emission from an IS is seen (from the time of the very first central engine activity) is

$$t_{\mathrm{obs}} \simeq t_{\mathrm{ej}} + \frac{R_{\mathrm{IS}}}{2\Gamma^2 c} \simeq t_{\mathrm{ej}} + \delta t \simeq t_{\mathrm{ej}}, \tag{9.75}$$

where t_{ej} is the ejection time of the central engine since the beginning of engine activity, which is typically much longer than δt except during the very early phase of the GRB. As a result, the observed time history essentially reflects the time history of the central engine (Kobayashi et al., 1997; Maxham and Zhang, 2009; Gao and Mészáros, 2015). The upper left panel of Fig. 9.10 shows a comparison between t_{ej} and t_{obs} (marked as $t_{\oplus}$ in the plot) values from many shells in a Monte Carlo simulation (Maxham and Zhang, 2009). One
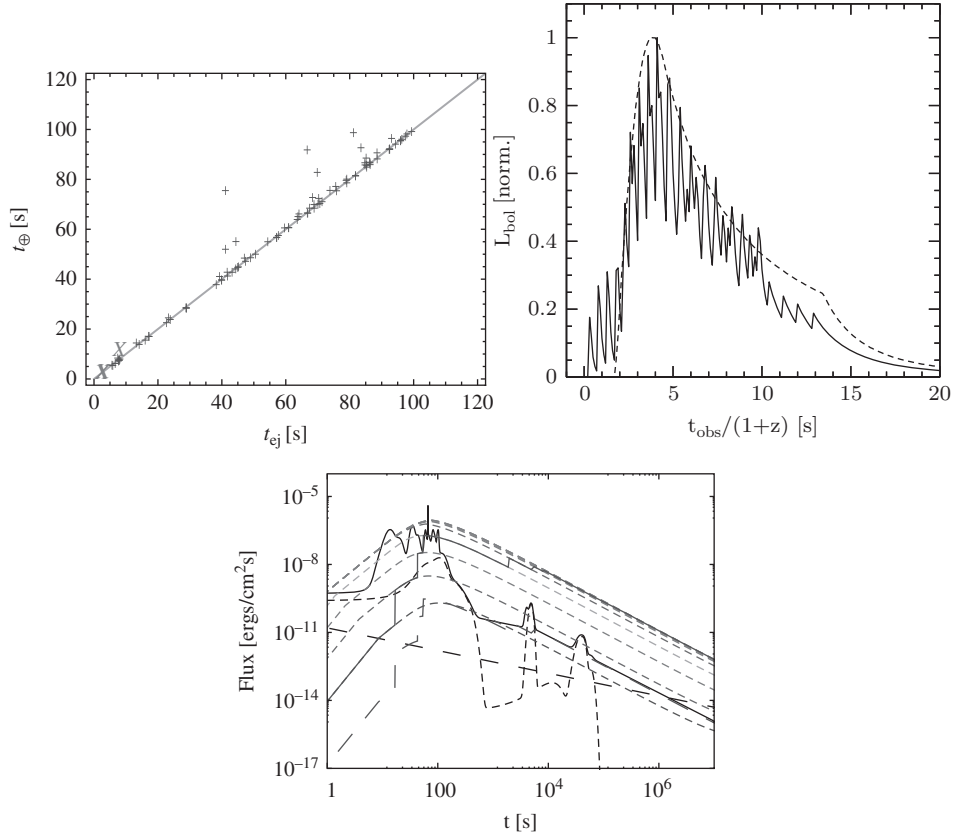
Figure 9.10 Some model predictions of the internal shock model. *Upper left:* A comparison between t_{ej} and t_{obs} in the internal shock model from a Monte Carlo simulation. From Maxham and Zhang (2009). *Upper right:* The simulated internal shock lightcurve invoking convolution of slow and fast variability components. From Hascoët et al. (2012a). *Lower:* The simulated lightcurve for the internal–external shock model: the spiky curve is the internal shock lightcurve, including X-ray flares; upper solid curve with zig-zag features is the indicative external shock lightcurve with energy injection with the typical parameter $\epsilon_e = 0.1$; and lower long-dashed zig-zagged curve that merges to the lower solid curve is the similar external shock lightcurve with an abnormally low parameter $\epsilon_e = 10^{-3}$. From Maxham and Zhang (2009).

can see that, except for a few collisions (which typically have very small relative Lorentz factors), the majority of collisions line up at the $t_{obs} = t_{ej}$ line.

Observationally, GRB lightcurves show the superposition of slow and fast variability components (e.g. Gao et al., 2012). The straightforward simulation of lightcurves (upper left panel of Fig. 9.2) does not show such a superposition effect. However, if one introduces a convolution of slow and fast variability components at the central engine, a lightcurve with both slow and fast variability components may be reproduced (Hascoët et al., 2012a) (upper right panel of Fig. 9.10).

A combined internal shock + external shock lightcurve is presented in the lower panel of Fig. 9.10. By ejecting random shells from the central engine with a certain distribution of shell masses and Lorentz factors, Maxham and Zhang (2009) tracked all the collisions

and assigned a pulse for each collision with the total energy defined by the total internal energy defined by η_{IS} (Eq. (9.72)). The left-over energy is dumped to the blastwave, which is tracked dynamically as more and more shells pile up onto it. Such a process naturally gives rise to several interesting observed properties of GRBs, including the spiky prompt emission lightcurves, X-ray flares (if multiple shell ejection episodes are assumed), as well as an early shallow decay phase of the X-ray afterglow (upper left panel of Fig. 9.10). However, in order to produce an early steep decay phase of X-ray afterglow, one has to artificially lower the afterglow level (e.g. decreasing ϵ_e from 0.1 to 0.001), or artificially increase the radiative efficiency of the prompt emission. This is another manifestation of the efficiency problem of the internal shock model.

9.6.4 Spectra

The leading radiation mechanism of the internal shock model is synchrotron radiation (Rees and Mészáros, 1994; Mészáros et al., 1994; Tavani, 1996; Daigne and Mochkovitch, 1998; Daigne et al., 2011). However, the simplest synchrotron model has several difficulties in interpreting the observations, which we discuss in the following.

E_p and Electron Number Problem

For synchrotron emission from internal shocks, usually electrons are in the fast cooling regime (see below). The E_p is therefore determined by the minimum injection energy of the electrons, i.e. (Exercise 9.3)

$$E_p \sim \hbar \Gamma \gamma_m^2 \frac{eB'}{mc}(1+z)^{-1}. \tag{9.76}$$

Assuming $\epsilon_e + \epsilon_B + \epsilon_p = 1$, $\xi_e = \xi_p = 1$ (ξ_e and ξ_p are the fractions of electrons and protons that are accelerated), one has[8]

$$\bar{\gamma}_p - 1 = (\gamma_{43} - 1)\epsilon_p, \tag{9.77}$$

$$\bar{\gamma}_e - 1 = \frac{\epsilon_e}{\epsilon_p} \frac{n_p}{n_e} \frac{m_p}{m_e}(\bar{\gamma}_p - 1), \tag{9.78}$$

and

$$\frac{B'^2}{8\pi} = \frac{L_w \eta_{\mathrm{IS}} \epsilon_B}{4\pi R^2 c \Gamma^2}. \tag{9.79}$$

Substituting them into (9.76), one gets

$$E_{p,\mathrm{IS}} \simeq 4.4\,\mathrm{keV}\, L_{\gamma,52}^{1/2} R_{\mathrm{IS},14}^{-1} \left(\frac{1+z}{2}\right)^{-1} \left(\frac{\epsilon_B}{\epsilon_e}\right)^{1/2} \left(\frac{\epsilon_e}{\epsilon_p}\frac{n_p}{n_e}\right)^2 (\bar{\gamma}_p - 1)^2 \ll 300\,\mathrm{keV}. \tag{9.80}$$

So the simplest internal shock model cannot correctly reproduce the observed E_p, which is in the sub-MeV range (typically $\sim$200–300 keV).

---

[8] The internal RS emission is calculated, since it typically has a higher E_p. Notations follow the convention, e.g. γ_{43} is the relative Lorentz factor between Regions IV (unshocked trailing shell) and III (shocked trailing shell) as defined in §4.3.

In order to match the observational constraints, one has to assume (Daigne and Mochkovitch, 1998) that only a small fraction of electrons are accelerated, or

$$\frac{n_p}{n_e} \gg 1, \quad \text{or} \quad \xi_e \ll 1. \tag{9.81}$$

The observed synchrotron flux also demands that only a small fraction of electrons are accelerated. This is also related to estimating the synchrotron self-absorption frequency ν_a in the internal shock model. As discussed in §5.1.6, there are two methods to estimate ν_a. One is to directly compare the observed synchrotron flux and a blackbody flux, and to find the frequency at which the two fluxes are equal (blackbody method). The other is to calculate the source function of the emitting electrons. The second method needs knowledge about the total number of electrons. If one uses the observed GRB luminosity and inferred Lorentz factor (assuming an efficiency) to calculate electron number and then calculate ν_a using the second method, it is often inconsistent with the ν_a derived from the first method (which replies on the observed flux only). In order to achieve self-consistency, again one has to assume that only a small fraction of electrons are accelerated (Shen and Zhang, 2009).

If indeed only a small fraction of electrons are accelerated, the majority of electrons would form a thermal bump, whose synchrotron radiation would show a bump feature in the spectrum. Such a feature is not observed from the GRB spectral data.

Fast Cooling Problem

In order to match the observed E_p, one needs to have

$$\gamma_{e,p} \simeq 2.3 \times 10^3 L_{\gamma,52}^{-1/4} r_{14}^{1/2} \left(\frac{\epsilon_e}{\epsilon_B}\right)^{1/4} \left(\frac{1+z}{2}\right)^{1/2} \left(\frac{E_p}{250\,\text{keV}}\right)^{1/2}. \tag{9.82}$$

With the B' field derived from L and r, one can estimate the comoving cooling time scale of the electrons:

$$t_c' \sim 0.008\,\text{s} \ll t_{\text{dyn}}'. \tag{9.83}$$

This suggests $\gamma_c \ll \gamma_{e,p}$. So the spectrum below E_p should be in the deep fast cooling regime. The predicted photon spectral index is $\alpha = -1.5$. This is inconsistent with the typically observed spectral index $\alpha \sim -1$. This fast cooling problem has been raised by various authors in several different forms (Ghisellini et al., 2000; Kumar and McMahon, 2008; Kumar and Zhang, 2015).

Several proposals have been suggested to alleviate the fast cooling problem of the internal shock models:

- Pe'er and Zhang (2006) suggested that magnetic fields generated at the shock front due to plasma instabilities rapidly decay with radius in the downstream, so that electrons only undergo rapid cooling briefly and soon enter the slow cooling regime further downstream. The introduction of SSC cooling within this scenario can help to reach harder spectra to match the data (Zhao et al., 2014).

- An alternative mechanism is to introduce post-shock-front turbulent heating of electrons, so that electrons are in the regime of slow heating balancing fast cooling (Asano and Terasawa, 2009; Xu and Zhang, 2017). Invoking hadronic processes from the accelerated protons introduces rich features, and can effectively harden the spectrum to the observed value under certain conditions (Asano and Mészáros, 2011; Murase et al., 2012). These hadronic models usually introduce a large proton-to-electron energy ratio in order to enhance hadronic processes, which is subject to the constraints from the progressively stringent upper limits of the high-energy neutrino flux from GRBs (see more discussion in §12.2).
- Uhm and Zhang (2014b) realized that the magnetic field strength of an expanding jet naturally decreases with time due to flux conservation as the jet cross section increases. For a large emission radius (e.g. $R_{\rm GRB} \sim 10^{15}$ cm, as inferred from various observational constraints in §9.2), fast cooling of electrons in such a decaying magnetic field modifies the emission spectrum from the standard $\alpha = -1.5$ value. Depending on the parameters, the α value can be ~ -1 and even as hard as -0.8 (§5.1.5).
- Along a different direction, Daigne et al. (2011) introduced IC cooling in the Klein–Nishina (KN) regime to modify the electron spectrum to reach $\alpha \sim -1$ in the internal shocks. In order to reach the KN regime, γ_e needs to be very large, which requires a very small ϵ_B ($\sim 10^{-5}$) to keep the same observed E_p. Such a low-B model would inflate the bright photosphere emission problem discussed next.

Bright Photosphere Emission Component

In order to drive internal shocks, the composition of the ejecta has to be a matter-dominated "fireball", which initially undergoes a thermally driven acceleration phase and generates a bright quasi-thermal emission as the jet passes the photosphere. Calculations show that usually this photosphere component is bright and outshines the non-thermal synchrotron component from the internal shocks in most of the parameter space (Mészáros and Rees, 2000b; Gao and Zhang, 2015).

Observationally, bright photosphere emission is indeed seen in a fraction of GRBs, such as GRB 090902B (Abdo et al., 2009b; Ryde et al., 2010). The dominant component in these GRBs is likely of a photospheric origin (Pe'er et al., 2012). The excess optical, X-ray, and possible GeV emission of these bursts could have an internal shock origin (Asano et al., 2010; Pe'er et al., 2012).

In the majority of GRBs, the photosphere component is not prominent (Guiriec et al., 2011; Axelsson et al., 2012; Guiriec et al., 2013) or not detected (Abdo et al., 2009c; Zhang et al., 2016b). This usually requires a magnetized central engine (Zhang and Pe'er, 2009; Gao and Zhang, 2015) to suppress the photosphere component such that the synchrotron component can become the dominant spectral component. In a certain parameter regime (σ_0 not too large and η not too small), one may have the photosphere emission suppressed but in the internal shock radius σ already drops below unity to allow internal shocks to develop (Hascoët et al., 2013; Gao and Zhang, 2015). For a wider range of (η, σ_0) initial conditions, it is likely that either the photosphere is not suppressed or σ at the typical internal shock radius is still above unity (Gao and Zhang, 2015). In these cases, the internal

shocks would not be the dominant particle acceleration and synchrotron emission sites. Rather, strong synchrotron emission may originate from regions where internal-collision-induced magnetic reconnection and turbulence (ICMART) occurs (Zhang and Yan, 2011).

9.6.5 Summary

Due to the intermittent nature of the GRB central engine, as evidenced in the observed erratic lightcurves, internal shocks are expected to develop at the internal shock radius (Eq. (9.74)).[9] As a result, synchrotron radiation in internal shocks should contribute to the observed GRB emission in most GRBs. The question is whether internal shocks power the dominant Band-function component as observed in most GRBs.

The issues and possible solutions of the internal shock model in interpreting the dominant GRB spectral component can be summarized as follows:

- Efficiency is not very high. The following two arguments may alleviate the problem. First, the relative Lorentz factors in GRBs may be larger than normally expected. Also the inelastic assumption may not be valid, so that shells can undergo multiple collisions to further dissipate energy. All these effects tend to increase the internal shock energy dissipation efficiency (Beloborodov, 2000; Kobayashi and Sari, 2000). Second, recent afterglow modeling suggests that ϵ_B and even ϵ_e in the afterglow region may not be as high as previously assumed (e.g. Kumar and Barniol Duran, 2009; Santana et al., 2014; Wang et al., 2015b; Gao et al., 2015a). This lowers the required GRB efficiency, so that internal shocks may provide adequate efficiency for interpreting at least some GRBs.
- The predicted E_p is too low unless only a small fraction of electrons are allowed to accelerate. For non-relativistic shocks, such as supernova remnant shocks, indeed a small fraction of electrons are accelerated, but the emission from the thermal electrons is clearly observed. For GRB afterglow emission, which invokes relativistic shocks, no evidence of emission from a thermal electron population is observed. This suggests that relativistic shocks may be more efficient in accelerating particles. Internal shocks are typically trans-relativistic, and might form a thermal population of electrons. However, the observed spectra also do not show evidence of this population.
- Since a small variability time scale with δt as small as milliseconds has been observed, given the inferred Lorentz factor, the internal shock radius R_{IS} is usually small enough that synchrotron radiation is in the deep fast cooling regime. The predicted spectrum is too soft to interpret the data. A very rapid magnetic field decay in the shock downstream or slow turbulent heating may help to solve the problem.
- The weak or vanishing thermal component in most GRBs requires $\sigma_0 \gg 1$ at the central engine, so that at the internal shock radius σ is likely still above unity. This would further reduce radiative efficiency from internal shocks by a factor $1 + \sigma$.
- The internal shock model also has difficulty in explaining the so-called Amati/Yonetoku relations observed in GRB prompt emission, which shows roughly $E_p \propto E^{1/2} \propto L^{1/2}$.

---

[9] For $\sigma > 1$, internal shocks can still develop. However, without dissipating magnetic energy, the internal shock efficiency is further suppressed by a factor of $(1 + \sigma)$, making internal shock emission even weaker (Zhang and Yan, 2011; Narayan et al., 2011).

The argument is the following. In the expression of E_p in the synchrotron radiation model (9.76), the factor

$$\Gamma B' = B \propto \left(\frac{L}{4\pi r^2 c} \right)^{1/2} \propto L^{1/2} r^{-1}. \tag{9.84}$$

So, in general, Eq. (9.76) can be expressed as (Zhang and Mészáros, 2002a)

$$E_p \propto \gamma_{e,p}^2 L^{1/2} R_{\mathrm{IS}}^{-1} \propto \gamma_{e,p}^2 L^{1/2} \Gamma^{-2} \delta t^{-1}, \tag{9.85}$$

with the last proportionality applying to the internal shock model. Since $\gamma_{e,p}$ depends on the strength of the internal shock (i.e. the relative Lorentz factor between the colliding shells), one may take it to be not strongly dependent on other parameters. In order to interpret the Amati/Yonetoku relation $E_p \propto L^{1/2}$, one has to assume $R_{\mathrm{IS}} \sim$ const for bursts with different L. Within the internal shock models this suggests $\Gamma \sim$ const, since the minimum variability time scale δt may be similar among bursts. However, observations show that there is a correlation between Γ and L, e.g. $\Gamma \propto E^{1/4} \propto L^{1/4}$ (Liang et al., 2010; Lü et al., 2012). Substituting this correlation in the expression of E_p, one gets $E_p \propto L^0$. This is inconsistent with the Amati/Yonetoku relations. Mochkovitch and Nava (2015) modeled the E_p–E_{iso} relation within the internal shock model, and suggested that the model can be made consistent with the data only if several strong constraints are satisfied on both the dynamics of the flow and the microphysics that governs the redistribution of the shock-dissipated energy.

In summary, even though the internal shock model has many attractive features in interpreting GRB prompt emission, there exist several drawbacks that are not easy to overcome. These were the main motivations to develop the photosphere models (as discussed in §9.3 and §9.4 above) and the optically thin magnetic dissipation models, which we discuss next.

9.7 Magnetic Dissipation in an Optically Thin Region

If the GRB central engine launches a Poynting-flux-dominated outflow, as expected in the central engine models invoking, e.g. the Blandford–Znajek mechanism to tap the BH spin energy, or a millisecond magnetar that spins down due to magnetic dipole radiation, magnetic dissipation likely happens within the jet. If significant dissipation is suppressed below the photosphere, it is likely that a large-scale magnetic field with an ordered magnetic field configuration would be advected with the jet to large radii, at which significant magnetic dissipation may happen to power GRB prompt emission through forced magnetic reconnection or current-driven instabilities. On the other hand, the transportation and dissipation of large-scale magnetic fields in GRB jets are complicated physical problems, and no conclusive results have been obtained. Below, we introduce several proposed scenarios.

9.7.1 GRB Ejecta as Plasma

Before discussing the details of magnetic dissipation in a GRB jet, it is informative to summarize the GRB ejecta parameters within the context of plasma physics (Exercise 9.4). The following derivations follow Zhang and Yan (2011). The typical GRB emission radius is normalized to $R_{\rm GRB} = 10^{15}$ cm, which corresponds to the radius for the emission with a "slow" (second-duration) variability component, e.g. $R_{\rm GRB} = \Gamma^2 c \delta t_{\rm slow} = 3 \times 10^{15}$ cm $\Gamma_{2.5}^2 \delta t_{\rm slow}$.

- *Length scales*: The "thickness" of the ejecta is

$$\Delta' \sim \frac{R_{\rm GRB}}{\Gamma} \simeq (3.2 \times 10^{12} \text{ cm}) R_{\rm GRB,15} \Gamma_{2.5}^{-1} \tag{9.86}$$

in the comoving frame, and

$$\Delta \sim \frac{R_{\rm GRB}}{\Gamma^2} \simeq 10^{10} \text{ cm} R_{\rm GRB,15} \Gamma_{2.5}^{-2} \tag{9.87}$$

in the lab frame. For a conical jet with opening angle θ_j, the cross section radius of the emission region is

$$R_\theta \sim R_{\rm GRB} \theta_j = 8.7 \times 10^{13} \left(\frac{\theta_j}{5^{\rm o}} \right) R_{\rm GRB,15} \text{ cm} \tag{9.88}$$

in both the lab frame and the comoving frame. The condition $R_\theta \gg \Delta' \gg \Delta$ is usually satisfied.

- *Plasma number density*: For a conical jet with total "wind" luminosity L_w, Lorentz factor Γ, and magnetization parameter σ, the comoving-frame ejecta proton number density is

$$n'_p = \frac{L_w}{4\pi(1+\sigma)R_{\rm GRB}^2 \Gamma^2 (m_p + \mathcal{Y} m_e)c^3}$$
$$\simeq (1.8 \times 10^7 \text{ cm}^{-3}) L_{w,52} \Gamma_{2.5}^{-2} R_{\rm GRB,15}^{-2} \hat{m}^{-1}(1+\sigma)_1^{-1} , \tag{9.89}$$

where $\mathcal{Y} \geq 1$ denotes the pair multiplicity (baryon-associated electrons included), and $\hat{m} = 1 + \mathcal{Y} m_e/m_p$ is the normalized mass, which is ~ 1 if the $\mathcal{Y} \ll m_p/m_e$. In the lab frame, the ejecta proton number density is

$$n_p = \Gamma n'_p \simeq (5.6 \times 10^9 \text{ cm}^{-3}) L_{w,52} \Gamma_{2.5}^{-1} R_{\rm GRB,15}^{-2} \hat{m}^{-1}(1+\sigma)_1^{-1} . \tag{9.90}$$

The lepton number densities in the ejecta are

$$n'_e = \mathcal{Y} n'_p, \quad n_e = \mathcal{Y} n_p, \tag{9.91}$$

in the comoving and lab frames, respectively.

- *Magnetic field strength*: The magnetic field strength in the emission region is

$$B' = \left(\frac{L_w}{\Gamma^2 R_{\rm GRB}^2 c} \frac{\sigma}{1+\sigma} \right)^{1/2}$$
$$\simeq (1.8 \times 10^3 \text{ G}) \left(\frac{\sigma}{1+\sigma} \right)^{1/2} L_{w,52}^{1/2} \Gamma_{\rm GRB,2.5}^{-1} R_{\rm GRB,15}^{-1} \tag{9.92}$$

in the comoving frame, and is

$$B = \Gamma B' \simeq (5.8 \times 10^5 \text{ G}) \left(\frac{\sigma}{1+\sigma} \right)^{1/2} L_{w,52}^{1/2} R_{\text{GRB},15}^{-1} \tag{9.93}$$

in the lab frame. In the latter case, this **B** field is accompanied by an induced $\mathbf{E} = -\mathbf{V} \times \mathbf{B}$ field for an ideal MHD fluid.

- *Collisional mean free path and time scale*: The Coulomb collision radius may be defined by $e^2/r_{\text{col}} \sim kT$ so that $r_{\text{col}} \sim e^2/kT \sim (1.7 \times 10^{-3} \text{ cm})/T$, where kT generally denotes the average energy of the particles. The comoving-frame collision mean free path of electrons can be estimated as

$$l'_{e,\text{col}} = (n'_e \pi r_{\text{col}}^2)^{-1} \simeq (6.5 \times 10^{17} \text{ cm})$$
$$\times L_{w,52}^{-1} \Gamma_{2.5}^2 R_{\text{GRB},15}^2 \hat{m} Y^{-1} (1+\sigma)_1 T_{e,10}^2 . \tag{9.94}$$

For a relativistic flow, the plasma temperature may be at least a relativistic temperature $T_e \sim m_e c^2/k = 5.9 \times 10^9$ K. Equation (9.94) shows that GRB jets are "collisionless" in the emission region. Since the comoving electron speed $v'_e \sim c$, the comoving-frame collisional time can be estimated as

$$\tau'_{\text{col, } R} = \frac{l'_{e,\text{col}}}{c} \simeq (2.2 \times 10^7 \text{ s})$$
$$\times L_{w,52}^{-1} \Gamma_{2.5}^2 R_{\text{GRB},15}^2 \hat{m} Y^{-1} (1+\sigma)_1 T_{e,10}^2 , \tag{9.95}$$

which is $\gg$ the dynamical time scale $t'_{\text{dyn}} = R_{\text{GRB}}/\Gamma c \sim (110 \text{ s}) R_{\text{GRB},15} \Gamma_{2.5}^{-1}$. This again suggests the collisionless nature of the ejecta.

- *Gyroradii and gyrofrequencies*: Without Coulomb collisions, the GRB plasma is communicated through magnetic fields microscopically, so that the ejecta can still be approximately described as a "fluid" macroscopically. The comoving-frame gyro(cyclotron)radii are

$$r'_{B,e} = \frac{\gamma_e m_e c^2}{eB'} \simeq 0.93 \text{ cm } \gamma_e L_{w,52}^{-1/2} \Gamma_{2.5} R_{\text{GRB},15} \left(\frac{1+\sigma}{\sigma} \right)^{1/2} \tag{9.96}$$

for electrons (where γ_e is the electron Lorentz factor), and

$$r'_{B,p} = \frac{\gamma_p m_p c^2}{eB'} \simeq (1.7 \times 10^3 \text{ cm}) \gamma_p L_{w,52}^{-1/2} \Gamma_{2.5} R_{\text{GRB},15} \left(\frac{1+\sigma}{\sigma} \right)^{1/2} \tag{9.97}$$

for protons (where γ_p is the proton Lorentz factor). For typical values of γ_e and γ_p, both radii are $\ll \Delta'$. The corresponding comoving gyrofrequencies are

$$\omega'_{B,e} = \frac{eB'}{m_e c} \simeq (3.2 \times 10^{10} \text{ s}^{-1}) L_{w,52}^{1/2} \Gamma_{2.5}^{-1} R_{\text{GRB},15}^{-1} \left(\frac{\sigma}{1+\sigma} \right)^{1/2} \tag{9.98}$$

for electrons, and

$$\omega'_{B,p} = \frac{eB'}{m_p c} = \omega'_{B,e} \frac{m_e}{m_p}$$
$$\simeq (1.7 \times 10^7 \text{ s}^{-1}) L_{w,52}^{1/2} \Gamma_{2.5}^{-1} R_{\text{GRB},15}^{-1} \left(\frac{\sigma}{1+\sigma} \right)^{1/2} \tag{9.99}$$

for protons. Both are much larger than the inverse of the comoving dynamical time, i.e. $(R_{\mathrm{GRB}}/\Gamma c)^{-1} \sim (9.5 \times 10^{-3}\ \mathrm{s}^{-1})\ \Gamma_{2.5}R_{\mathrm{GRB},15}^{-1}$. This justifies the fluid description of the GRB ejecta.

- *Plasma frequencies and plasma skin depths*: The comoving-frame relativistic plasma frequencies are

$$\omega'_{p,e} = \left(\frac{4\pi n'_e e^2}{\bar{\gamma}_e m_e}\right)^{1/2} \simeq (2.4 \times 10^8\ \mathrm{s}^{-1})$$
$$\times\ \bar{\gamma}_e^{-1/2}Y^{1/2}L_{w,52}^{1/2}\Gamma_{2.5}^{-1}R_{15}^{-1}\hat{m}^{-1/2}(1+\sigma)_1^{-1/2} \quad (9.100)$$

for electrons, and

$$\omega'_{p,p} = \left(\frac{4\pi n'_p e^2}{\bar{\gamma}_p m_p}\right)^{1/2} \simeq (5.5 \times 10^6\ \mathrm{s}^{-1})$$
$$\times\ \bar{\gamma}_p^{-1/2}L_{w,52}^{1/2}\Gamma_{2.5}^{-1}R_{15}^{-1}\hat{m}^{-1/2}(1+\sigma)_1^{-1/2} \quad (9.101)$$

for protons, where $\bar{\gamma}_e$ and $\bar{\gamma}_p$ denote the mean Lorentz factors of the relativistic electrons and protons, respectively. The corresponding plasma skin depths are

$$\delta'_e = \frac{c}{\omega'_{p,e}} \simeq 130\ \mathrm{cm}\ \bar{\gamma}_e^{1/2}Y^{-1/2}L_{w,52}^{-1/2}\Gamma_{2.5}R_{15}\hat{m}^{1/2}(1+\sigma)_1^{1/2} \quad (9.102)$$

for electrons, and

$$\delta'_p = \frac{c}{\omega'_{p,p}} \simeq (5.4 \times 10^3\ \mathrm{cm})\ \bar{\gamma}_p^{1/2}L_{w,52}^{-1/2}\Gamma_{2.5}R_{15}\hat{m}^{1/2}(1+\sigma)_1^{1/2} \quad (9.103)$$

for protons. For an outflow entrained with an ordered magnetic field, the plasma oscillation frequencies and skin depths are relevant only in the direction parallel to the magnetic field lines, while the gyrofrequencies and gyroradii are more relevant in the direction perpendicular to the magnetic field lines.

- *Reynolds number*: The Reynolds number is defined as the ratio between the viscous diffusion time $\tau_\nu = L^2/\nu$ and the relative flow time scale $\tau_f = L/\delta v$, i.e.

$$R_e \equiv \frac{L\delta v}{\nu}\ . \quad (9.104)$$

Here L and δv are the characteristic length and relative velocity of the flow, and

$$\nu \sim c_s l \quad (9.105)$$

is the kinematic viscosity, where c_s is sound speed, and l is the mean free path of microscopic interactions which defines the viscosity. For a relativistic flow, the relative speed is usually also relativistic, so that $\delta v \sim c$. The sound speed is also close to the speed of light $c_s \sim c/\sqrt{3}$ (Eq. (4.55)). As a result, one may estimate $R_e \sim L/l \sim 10^{28} \gg 1$. This suggests that the GRB outflow is highly turbulent. Such turbulence has been seen in numerical simulations of weakly magnetized fluids within the GRB context (e.g. Zhang et al., 2009c).

- *Magnetic Reynolds number*: The magnetic Reynolds number is defined as the ratio between the magnetic resistive diffusion time $\tau_{\mathrm{dif}} = L^2/\eta$ and the flow time $\tau_f = L/\delta v$, i.e.

$$R_m \equiv \frac{L\delta v}{\eta},$$ (9.106)

where η is the magnetic diffusion coefficient. The maximum resistivity is for the "Bohm" diffusion, i.e.

$$\eta_{\mathrm{B}} \lesssim r_B v \sim r'_{B,e} c,$$ (9.107)

which gives

$$R_{m,B} \simeq \Delta'/r'_{B,e} \simeq 3.4 \times 10^{12} \gamma_e^{-1} L_{w,52}^{1/2} \Gamma_{2.5}^{-2} \left(\frac{\sigma}{1+\sigma} \right)^{1/2} \gg 1.$$ (9.108)

This suggests that magnetic field lines can be highly distorted and turbulent if the turbulence-triggering condition is satisfied. The triggering condition, however, is increasingly stringent with increasingly higher σ of the fluid.

9.7.2 Magnetic Reconnection Physics

Since magnetic reconnection is at the core of most magnetic dissipation models, in the following we briefly introduce the general physics of magnetic reconnection.

Sweet–Parker Reconnection Theory

The standard non-relativistic reconnection theory is the Sweet–Parker theory (Sweet, 1958; Parker, 1957). According to this theory, two sets of field lines with opposite orientations approach each other and reconnect within a layer of thickness δ and length $\mathcal{L}$ (top left panel of Fig. 9.11), with the relationship

$$\frac{\delta}{\mathcal{L}} = \frac{v_{\mathrm{in}}}{v_{\mathrm{A}}} = S^{-1/2}$$ (9.109)

satisfied, where

$$S \equiv \frac{\mathcal{L} v_{\mathrm{A}}}{\eta}$$ (9.110)

is the Lundquist number, η is the magnetic diffusion coefficient, v_{in} is the inflow speed, and v_{A} is the Alfvén speed (Eq. (4.108)):

$$v_{\mathrm{A}} = \frac{c v_{\mathrm{A,NR}}}{(c^2 + v_{\mathrm{A,NR}}^2)^{1/2}},$$ (9.111)

and

$$v_{\mathrm{A,NR}} = \frac{B}{\sqrt{4\pi\rho}} = \sqrt{\sigma} c$$ (9.112)

is the Alfvén speed in the non-relativistic regime. For $\sigma \gg 1$, one has $v_{\mathrm{A}} \sim c$, and

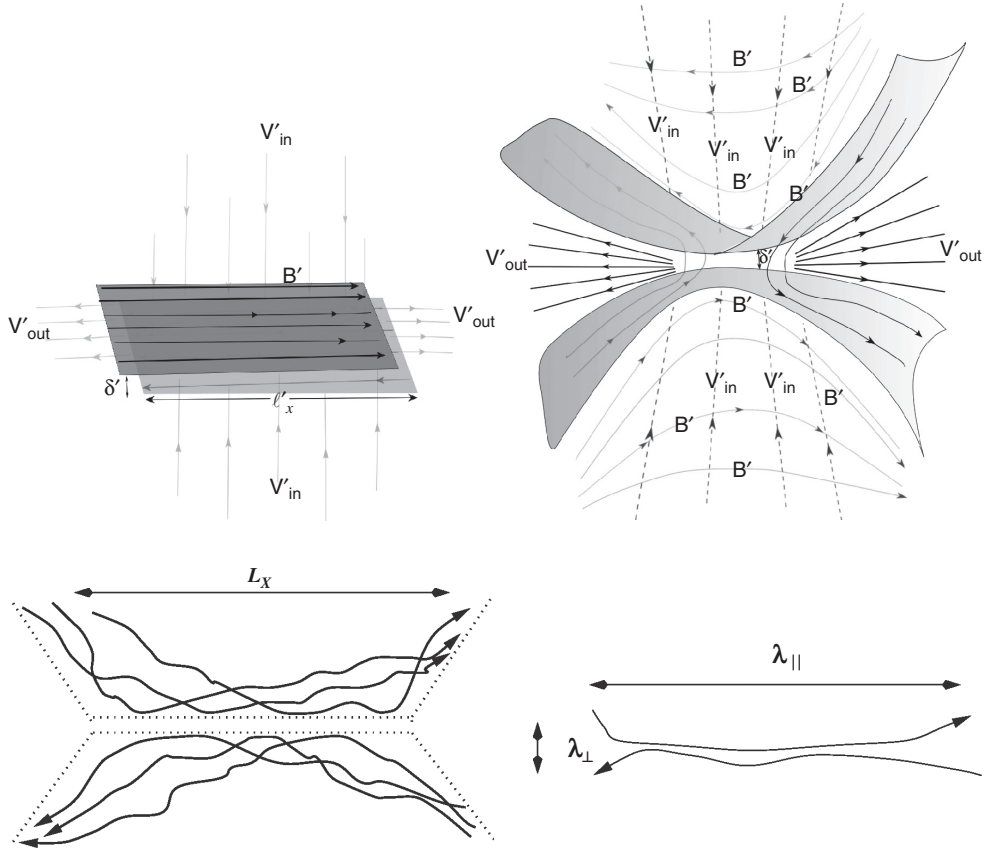$$\gamma_{\mathrm{A}} = (1 + \sigma)^{1/2}.$$ (9.113)

Figure 9.11 Various configurations of magnetic reconnections. *Top left:* The Sweet–Parker geometry. *Top right:* The Petschek geometry. From Kumar and Zhang (2015). *Lower panels:* Turbulent reconnection. Reproduced from Figure 2 in Lazarian and Vishniac (1999) with permission. ©AAS.

In general, $v_{in} = v_A S^{-1/2} \ll v_A$ is satisfied, so that the Sweet–Parker reconnection process is an extremely slow process.

In the high-σ, relativistic regime, the reconnection physics has some novel features. The following discussion closely follows Kumar and Zhang (2015). Other relevant work in the literature includes Blackman and Field (1994), Lyutikov and Uzdensky (2003), and Lyubarsky (2005).

For a relativistic, high-σ Sweet–Parker reconnection configuration, conservations of mass and energy flux give

$$n_1 \mathcal{L} v_{in} = n_2 \delta v_{out} \gamma_{out}, \qquad (9.114)$$

$$(B^2/4\pi)\mathcal{L} v_{in} = n_2 \delta m_p c^2 v_{out} \gamma_{out}^2 \Theta, \qquad (9.115)$$

where n_1 and n_2 are plasma comoving-frame densities outside and inside of the reconnection layer, respectively, v_{in} and v_{out} are the inflow and outflow velocities of the reconnecting fluids, γ_{out} is the outflow Lorentz factor (the inflow speed has to be non-relativistic, say at

most $0.1c$, due to the thermal pressure in the reconnection layer), and Θ is the internal random Lorentz factor of protons in the reconnection layer (current sheet). The ratio between these two relations gives

$$\frac{B^2}{4\pi n_1 m_p c^2} \equiv \sigma \simeq \gamma_A^2 = \gamma_{\text{out}}\Theta. \tag{9.116}$$

So the outflow from the reconnection layer can be relativistic for $\sigma > 1$. This provides the physical basis for the moderately relativistic mini-jets in the bulk jet.

One can prove that the Sweet–Parker reconnection is very inefficient, i.e. $v_{\text{in}} \ll c$. The time scale for magnetic dissipation in the current sheet is

$$t_{B,\text{dis}} \simeq \frac{\delta^2}{\eta}, \tag{9.117}$$

so that

$$v_{\text{in}} = \frac{\delta}{t_{B,\text{dis}}} \sim \frac{\eta}{\delta}. \tag{9.118}$$

Balancing the external magnetic pressure and the thermal pressure in the reconnection layer,

$$\frac{B^2}{8\pi} \sim n_2 \Theta m_p c^2, \tag{9.119}$$

with Eq. (9.116), one gets

$$\frac{n_2}{n_1} \sim \frac{\sigma}{\Theta} \sim \gamma_{\text{out}}. \tag{9.120}$$

With Eqs. (9.114) and (9.118), one finally gets

$$v_{\text{in}} \sim (v_A v_{\text{out}})^{1/2} \gamma_{\text{out}} S^{-1/2}. \tag{9.121}$$

For typical parameters of GRB emission due to forced magnetic dissipation (e.g. Zhang and Yan, 2011), the scale length of the emission region may be estimated as

$$\mathcal{L} \sim r/\Gamma \sim 10^{13} \text{ cm } r_{15}\Gamma_2^{-1}. \tag{9.122}$$

In the Bohm diffusion limit, the magnetic diffusion coefficient can be estimated as

$$\eta \simeq c r'_{B,p}, \tag{9.123}$$

where

$$r'_{B,p} = \frac{\gamma_p m_p c^2}{eB'} \simeq (5.6 \times 10^2 \text{ cm}) \gamma_p L_{w,52}^{-1/2} \Gamma_2 R_{15} \left(\frac{1+\sigma}{\sigma}\right)^{1/2} \tag{9.124}$$

is the comoving proton Larmor gyration radius. Noting $v_A \sim v_{\text{out}} \sim c$ when $\sigma \gg 1$, one finally obtains from Eq. (9.121)

$$v_{\text{in}} \sim c\gamma_{\text{out}} \left(\frac{r_{B,p}}{\mathcal{L}}\right)^{1/2} \sim 10^{-5} c\gamma_{\text{out},1} \ll c. \tag{9.125}$$

Petschek Reconnection

Petschek (1964) proposed a fast steady-state reconnection scenario, which invokes a much shorter width $\mathcal{L}$ of the resistive layer, so that v_{in} is significantly increased. Instead of a planar geometry as invoked in the Sweet–Parker geometry, Petschek invokes an "X"-shaped geometry, with rapid reconnection proceeding at the *X-point* (top right panel of Fig. 9.11). Resistive MHD numerical simulations, on the other hand, indicate that the Petschek reconnection model is unstable, unless η keeps increasing near the X-point (Uzdensky and Kulsrud, 2000). There is no straightforward reason why such a condition can be satisfied.

Turbulent Reconnection

Lazarian and Vishniac (1999) proposed that rapid reconnection may proceed with the presence of turbulence (lower panels of Fig. 9.11). This allows multiple reconnection events to occur simultaneously.

As shown in Eqs. (9.104) and (9.108), both the Reynolds number R_e and the magnetic Reynolds number $R_{m,B}$ are $\gg 1$ in a GRB environment, suggesting that the magnetized GRB outflow can be highly distorted and turbulent.

Many numerical simulations independently showed that reconnection proceeds stochastically via turbulence. Simulations of Sweet–Parker reconnections show that magnetic islands are formed due to the tearing instability, and that multiple reconnection sites quickly appear to speed up reconnection (e.g. Loureiro et al. 2007; Hesse and Zenitani 2007; Samtaney et al. 2009; Guo et al. 2014, 2016). Three-dimensional simulations (e.g. Kowal et al., 2009) showed turbulent reconnection as suggested by Lazarian and Vishniac (1999).

In the presence of turbulence, magnetic reconnections proceed over a local scale length $\lambda_{\parallel}$ rather than the global scale length $\mathcal{L}$. Accordingly, it is the parameter

$$s \equiv \frac{\lambda_{\parallel} v_A}{\eta} \tag{9.126}$$

rather than the Lundquist number S that defines how efficiently reconnection proceeds. Since $\lambda_{\parallel} \ll \mathcal{L}$, one has $s \ll S$. The inflow speed v_{in} is greatly increased, so that the magnetic dissipation rate is significantly increased. This would facilitate significant magnetic energy dissipation to power efficient GRB prompt emission.

9.7.3 Overview of Optically Thin Magnetic Dissipation Models

Several magnetic dissipation models in an optically thin region have been discussed in the literature.

- Usov (1994), within the framework of a millisecond magnetar central engine, proposed that, at a radius of $\sim 10^{13}$ cm, the MHD approximation of the pulsar wind breaks down. Intense electromagnetic waves are generated. Outflowing particles (electron–positron pairs) are accelerated to a Lorentz factor of $\sim 10^6$, and power non-thermal

synchro-Compton radiation. Other variants of such MHD-condition-broken scenarios include Lyutikov and Blackman (2001) and Zhang and Mészáros (2002a).

- Lyutikov and Blandford (2003) proposed an "electromagnetic model" of GRBs by invoking a Poynting-flux-dominated outflow with extremely high σ_0 at the central engine. The dissipation radius is at $\sim 10^{16}$ cm due to current-driven instabilities. This model is discussed below in §9.7.4.
- McKinney and Uzdensky (2012) proposed a reconnection switch model, which conjectures a dissipation radius of $\sim 10^{14}$ cm, at which reconnection is switched from the collisional to collisionless regime. This model is discussed below in §9.7.5.
- Zhang and Yan (2011) proposed an Internal-Collision-induced MAgnetic Reconnection and Turbulence (ICMART) model by invoking collision-triggered magnetic dissipation at a radius of $\sim 10^{15}$ cm. This model will be discussed separately in detail in §9.8.

9.7.4 The Electromagnetic Model

The eletromagnetic model of GRBs was proposed by Lyutikov and Blandford (2003), and re-stated by Lyutikov (2006). The main features of the model include the following:

- The model invokes a Poynting-flux-dominated outflow launched from a rotating, relativistic, stellar progenitor, which loses much of its rotational energy in the form of a Poynting flux. The magnetization parameter σ_0 at the central engine is extremely high. At the deceleration radius ($\sim 3 \times 10^{16}$ cm), the outflow is still "sub-Alfvénic", i.e. $\Gamma < \gamma_A = (1 + \sigma)^{1/2}$, which means

$$\sigma > \sigma_c \equiv \Gamma^2 - 1 \sim 10^5 \Gamma_{2.5}^2 \qquad (9.127)$$

at the deceleration radius.
- The central engine lasts for ~ 100 s. Initially a non-spherically symmetric, non-relativistic, electromagnetically dominated bubble expands inside the star, most rapidly along the spin axis. Later the bubble breaks out from the stellar surface and accelerates itself to an extreme relativistic speed due to a magnetic pressure gradient within the bubble. At the end of central engine activity, most of the electromagnetic energy is concentrated in a thin shell, which expands into the circumburst medium. No intermittent central engine activities are invoked in the model to interpret the observed GRB variability.
- Inside the star some magnetic dissipation occurs, so that thermal photons are released at the photosphere to produce a weak thermal precursor. The bulk of the electromagnetic energy remains undissipated until the outflow reaches the deceleration radius of $\sim 3 \times 10^{16}$ cm, at which current-driven instabilities develop, leading to significant magnetic energy dissipation, particle acceleration, and radiation. Lyutikov and Blandford (2003) argued that the reason that magnetic disspation does not happen at smaller radii is "because the particle acceleration is suppressed near the central engine by efficient pair production which screens out the electric field that led to the particle acceleration".

- The observed variabilities are not interpreted as the intermittent behavior of the central engine. Rather, it is suggested that in the emission region many fundamental emitters (or mini-jets) due to magnetic dissipation (e.g. reconnection) exist. These mini-jets move relativistically in the bulk comoving frame (since the Alfvén Lorentz factor $\gamma_A \gg 1$), so that they are Doppler-boosted locally. Due to the random orientations of these fundamental emitters, rapid variability in the GRB lightcurves is observed. Such an interpretation of GRB variability was investigated later by various authors from different angles (Narayan and Kumar, 2009; Lazar et al., 2009; Zhang and Yan, 2011; Zhang and Zhang, 2014).

This electromagntic model opens a new window for considering the GRB problem in the opposite regime of the fireball model, i.e. the electromagnetic regime. It suffers, however, from some problems both theoretically and observationally.

- From the theoretical point of view, σ is expected to decrease as a function of radius r (§7.4). In order to have $\sigma > 10^5$ at the deceleration radius, one demands an extremely high σ_0 which is likely unachievable at the central engine. It is also hard to avoid various kinds of magnetic dissipation processes during the propagation of such a Poynting-flux-dominated outflow. Such dissipation processes would significantly bring down the σ value of the outflow to be below σ_c defined in Eq. (9.127).
- Observationally, GRB lightcurves suggest that the GRB central engine is intermittent and is active in multiple emission episodes. This is especially illustrated by the data on X-ray flares, which are clearly an extension of the prompt emission in the weaker and softer regime. Interpreting the decay segment of the X-ray flares as due to the curvature effect, Liang et al. (2006b) showed that the clock is restarting every time a new episode of emission is released. This demands multiple episodes of central engine activity to interpret GRB emission. Attributing the GRB variability to mini-jets only is not observationally justified.
- Some GRBs show a dominant or sub-dominant quasi-thermal spectral component, which can be attributed to the emission from the photosphere of a fireball or a hybrid outflow. This is inconsistent with the assumption of a purely Poynting-flux-dominated (extemely high-σ) outflow.

9.7.5 The Reconnection Switch Model

Another magnetic reconnection model was proposed by McKinney and Uzdensky (2012) (see the cartoon picture, Fig. 9.12). Adopting a striped wind magnetic field configuration, McKinney and Uzdensky (2012) identified two regimes of magnetic reconnection separated by a critical radius R_{trans}, which is defined by

$$\delta_{\text{SP}}(R_{\text{trans}}) = d_i(R_{\text{trans}}), \tag{9.128}$$

where $\delta_{\text{SP}} \sim \mathcal{L}S^{-1/2}$ is the Sweet–Parker current sheet thickness, and

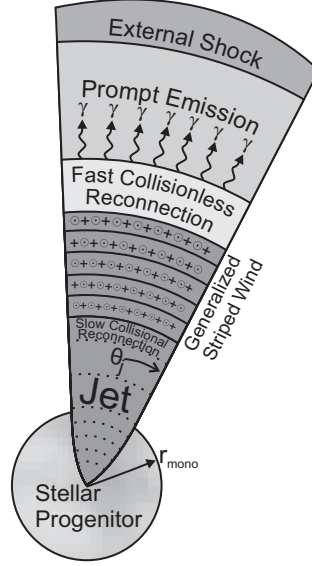$$d_i = \frac{c}{\omega_{pi}} \tag{9.129}$$

Figure 9.12   A cartoon picture of the reconnection switch model of GRBs. From McKinney and Uzdensky (2012).

is the characteristic length of the ion plasma, with ω_{pi} being the ion plasma frequency. At small radii $r < R_{\text{trans}}$, one has $\delta_{\text{SP}} > d_i$ (at small radii, ω_{pi} is large due to the large density of the outflow), so that magnetic reconnections are *collisional* with a Sweet–Parker geometry. The reconnection speed v_{in} is therefore small, so that magnetic dissipation is very inefficient. At $r > R_{\text{trans}}$, the condition $\delta_{\text{SP}} < d_i$ is satisfied (the fluid density drops to a critical value so that d_i is small enough to be below δ_{SP}). The plasma no longer obeys the resistive MHD equations, so that reconnections proceed rapidly in the *collisionless* regime, with a Petschek-like geometry. The reconnection rate increases suddenly, leading to efficient magnetic dissipation. The transition radius is estimated to be

$$R_{\text{trans}} \sim 10^{13}\text{–}10^{14} \text{ cm.} \tag{9.130}$$

Similar to the ICMART model (see §9.8 below in detail), the reconnection switch model presents a mechanism to trigger rapid magnetic dissipation beyond a critical radius. This radius is close to the standard internal shock radius, which is above the photosphere. Even though McKinney and Uzdensky (2012) stated that this mechanism would enhance the photosphere emission, with the standard parameters, the dissipation site is actually usually in the optically thin region. The radiation mechanism therefore should be non-thermal, e.g. via synchrotron radiation. The authors did not perform detailed modeling to compare the model predictions with the data. One would expect that synchrotron radiation would be in the fast cooling regime. The radiative efficiency is likely higher than that in the internal shock model. The lightcurve variability directly tracks the central engine activities (similar to the photospheric models), which is somewhat different from the internal shock model, where the observed luminosity is a convolution of the luminosity and Lorentz factor histories of the central engine.

9.8 Internal-Collision-Induced MAgnetic Reconnection and Turbulence (ICMART)

9.8.1 General Considerations

The Internal-Collision-induced MAgnetic Reconnection and Turbulence (ICMART) model was proposed by Zhang and Yan (2011) in view of the weak photosphere emission detected in some bright *Fermi* GRBs (e.g. GRB 080916C, Abdo et al. 2009c), which suggests that the jet composition of at least some GRBs is not a matter-dominated fireball (Zhang and Pe'er, 2009). On the other hand, the photosphere emission is indeed detected in some GRBs, mostly sub-dominant as compared with the synchrotron component (e.g. Guiriec et al., 2011; Axelsson et al., 2012; Guiriec et al., 2013), suggesting that the jet composition is not extremely Poynting flux dominated. The ICMART model is relevant for GRBs in the intermediate regime, with $\sigma > 1$ but not $\gg 1$ in the emission region.

The key ingredients of the ICMART model include the following:

- The GRB central engine carries a large σ_0 (e.g. >100), and may also carry a moderate fireball parameter η as well (i.e. a hybrid jet).
- The jet remains Poynting flux dominated and undissipated until reaching a large enough distance, e.g. $R_{GRB} \sim 10^{15}$ cm, as suggested by the observations. At R_{GRB}, the magnetization parameter σ is still above unity (Fig. 9.13 upper panel), in contrast to the internal shock model which requires $\sigma < 1$. Since the kinetic energy tapped from internal shocks is smaller by a factor of $(1+\sigma)$ with respect to the already small internal energy available in the traditional internal shock models, internal shock dissipation is not considered as the primary energy dissipation mechanism.
- The central engine is assumed intermittent, ejecting Poynting-flux-dominated "shells" intermittently. Internal collisions among these highly magnetized shells would trigger dissipation of *magnetic energy* in the outflow (rather than kinetic energy) through rapid turbulent reconnections, powering an efficient energy dissipation. Such collisions may happen at $R_{GRB} \sim 10^{15}$ cm.
- For a helical magnetic configuration, repeated collisions may be needed to destroy the ordered magnetic fields and eventually trigger an ICMART event (Zhang and Yan, 2011) (Fig. 9.13 lower panel). Another possibility is that collisions may trigger kink instability in the helical jet leading to magnetic dissipation (Lazarian et al., 2018). For the case of colliding discrete magnetic blobs (Yuan and Zhang, 2012), one collision can trigger rapid dissipation (Deng et al., 2015).
- Within this scenario, the magnetic energy is essentially not dissipated until an ICMART event is triggered. At the photosphere, the outflow is Poynting flux dominated, so that the photosphere emission is suppressed. The model predicts a bright non-thermal emission component and a weak or non-detectable thermal emission component.
- An ICMART event is envisaged to proceed in a runaway manner. Seed rapid reconnections would trigger turbulence, which facilitates more reconnections so that the reconnection regions increase exponentially until most of the magnetic energy is dissipated.
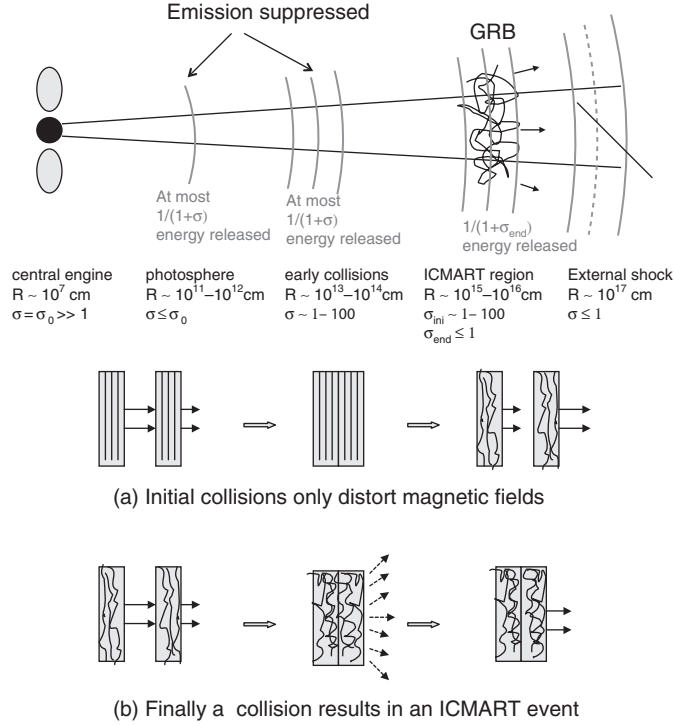
(a) Initial collisions only distort magnetic fields

(b) Finally a collision results in an ICMART event

Figure 9.13 *Upper:* A cartoon picture of the ICMART model showing various distance scales. *Lower:* One possible way to trigger ICMART events. From Zhang and Yan (2011).

Since the undissipated plasma has $\sigma_{ini} > 1$ in the emission region, each reconnection event would make a "mini-jet" with a moderate Lorentz factor $\gamma_{out} \lesssim \sqrt{1 + \sigma_{ini}}$ in the comoving frame of a bulk jet. The observed emission is the superposition of many mini-jets (Zhang and Zhang, 2014).

9.8.2 Model Features

The ICMART model has a list of features that distinguish it from other models (e.g. photosphere and internal shock models).

Efficiency

Suppose that before collision the shells have a magnetization parameter $\sigma_{ini} > 1$. After the ICMART event, the magnetization parameter is brought to $\sigma_{end} \sim 1$. Similar to the internal shock model, one can write down the energy conservation and momentum conservation laws (Zhang and Yan, 2011):

$$(\Gamma_2 m_2 + \Gamma_1 m_1)(1 + \sigma_{ini}) = \Gamma_m (m_1 + m_2 + \hat{\gamma} U)(1 + \sigma_{end}), \qquad (9.131)$$

$$(\Gamma_2 \beta_2 m_2 + \Gamma_1 \beta_1 m_1)(1 + \sigma_{ini}) = \Gamma_m \beta_m (m_1 + m_2 + \hat{\gamma} U)(1 + \sigma_{end}). \qquad (9.132)$$

This gives the same solution for Γ_m (Eq. (9.71)) as the internal shock model, i.e.

$$\Gamma_m = \left(\frac{\Gamma_1 m_1 + \Gamma_2 m_2}{m_1/\Gamma_1 + m_2/\Gamma_2} \right)^{1/2} . \tag{9.133}$$

The energy dissipation efficiency, on the other hand, is much larger, i.e. (Exercise 9.2)

$$\begin{aligned}
\eta_{\text{ICMART}} &= \frac{\Gamma_m \hat{\gamma} U}{(\Gamma_1 m_1 c^2 + \Gamma_2 m_2 c^2)(1 + \sigma_{\text{ini}})} \\
&= \frac{1}{1 + \sigma_{\text{end}}} - \frac{\Gamma_m(m_1 + m_2)}{(\Gamma_1 m_1 + \Gamma_2 m_2)(1 + \sigma_{\text{ini}})} \\
&\simeq \frac{1}{1 + \sigma_{\text{end}}} \quad (\text{if } \sigma_{\text{ini}} \gg 1).
\end{aligned} \tag{9.134}$$

This gives $\sim 50\%$ if $\sigma_{\text{end}} \sim 1$ and $\sigma_{\text{ini}} \gg 1$. This may account for the observed high radiative efficiency of some GRBs (e.g. Zhang et al., 2007a).

Numerical simulations of the collisions of two high-σ blobs indeed reveal significant magnetic dissipation triggered by collisions (Figs. 9.14 and 9.15). The typical dissipation efficiency is $\sim 35\%$. Evidence of collision-induced reconnections is directly observed in the simulations (Deng et al., 2015).

Two-Component Variability

Observationally, at least some GRB lightcurves show the superposition of fast and slow variability components (e.g. Vetere et al., 2006; Gao et al., 2012). The ICMART model may account for the existence of two emission components: the fast variability component is related to mini-jets due to multiple reconnection sites in the emission region due to turbulent reconnections, whereas the slow variability component is related to central engine activity. Within this hypothesis, a GRB is composed of multiple ICMART events, each representing one broad "pulse" in the lightcurve. Monte Carlo simulations of ICMART lightcurves can reproduce the basic features of the observed GRB lightcurves and power density spectra (Zhang and Zhang 2014, see the lower right panel of Fig. 9.2). Some GRBs
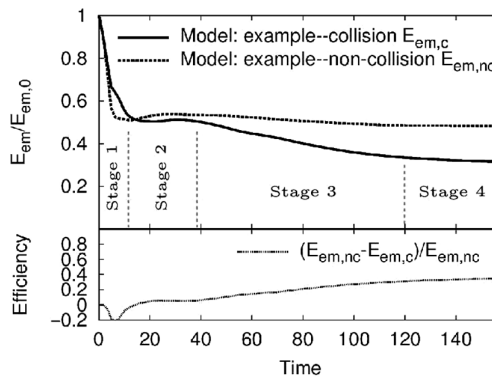

Figure 9.14 Numerical simulation results of ICMART events. The magnetic energy evolution of one magnetic blob (dashed line marked "non-collision") and two colliding magnetic blobs (solid line marked "collision"). Significant magnetic dissipation is observed. From Deng et al. (2015).

Panel A: stage 1, cut at t=4, zoom in

Panel B: stage 2, cut at t=18, zoom in

Panel C: stage 3, cut at t=58, zoom in

Panel D: stage 4, cut at t=94, zoom in

0   0.5   1.0   -0.75   0   0.75   0   0.15   0.3
Color bar of current   Color bar of V_x/V_y Color bar of density
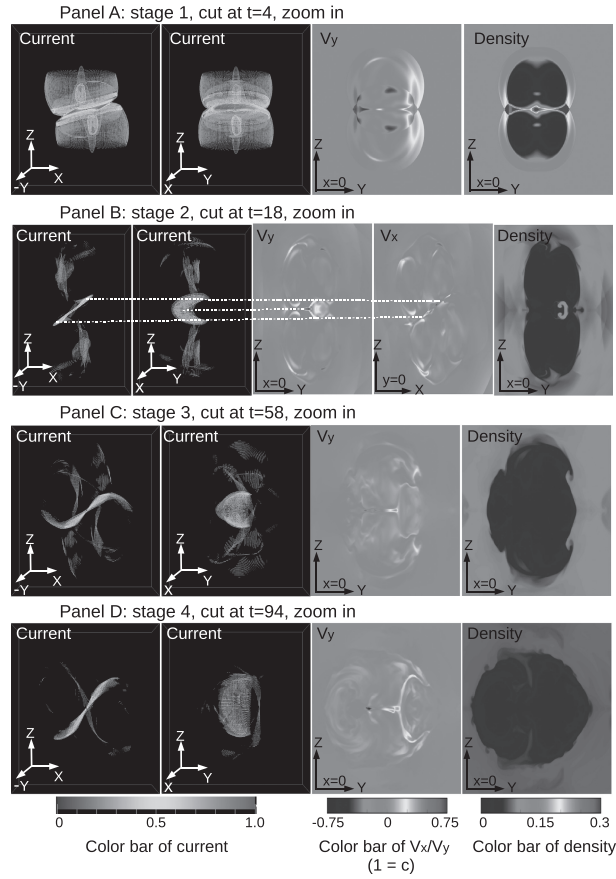   (1 = c)

Figure 9.15 Numerical simulation results of ICMART events. Representative cuts of current, velocity, and density of four different stages (as defined in the upper panel) during one ICMART event, showing collision-triggered reconnection of high-σ blobs. From Deng et al. (2015). A black and white version of this figure will appear in some formats. For the color version, please refer to the plate section.

have erratic lightcurves without clear pulses identified. Since each ICMART event lasts for a time scale of $R_{\mathrm{GRB}}/\Gamma^2 c \sim$ seconds, these erratic lightcurves may be understood by invoking multiple ICMART events triggered with slight time delays at slightly different emission regions, whose emission reaches the observer in an overlapping manner.

Electron Number Problem Alleviated

Recall that the internal shock model predicts too small an E_p and requires a small fraction of electrons being accelerated. The ICMART model may overcome this problem more naturally because the number of electrons in the emission region is a factor $(1+\sigma_{\mathrm{ini}}(R_{\mathrm{GRB}}))$ smaller than in the internal shock model. There is no need to assume that only a small fraction of electrons are accelerated. Each electron receives a larger Lorentz factor, and E_p may be estimated as (Zhang and Yan, 2011) (Exercise 9.3)

$$E_{p,\text{ICMART}} \simeq 160 \text{ keV } L_{\gamma,52}^{1/2} R_{\text{ICMART},15}^{-1} (\eta \epsilon_e)^{3/2} \sigma_{1.5}^2 \left(\frac{1+z}{2} \right)^{-1}, \qquad (9.135)$$

which is more consistent with the observations than the internal shock model (without assuming ξ < 1, Eq. (9.80)). One then avoids the problem of invoking an unobserved thermal electron population and the inconsistency in calculating the synchrotron self-absorption frequency faced by the internal shock model (Shen and Zhang, 2009).

Fast Cooling Problem Alleviated

Within the ICMART scenario, there are two ways to alleviate the fast synchrotron cooling problem faced by the internal shock model. First, the emission site is at a larger emission radius R_{GRB} where the magnetic field strength is lower. Fast cooling in a decaying magnetic field would make the electron spectrum below the injection frequency close to $p = 1$, which would make a low-energy photon index $\alpha \sim -1$ (Uhm and Zhang, 2014b; Geng et al., 2018). One issue is that this scenario requires that the σ value in the emission region is lower than unity, in apparent inconsistency with the hypothesis of moderate σ_{ini} in the outflow. One possibility would be that σ_{end} becomes smaller than unity in the current sheet, and electrons radiate from such a low-B region.

The second way to alleviate the fast cooling problem is to introduce a turbulence-induced, second-order Fermi, slow heating mechanism in the inter-reconnection regions (Zhang and Yan, 2011). Xu and Zhang (2017) considered adiabatic non-resonant acceleration in magnetic turbulence and derived a low-energy photon index matching the typical value $\alpha \sim -1$. Balancing acceleration and synchrotron cooling, electrons have a characteristic Lorentz factor that gives E_p in hundreds of keV through synchrotron radiation. If one also introduces a first-order Fermi acceleration mechanism and synchrotron cooling, a range of broken power-law spectra of GRB prompt emission are predicted (Xu et al., 2018), which are generally consistent with the observations.

Suppressed Photosphere

The ICMART model envisages that the photosphere emission is suppressed. Following the top-down approach developed by Gao and Zhang (2015) to diagnose central engine properties of GRBs with observed weak photosphere emission, one often gets a relatively large σ_0, and a $\sigma(R_{15}) > 1$ at $R \sim 10^{15}$ cm, which requires ICMART to dissipate magnetic energy and power GRB prompt emission.

Amati/Yonetoku Relations

Equation (9.135) has an apparent $E_p \propto L^{1/2}$ (Amati/Yonetoku) correlation. Other parameters, i.e. both the emission radius R_{ICMART} and σ, are involved in the problem. The effects of those two parameters may cancel each other out. A higher σ flow may give rise to a higher bulk Lorentz factor and, hence, a larger R_{ICMART}, so that the factor $R_{\text{ICMART}}^{-1}\sigma^2$ may cancel out the effects to allow a broad Amati/Yonetoku correlation to be satisfied. This is different from the internal shock model where $R_{\text{IS}} \propto \Gamma^2$ is expected (§9.6.5).

9.8.3 Model Predictions

The ICMART model has several interesting predictions, which may be used to verify or disprove this model.

Dynamical Acceleration During the Prompt Emission Phase

During an ICMART event significant magnetic energy is dissipated to power non-thermal emission, and part of the dissipated energy is converted to kinetic energy. So a prediction of the ICMART model is that the emission site is undergoing *bulk acceleration* during the prompt emission phase.

Two independent pieces of observational evidence seem to suggest that the GRB emission region undergoes bulk acceleration.

The first is related to the decay slope of the lightcurves. If the falling tail of an emission episode is controlled by the curvature effect, the decay slope α is steeper than the nominal value $2 + \beta$ (convention $F_\nu \propto t^{-\alpha} \nu^{-\beta}$) if the emission region undergoes bulk acceleration, see Fig. 3.5 (Uhm and Zhang, 2015). A study (Uhm and Zhang, 2016a) of the decay phase of several X-ray flares indicates that the decay slope is steeper than $2 + \beta$, even if the T_0 effect (properly choosing the zero time to plot the lightcurves in a logarithmic scale, Zhang et al. 2006) is corrected in the most conservative manner. Detailed modeling of the X-ray flare lightcurve and hardness evolution based on an accelerating emission region can well explain the data (Uhm and Zhang, 2016a). A systematic analysis of bright *Swift* X-ray flares (Jia et al., 2016) suggests that such an acceleration feature is ubiquitous. An alternative way to steepen the decay slope of an X-ray flare is to introduce comoving anisotropic emitters (Beloborodov et al., 2011; Beniamini and Granot, 2016; Barniol Duran et al., 2016). However, with the anisotropic effect only, it is difficult to reproduce both the lightcurve and hardness-ratio evolution curve (Geng et al., 2017). In reality, both effects (bulk acceleration and anisotropy) may play a role in defining the decay phase of X-ray flares. In any case, both features demand dissipating magnetic fields in a Poynting-flux-dominated flow, which is consistent with the prediction of the ICMART model.

Due to the erratic and overlapping features of the GRB prompt emission lightcurves, it is not easy to test the bulk acceleration scheme with the prompt emission data. On the other hand, evidence of bulk acceleration is independently collected through modeling the spectral-lag behavior of GRB prompt emission pulses. Uhm and Zhang (2016b) showed that the curvature effect alone cannot interpret the spectral lags, since the emission is always dominated by the on-axis emitter unless the spectrum is unrealistically narrow. For reasonable emission mechanisms (e.g. synchrotron or Comptonization), spectral lags demand systematic sweeping of E_p across an emission band during the course of a broad pulse. This suggests that a broad pulse is one radiation unit, with emission released when one single fluid unit streams in space. By modeling E_p evolution and the spectral lags of such an emitter, Uhm and Zhang (2016b) showed that the emitter needs to undergo bulk acceleration in order to reproduce the correct pulse shape and the energy-dependent pulse width and lags.

E_p Evolution Patterns

Along the same lines, since in the ICMART model a broad pulse is a consequence of one emission region streaming out from the central engine, the spectral peak E_p evolution is expected to be associated with the evolution of one broad pulse. In particular, since B decreases with radius (and hence, time) as the emitting region expands in space, a straight-forward prediction is hard-to-soft evolution (Uhm and Zhang, 2014b), as is commonly observed in GRB broad pulses, especially for the first pulse (Lu et al., 2012; Hakkila et al., 2015).[10] Under special conditions (e.g. the typical electron Lorentz factor γ_e also evolving rapidly with time), this pattern may be reversed, to allow an E_p tracking behavior.

Evolution of Polarization Properties

Since synchrotron radiation in ordered magnetic fields is invoked in the ICMART model, a relatively high average polarization degree (Π of tens of percent) is expected. Since an ICMART event is a process that destroys ordered magnetic fields, it is expected that the linear polarization degree Π of GRB emission reduces with time during each pulse (Zhang and Yan, 2011). For GRBs that have multiple ICMART events, the Π curve may go up and down multiple times. A clear evolution pattern is expected for bright GRBs with clearly separated emission episodes. Since different ICMART events may have different magnetic field configurations, the polarization angle may evolve in the same burst within different pulses. Considering the oscillation of the merged emitter after two magnetic blobs collide, one would also expect a switch of the polarization angle by $90^{\rm o}$ (Deng et al., 2016), as was observed in GRB 100826A (Yonetoku et al., 2011).

9.8.4 Summary

The ICMART model invokes a moderate σ in the emission region, which is designated to interpret GRB emission in the intermediate regime of jet composition between a matter-dominated fireball and a pure Poynting-flux-dominated flow. It cannot interpret thermally dominated GRBs such as GRB 090902B, which suggests a photosphere origin of emission. On the other hand, for the GRBs that show non-thermal emission, especially those with a weak or suppressed thermal emission, this model overcomes some difficulties faced by the internal shock model and is better positioned to satisfy observational constraints.

Since the ICMART model invokes the complicated physics of dissipative MHD in the high-σ, relativistic regime, many qualitiative speculations raised in the model (Zhang and Yan, 2011) demand proof from numerical simulations. Progress has been made in numerical simulations to validate some tentative suggestions in the original ICMART model. This includes the relatively high ($\sim$35%) energy dissipation efficiency, the existence of mini-jets (Deng et al., 2015), and high polarization degree, its temporal evolution, and change of polarization angle (Deng et al., 2016). However, more detailed numerical simulations are needed to investigate particle acceleration and radiation in a self-consistent manner.

---

[10] Later pulses may be subject to the overlapping effect, i.e. at any observational epoch, the received emission may come from both the tail of an earlier pulse and the beginning of a new pulse.

9.9 Other Prompt Emission Models

Besides the above three general categories of models (photosphere, internal shocks, and magnetic reconnection in an optically thin region), there are also several other suggestions for interpreting prompt emission in the literature. We mainly outline the basic theoretical picture of these models in the following, and offer some critical comments.

9.9.1 Synchrotron Self-Compton

The SSC model of GRB prompt emission was proposed to explain the "naked-eye" GRB 080319B, which showed a rough tracking behavior between the optical and γ-ray emission, but with an optical flux much higher than the spectral extrapolation of γ-ray emission to the softer energy band (e.g. Kumar and Panaitescu, 2008; Racusin et al., 2008).

However, various arguments are now against the SSC mechanism as the dominant mechanism for GRB prompt emission: (1) The same model would predict a bright GeV component due to second-order SSC. This would greatly increase the total energy budget of GRBs (Derishev et al., 2001; Piran et al., 2009). *Fermi* was not launched at the time of detection of the naked-eye GRB so that one could not test the existence of the second-order SSC. Another bright GRB (even though the prompt optical was not as bright as the naked-eye one), GRB 160625B, was detected to have a prompt optical counterpart with flux in excess of the extrapolation of γ-rays to the softer band (Zhang et al., 2018b). The GeV flux, however, was well consistent with the extrapolation of the sub-MeV spectrum to higher energies, probably with an exponential cutoff in some energy bins. This directly rules out the SSC origin of the sub-MeV component at least for this burst. (2) For both GRB 080319B and GRB 160625B, the optical emission is delayed with respect to the γ-ray emission by a few seconds (Beskin et al., 2010; Zhang et al., 2018b), suggesting that the optical and γ-ray emissions are likely from two different emission zones. (3) Since $E_p \propto \gamma_e^4$ (γ_e^2 from synchrotron, and another γ_e^2 from the SSC) for the SSC models, a slight change of γ_e would introduce a large swing of E_p, leading to a very wide distribution of E_p among GRBs, which is inconsistent with the observations (Zhang and Mészáros, 2002a). (4) Simulations suggested that the SSC lightcurves cannot be much spikier than the synchrotron lightcurves, in contrast with the observational data of GRB 080319B (Resmi and Zhang, 2012).

9.9.2 Compton Drag

Several authors have suggested that GRBs can be generated through bulk scattering of the background photons by a relativistic jet. In the cannonball model (see §7.7.1 for more discussion), GRB emission is interpreted as bulk Compton scattering of the electrons in the jet off seed photons from the so-called "glory" (or echo) of the progenitor star or associated supernova (e.g. Shaviv and Dar, 1995; Dar and de Rújula, 2004). Lazzati et al. (2000) proposed that GRB emission is powered by cold electrons in a relativistic jet upscattering seed photons from the GRB progenitor star, or from the cocoon produced by the passage of the jet through the star. Broderick (2005) investigated a similar IC scenario within the

framework of a specific progenitor system, i.e. a helium star–neutron star binary. Titarchuk et al. (2012) introduced a two-step Comptonization model and claimed that the model can reproduce a typical Band-function spectrum.

More detailed calculations are needed to calculate dynamical evolution of the jet and the resulting emission spectra in the Compton drag model. Since short GRBs are envisaged to be produced from compact-star-merger systems, from which no bright seed photon sources are expected, they cannot be natually interpreted with the Compton drag model.

9.9.3 Hadronic Models

The contributions of hadronic interaction processes within the internal shock model have been discussed by several groups (Gupta and Zhang, 2007b; Asano et al., 2009; Asano and Mészáros, 2012; Murase et al., 2012). These models usually invoke $p\gamma$ interactions to produce pions. Neutron pions would directly decay to γ-rays, while charged pions would decay to muons and eventually $e^{\pm}$, which would generate radiation via synchrotron radiation. Two-photon pair production usually occurs, so that such a model usually invokes detailed pair–photon cascades.

Kazanas et al. (2002) proposed a "super-critical pile" model for GRBs. The general idea is that as a relativistic jet propagates in a circumburst medium, the Bethe–Heitler process ($p\gamma \rightarrow pe^{+}e^{-}$) may reach a resonance condition, namely, the typical synchrotron radiation energy of the pairs is such that it ensures the Bethe–Heitler kinetic condition, and the column density of the photons also satisfies the condition of runaway production of the pairs. This model invokes an external site to discharge the kinetic energy of the jet, so it is similar to the external shock model and suffers from the same difficulty of the external shock model in accounting for the observed GRB variability, and is at odds with the growing evidence that GRB variabilities are connected to intermittent central engine activities.

Petropoulou et al. (2014) considered another hadronic supercriticality of GRB emission. They considered hadronic interactions only. They studied the interaction of the γ-rays produced by proton synchrotron radiation with the protons/neutrons in the ejecta and the subsequent radiation of the produced electron–positron pairs, pions, kaons, and muons. They identified a feedback loop and a critical condition to separate the hadronic interactions into the sub-critical and super-critical regimes. A key parameter is the compactness parameter of the proton-synchrotron-generated γ-rays. If it exceeds a critical value, a positive feedback is triggered and all the hadronic processes are enhanced, giving rise to significant GRB emission. In the opposite regime, when the seed photon compactness is below a critical value, the entire reaction chain is quenched, and hadronic emission becomes inefficient.

There are several issues in general for the hadronic models. First, since hadronic emission processes are less efficient than leptonic processes, leptonic emission usually outshines hadronic emission, unless the electrons carry a very small fraction of the internal energy in the emission region (e.g. $\epsilon_e \ll 1$) as compared to protons (e.g. $\epsilon_p \sim 1$) (Gupta and Zhang, 2007b). Next, this in general demands a very high energy budget in the GRB ejecta. Afterglow modeling, on the other hand, suggests that the kinetic energy in the blast-wave is usually not much larger than the γ-ray energy emitted during the prompt emission

phase, i.e. the prompt emission is efficient (e.g. Panaitescu and Kumar, 2001, 2002; Zhang et al., 2007a). This suggests that the leptonic component is likely the dominant emission component for GRB prompt emission. Finally, invoking a large proton energy fraction ϵ_p usually suggests significant neutrino emission. The non-detection of any neutrino spatially and temporally associated with any GRB (e.g. Abbasi et al., 2012; Aartsen et al., 2015, 2016) places progressively more stringent constraints on the parameter space allowed for hadronic models (see Chapter 12 for more discussion).

9.9.4 Effect of Neutron Decay

If the GRB outflow is a neutron-rich fireball, free neutrons will decouple from protons without participating in internal shock dissipation. They will however decay to produce protons and electrons at a characteristic radius $R_\beta \simeq (8 \times 10^{15}$ cm$)(\Gamma_n/300)$ (Eq. (7.107)). Besides modifying the afterglow behavior as discussed in §8.8.1, they will also leave interesting signatures during the prompt emission phase. In particular, neutron shells may follow a similar luminosity/Lorentz factor history as proton shells, so that after they decay the interactions among these decayed neutron shells may drive internal shocks and produce optical emission through synchrotron radiation (due to weaker B fields at larger radii), which roughly tracks the γ-ray emission, delayed by a time $R_\beta/\Gamma^2 c \sim 3$ s $(\Gamma_n/300)^{-1}$.

The "naked-eye" GRB 080319B has bright prompt optical emission roughly tracking the γ-ray emission (Racusin et al., 2008). The physical origin of the optical emission is still unknown. One possibility is that the source is a neutron-rich fireball. While the proton-shell internal shocks produce the γ-ray emission, the collisions of proton shells with the neutron-decay trails may have produced the optical emission (Fan et al., 2009). The same model may apply to other GRBs with "lagged-tracking" optical emission with respect to the γ-rays, e.g. GRB 160625B (Zhang et al., 2018b).

9.10 Theory Confronting Observations

Unlike the afterglow theory, which is relatively simple and well tested by the data, the GRB prompt emission theory is much more complicated. Due to the diverse observational data and the uncertainties inherent in the models (e.g. jet composition, energy dissipation mechanism, and particle acceleration and radiation mechanisms), there is no single model that can account for all the prompt emission observations.

In principle, one could summarize a list of observational properties, and make a grading chart of all the prompt emission models with each criterion. This is, however, not practical due to the following reasons: (1) most models have not been studied in enough detail to make predictions regarding each observational criterion; (2) some models have been identified to contain serious flaws that are disfavored by the data of most GRBs; (3) many models have overlapping theoretical ingredients and predicted properties, so that they may be grouped together as the same type of model. In the following, we focus on three well-motivated and studied models that have distinct predicted properties as representatives of

the models: the dissipative photosphere model, the internal shock model, and the ICMART model. The last model may be regarded as representative of a broader category of GRB models that invoke magnetic dissipation of a moderately high σ at a large emission radius from the central engine.

Another complication is that each of the models discussed in this chapter invokes one mechanism at one emission site. It is very likely that multiple mechanisms operate in multiple emission sites for different GRBs or even in one GRB. For example, for the jet composition relevant to GRBs (fireballs or hybrid jets), there are at least two emission sites for GRB prompt emission: the photosphere and a non-thermal site in the optically thin region (e.g. the internal shocks or the ICMART regions). Adding neutrons to the jet would introduce another emission site where decayed neutron shells dissipate.

Table 9.1 lists three representative models. The basic ingredients of the three models are summarized below:

- *Dissipative photosphere model:* Smallest emission radius ($R_{\mathrm{GRB}} = R_{\mathrm{ph}} \sim 10^{10}$–$10^{12}$ cm), optically thick, quasi-thermal, Comptonized spectrum, with synchrotron contribution, low σ;
- *Internal shock (IS) model:* Intermediate emission radius ($R_{\mathrm{GRB}} = R_{\mathrm{IS}} \sim 10^{13}$–$10^{14}$ cm), optically thin, low σ, synchrotron radiation, likely randomized magnetic fields;
- *ICMART model:* Large emission radius ($R_{\mathrm{GRB}} = R_{\mathrm{ICMART}} \sim 10^{15}$–$10^{16}$ cm), optically thin, moderately high σ before dissipation, synchrotron radiation, ordered magnetic fields.

In the following, we discuss how these three models confront various observational criteria. The grade "Yes" or "No" indicates that a particular model satisfies or does not satisfy that particular criterion. The grade "Yes(?)" indicates that the model in general can satisfy the observational constraint, but there may be issues or far-stretching of the parameters. The grade "No(?)" indicates that the model in general cannot satisfy the observational constraint, but may be modified by introducing special conditions.

1. *Lightcurve (I): slow variability component:* GRB lightcurves have broad pulses, usually lasting for an order $\Delta t_{\mathrm{slow}} \sim$ seconds. Using the standard formula, $R_{\mathrm{GRB}} \sim \Gamma^2 c \Delta t_{\mathrm{slow}} \sim 3 \times 10^{15}$ cm $\Gamma_{2.5}^2 \Delta t_{\mathrm{slow}}$. The ICMART model invokes such a large distance as the GRB site. Δt_{slow} is interpreted as the time scale for the magnetized blobs to travel to this distance and to radiate. The slow component may also be interpreted in the photosphere model and the internal shock model, but both require that the central engine has an *intrinsic* time scale of seconds, which is convolved with a much smaller variability time scale. One way to produce such a long time scale may be through interaction with the progenitor stellar envelope (Morsony et al., 2010). For the internal shock model, such a large time scale would correspond to internal shocks at large radii ($\sim 10^{15}$ cm).

2. *Lightcurve (II): fast variability component:* Some GRB lightcurves have rapid variabilities with durations $\Delta t_{\mathrm{fast}} \sim$ several milliseconds. The standard formula gives $R_{\mathrm{GRB}} \sim \Gamma^2 c \Delta t_{\mathrm{fast}} \sim 3 \times 10^{13}$ cm $\Gamma_{2.5}^2 \Delta t_{\mathrm{fast},-2}$. Within the photosphere and internal shock models, these time scales are also related to the intrinsic time scales at

Table 9.1 Grading chart for three representative GRB prompt emission models

Criterion	Photosphere	IS	ICMART
Lightcurve properties:			
Slow variability	Yes	Yes	Yes
Fast variability	Yes	Yes	Yes
Superposition	Yes	Yes	Yes
E_p evolution: hard-to-soft	No	Yes(?)	Yes
E_p evolution: tracking	Yes	Yes(?)	Yes(?)
Spectral lags	No	No(?)	Yes
Power density spectrum	Yes	Yes	Yes
Spectral properties:			
Origin of E_p	Yes	Yes	Yes
$\alpha \sim -1$	Yes(?)	Yes(?)	Yes(?)
$\alpha > -2/3$	Yes	No(?)	No(?)
β	Yes	Yes	Yes
Narrowness	Yes	Yes(?)	Yes(?)
E_p distribution	Yes(?)	Yes(?)	Yes(?)
Thermal component	Yes	No	No
High-energy component	No(?)	Yes(?)	Yes(?)
Other properties:			
γ-ray radiative efficiency	Yes	Yes(?)	Yes
γ-ray polarization	Yes(?)	Yes(?)	Yes
Optical polarization	No(?)	No(?)	Yes
Neutrino upper limit	No(?)	No(?)	Yes
Three-parameter correlations	No(?)	No(?)	Yes(?)

the central engine. Within the ICMART model, these time scales are related to the operation time scales of the mini-jets in the observer frame (Zhang and Zhang, 2014).

3. *Lightcurve (III): superposition:* GRB lightcurves often show superposition of fast variabilities on top of slow variabilities. In the photosphere model, this is due to the convolution of the intrinsically fast and slow variabilities at the central engine. *Different observational times correspond to emission from different fluid elements in a continuous, variable outflow.* In the internal shock model, the observed emission is a result of superposition of emission from internal shocks at different radii. For a slow pulse with superposed rapid variability, *at any observational epoch, the observed emission is from two different emission elements at two distinct emission regions (a closer-in internal shock and a further-out internal shock).* In the ICMART model, the slow component is the emission from one single fluid unit which is streaming outward as a function of time. The fast component is the variable emission of local mini-jets within the bulk ejecta. *For the slow-component pulse, different observational times correspond to different emission times from the same fluid.*

4. *E_p evolution:* Even though all three models can successfully explain GRB lightcurves with very different assumptions, when combining spectral and temporal information

the issues of the models are exposed. Observationally, one can see two E_p-evolution patterns: hard-to-soft evolution and intensity tracking (§2.1.3). Both patterns are related to the broad pulses (the slow variability component), which indirectly suggests that the entire broad pulse may be one radiation unit. This raises challenges to the photosphere model and small-radius internal shocks, which attribute the broad pulses as the time history of the central engine. Whereas the tracking pattern may be explained in these models (for both thermal and synchrotron emission the peak photon energy and luminosity are positively related), the hard-to-soft evolution pattern is at odds with these models, since it is highly contrived to produce both high-E_p emission (beginning of the pulse) and low-E_p emission (end of the pulse) when emission luminosity is low (Deng and Zhang, 2014b). Within the ICMART model, the hard-to-soft E_p evolution is naturally expected, since one pulse corresponds to one radiation unit and magnetic field strength naturally drops as the emitter expands (Uhm and Zhang, 2014b). The internal shock model can also interpret the slow component as one radiation unit by invoking a large-radius internal shock. However, if there are fast-variability components superposed on the slow components, one then has to also invoke small-radius internal shocks. It is then unclear how E_p would behave in accordance with the slow component and, in particular, produce a hard-to-soft evolution pattern.

5. *Spectral lags:* Soft pulses are usually broader and lag behind hard pulses (§2.1.2). This again requires the entire pulse to be one radiation unit, and E_p sweeps across the band as a function of time. This is natural for the ICMART model and the large-radius internal shock model, but essentially impossible for photosphere and small-radius internal shock models (Uhm and Zhang, 2016b). The large-radius emission models also predict a unique connection between the spectral lag behavior and the E_p-evolution pattern, e.g. a positive lag is related to hard-to-soft evolution of E_p (e.g. Uhm et al., 2018). Such a feature is seen in the data. On the other hand, there is no such prediction within the photosphere models or the small-radius internal shock models.

6. *Power density spectrum:* The power density spectrum of GRBs is essentially a power law. For photosphere and internal shock models, it depends on the central engine variability time scale distribution. For the ICMART model, it mostly depends on the distribution of σ and orientations of the mini-jets in the emission region. Since many uncertainties are involved, all the models can be made consistent with the data by adjusting input parameters of the models.

7. *Spectra (I): origin of E_p:* The photosphere model usually interprets E_p as related to the photon temperature of an adiabatically cooled fireball. Some dissipative photosphere models define E_p as the electron temperature, which could be higher than the seed photon temperature. The IS and ICMART models invoke synchrotron radiation as the dominant emission mechanism. The peak energy E_p is usually defined as the electron injection frequency ($E_p = h\nu_m$) since fast cooling is usually expected. In the case of slow-heating, E_p is defined as the synchrotron frequency of thermal electrons due to heating/cooling balance.

8. *Spectra (II): low-energy photon index of the Band-function component $\alpha \sim -1$:* The simplest photosphere model predicts too hard a spectrum, with $\alpha \sim +0.5$. Superposition of many radiation units with very different E_p's may soften the spectrum to the

desired value, but significant variation of the fireball launch radius r_0 may be needed (Deng and Zhang, 2014b). Alternatively, a special structured jet (much smaller Γ at the wing but a top-hat luminosity distribution) may help to achieve $\alpha \sim -1$ (Lundman et al., 2013). For synchrotron radiation in the deep fast cooling regime, $\alpha = -1.5$ is predicted, which is too soft. The problem can be solved within the ICMART model or large-radius internal shock model, when considering fast cooling in a decreasing magnetic field at a large radius from the central engine (Uhm and Zhang, 2014b).

9. *Spectra (III): $\alpha > -2/3$:* A fraction of GRBs have $\alpha > -2/3$, which exceeds the "synchrotron line of death" (Preece et al., 1998). For these GRBs, the IS and ICMART models are ruled out, and the photosphere model is validated.

10. *Spectra (IV): high-energy spectral index β:* The β index can be explained in all the models; it is related to the electron energy spectral index in the emission region for IS and ICMART models, and is related to the optical depth and energy gain in multiple IC scattering in the photosphere Comptonization model (Eq. (5.141)), but notice the different definitions of β).

11. *Spectra (V): the "narrowness" of spectra:* It has been claimed in several papers (e.g. Axelsson and Borgonovo, 2015; Yu et al., 2015) that the shape of the spectrum near E_p is too narrow to be explained by the synchrotron models. These authors used the Band function to fit the data and then compared the synchrotron model with the Band-function results. Such an approach is flawed. In spectral fits of GRBs, due to the limited photon numbers, multiple models can reach an equally good fit to the same data, since the data adjust to the input models to reach the best fit. Zhang et al. (2016b) have shown that the empirical Band function and the more detailed synchrotron model (Uhm and Zhang, 2014b) can equally well fit the time-resolved spectra of the bright *Fermi* GRB 130606B, which has typical Band spectra in essentially all time bins. This suggests that the synchrotron radiation model can interpret the typical Band-function spectra of GRBs. One issue for the ICMART model is that the observed spectrum is the superposition of emission from many mini-jets, which tends to further broaden the observed spectrum. Similarly, the internal shock model that invokes both small- and large-radius internal shocks also faces the problem that emission comes from different regions, so that the superposed spectra could be further broadened.

12. *Spectra (VI): distribution of E_p among GRBs:* Even though bright GRBs have a relatively narrow E_p distribution around 200–300 keV (Preece et al., 2000), the global distribution of E_p is wide, from several keV for GRB 060218-like X-ray flashes (Campana et al., 2006; Sakamoto et al., 2005) to $\sim$15 MeV for GRB 110721A (Axelsson et al., 2012). Such a distribution may be good for both photosphere and synchrotron models (Zhang and Mészáros, 2002a). The 15 MeV E_p of GRB 110721A is beyond the "death line" of the photosphere models that interpret E_p as the adiabatically cooled fireball temperature (Zhang et al., 2012a). It may be accommodated if one introduces a higher electron temperature in the dissipative photosphere scenario.

13. *Spectra (VII): thermal component:* A quasi-thermal component, usually characterized as a quasi-blackbody in the time-resolved spectra and a multi-color blackbody in the time-integrated spectra, is claimed in some GRBs. Sometimes it is the dominant component of the burst (e.g. GRB 090902B, Ryde et al. 2010). In some other

cases, it is a sub-dominant component (e.g. Guiriec et al., 2011; Axelsson et al., 2012; Guiriec et al., 2013). This component is clearly of a photospheric origin, and the IS and ICMART models cannot interpret it.

14. *Spectra (VIII): the high-energy component:* Some GRBs show a high-energy component extending to high-energy bands (Abdo et al., 2009a; Ackermann et al., 2010). This component clearly arises from a different emission region. In several cases, it even extends to much lower energies (Guiriec et al., 2015). This component is not straightforwardly expected from the models, and its origin is a mystery. For GRB 090902B, the power-law component extending from low to high energies can be explained as the synchrotron + SSC/EIC components from the internal shocks (Pe'er et al., 2012).

15. *γ-ray radiative efficiency:* Both the photosphere model and the ICMART model can give high γ-ray radiative efficiency as observed in some GRBs (Lloyd-Ronning and Zhang, 2004; Zhang et al., 2007a). The IS model may reach a high efficiency under some contrived conditions. Some GRBs do have a low efficiency (Wang et al., 2015b), which is consistent with the IS model.

16. *γ-ray polarization:* Strong linear polarization (Π equals tens of percent) of GRB γ-ray emission has been claimed in several bright GRBs, even though the significance of the detections was not high (§2.1.5). The ICMART model invokes ordered magnetic field lines in the emission region, and naturally explains the polarization data. The internal shock models without an ordered magnetic field configuration and the photosphere models cannot account for such large polarization degrees, unless significant geometric effects (viewing angle near or outside the jet cone or the jet is structured) are invoked. A statistical analysis of many GRBs with polarization measurements in the future (with a sensitive γ-ray polarimeter) may differentiate among the models (Toma et al., 2009).

17. *Polarization of early optical afterglows:* Polarized early optical afterglows have been observed in several cases (§8.9). The lightcurves of most of these cases are consistent with them being of a reverse shock origin. The polarization degree is around $10-30\%$, signficant but on average lower than that in the prompt γ-ray phase. This is consistent with the ICMART model, which suggests that the prompt emission phase is a process of destroying the local ordered magnetic fields, so that the post-dissipation ejecta have less-ordered magnetic field configurations in the emission region. Within the photosphere and internal shock models, one needs to invoke an off-axis geometry (e.g. Waxman, 2003) to account for the data. However, these GRBs that show polarized optical emission are typically bright GRBs, suggesting a nearly on-axis geometry. One may also introduce an ordered magnetic field in a low-σ flow relevant to the photosphere and internal shock models. However, due to the large Reynolds and magnetic Reynolds numbers of the flow, the ordered field would quickly be randomized and cannot be retained until the deceleration radius unless σ is large enough (Deng et al., 2017).

18. *Neutrino flux upper limit:* A matter-dominated GRB with small dissipation radii (photosphere and IS) would give rise to bright neutrino emission from GRBs (see Chapter 12 for details). The ICMART model invokes a large emission radius, so that the neutrino flux is greatly reduced (Zhang and Kumar, 2013). *IceCube* now places

progressively stringent upper limits on the GRB neutrino flux (Abbasi et al., 2012; Aartsen et al., 2015, 2016, 2017a,b). The parameter spaces of both photosphere and internal shock models are greatly constrained, while the ICMART model comfortably satisfies the observational constraints with typical parameters.

19. *Correlations:* Observationally, there are several empirical correlations among $E_{\gamma,\mathrm{iso}}$ or $L_{\gamma,p,\mathrm{iso}}$ with E_p, Γ_0, etc. (§2.6). The Amati/Yonetoku relation, i.e. $E_p \propto E_{\gamma,\mathrm{iso}}^{1/2}$ or $E_p \propto L_{\gamma,p,\mathrm{iso}}^{1/2}$ can be explained in all three models (Thompson, 2006; Zhang and Yan, 2011; Fan et al., 2012; Mochkovitch and Nava, 2015). This is because the expressions for E_p in these models all have a Γ_0 dependence. The freedom of allowing Γ_0 to be related to other parameters may allow interpretation of the relations (Zhang and Mészáros, 2002a). When the correlation between Γ_0 vs. $E_{\gamma,\mathrm{iso}}$ or $L_{\gamma,p,\mathrm{iso}}$ is taken into account, and especially the three-parameter correlations among E_p, $E_{\gamma,\mathrm{iso}}$ or $L_{\gamma,p,\mathrm{iso}}$, and Γ_0, the photosphere and the internal shock models both fail to account for the correlations (Liang et al., 2015).

From Table 9.1 one can see that none of the three models can pass all observational criteria. This suggests that GRB emission is not powered by one single emission mechanism, and different bursts may require different mechanisms to power the observed γ-ray emission. Since these models differ from each other by their jet compositions, one can draw the following tentative conclusion.

GRB jets have a distribution of jet compositions, in particular, the σ_0 and η parameters at the central engine may differ from burst to burst. For low-σ_0 events, the evolution of the outflow follows the standard fireball shock model, with emission powered by the photosphere and internal shocks. For high-σ_0 cases, the photosphere emission is suppressed, and the non-thermal emission is powered by magnetic dissipation, possibly triggered by internal collisions or other instabilities. Based on the observational data, one may draw the conclusion that the pure fireball cases are not common, and the magnetically dominated outflows or hybrid jets may make up the majority of bursts.

Exercises

9.1   Following the standard fireball model, derive R_0 and η based on the observed photosphere properties, i.e. derive Eqs. (9.29) and (9.30).

9.2   Derive the energy dissipation efficiency of the internal shock (Eq. (9.72)) and the ICMART (Eq. (9.134)) models.

9.3   Derive the expression of E_p in the internal shock (Eq. (9.76)) and ICMART (Eq. (9.135)) models.

9.4   Derive the numerical values of the typical GRB ejecta parameters within the context of the plasma physics in §9.7.1.