THE STRUCTURE OF THE BROAD EMISSION LINE REGION AS SHOWN BY VARIABILITY MONITORING

JULIAN H. KROLIK
Johns Hopkins University

Abstract.
The various methods used to infer the physical conditions and location of the material responsible for broad line emission in AGN are reviewed. Recent efforts have focussed on reverberation mapping, whose basic concepts and experimental constraints are discussed. A new method for analyzing the results of monitoring experiments, regularized linear inversion, is presented. This method is then applied to published data from the 1989 IUE campaign on NGC 5548, and the results found contrasted with those obtained by the previous standard method, maximum entropy.

1. Introduction

Studies of the broad emission lines in active galactic nuclei have engaged very large numbers of astronomers ever since the discovery of AGN. Precisely because there is such a large volume of work in this field, it is very easy to get lost in the details. For this reason, I will begin by summarizing why this work is felt to be valuable, and will try to highlight what our principal goals are.

The first and lowest level goal is to understand the physical conditions in the source(s) of these emission lines, and to determine their principal production mechanisms. There has been, I believe, substantial progress toward accomplishing this task. A related problem is to locate the line emission region with respect to the other structures in AGN. While there has been much recent work in this area (whose discussion will form the bulk of this review), this problem is by no means completely solved.

The next step up in sophistication is to understand the dynamics of the line-emitting material. At present, while we have a measure of the volume-integrated velocity distribution function in the form of line profiles, we do not know whether the *local* velocity distribution has any special spatial orientation with respect to the center of the AGN, *i.e.* , we don't know whether the matter is falling in, streaming out, or following stable bound orbits. Nor do we know which forces are responsible for accelerating the matter. Whatever these are, there must be some very special constraints built into the hydrodynamics because the observed 1-d velocity dispersion (typically several thousand km/s) corresponds to Mach numbers of several hundred with respect to the sound speed inside the line-emitting gas.

An area about which we would certainly like to learn, but about which even less is known is the "natural history" or life-cycle of the emission line gas. While there have been many speculations voiced about the origin and ultimate fate of the emission line gas, none has achieved any wide acceptance. To give a sense of the vast range of uncertainty in this area, I'll simply list a sampling of these

T. J.-L. Courvoisier and A. Blecha: Multi-Wavelength Continuum Emission of AGN, 163–172.

suggestions: thermal instabilities in an accretion flow (Krolik 1988); winds from red giant stars in a dense stellar cluster (Scoville and Norman 1988; Kazanas 1989); or the surface of the accretion disk in which many of us would like to be able to believe (Dumont and Collin-Souffrin 1990).

Finally, if we are fortunate, we will be able to study AGN emission lines not just to learn about their own origin and character, but as a device to learn about other elements of the AGN system. It has long been a hope in this field that once we understand well the properties of the emission line region itself, we will be able to use the emission lines as "plasma diagnostics" of surrounding regions. If the line-emitting matter is in pressure balance with ambient, less-observable gas (or even if it is not in pressure balance, if there is a calculable relation between the two pressures), then knowledge of the physical conditions in the emission line region carries over into knowledge of those in the hot gas. If in addition there is dynamical coupling between emission line gas and nearby unseen gas (through drag, for example, or viscosity in an accretion disk), then the emission lines tell us about the ambient gas's motions. Reliable results in this area still remains elusive.

2. Techniques for Reaching These Goals

A variety of methods have been invented for accomplishing these objectives. The simplest—and oldest—way to find the physical conditions prevailing in emission region is to construct a single-zone photoionization model, an idea that goes back almost twenty-five years to Bahcall and Kozlovsky (1969), and, in its first detailed realization, to work by Davidson (1972) and MacAlpine (1973). The fundamental idea behind this method is to assume that the emission lines are powered by photoionization. While this idea was initially controversial, the success of these models in crudely reproducing the observed relative line strengths, and, more recently, the excellent correlation (at a delay) found between fluctuations in the continuum and fluctuations in the lines (*e.g.* Clavel *et al.* 1991; Peterson *et al.* 1991; Reichert *et al.* 1993), very strongly support its basic assumption.

To carry out this method, one varies three free parameters: the gas pressure (or density), its thickness (either in length or column density), and the ionization parameter at its exposed edge (defined variously as the ratio of ionizing photon density to gas density, or ionizing photon energy density to gas pressure (this version is denoted Ξ), or in one of several other almost equivalent forms). Using known atomic physics data, one then computes the local ionization equilibrium, local thermal balance, and local excited state population balance in selected atoms and ions for gas subjected to a continuum of a specified spectral shape. It is, of course, a weakness of this method that while we know a great deal about AGN spectra at energies of several to 10 eV, and also 0.5 to 20 keV, we have very little direct knowledge of their spectra in the range between these energies, which is exactly the range of most interest for photoionization. From the computed local physical conditions, one then calculates the total photon emission from the gas,

and compares the relative strengths of the emission lines to the values observed. By varying the free parameters, one eventually finds values which more or less reproduce the observed line strengths. In order to get the best fit to the greatest amount of line flux using only a single zone, one usually finds pressures $p \sim 0.01 - 0.1$ dyne cm^{-2}, column densities $N \sim 10^{22} - 10^{23}$ cm^{-2}, and ionization parameters (defined in the pressure ratio form) $\Xi \lesssim 1$ (*e.g.* Kwan and Krolik 1981; Krolik and Kallman 1988; Rees, Ferland, and Netzer 1990). Strikingly, because the relative line strengths vary comparatively little from object to object, these parameters are crudely similar for all AGN. They may be combined with the luminosity in any particular object to estimate the mean distance of the emission line region from the source of the continuum, assuming free-streaming propagation in between:

$$\langle r \rangle \simeq \left[\frac{L_{ion}}{4\pi c p \Xi} \right]^{1/2} \sim 100 L_{ion,44}^{1/2} \text{lt-d}, \tag{1}$$

where L_{ion} is the luminosity in the ionizing band.

However, this method has a number of unsatisfactory points. First of all, while the quality of fit to the line strengths one obtains with a single zone model is surprisingly good (there is no *a priori* reason why a single zone model should come anywhere close), it is by no means perfect, and because there is no clear statistical measure of the quality of fit, one can never be quite sure how well determined the mean parameters are, or whether the quality of the fit is as good as could be expected from data of a given signal/noise. In addition, there are a number of hidden assumptions making it somewhat model-dependent. For example, nearly all these calculations assume slab geometry. In addition, while assuming solar abundances gets one close to the observed line strengths, there is the possibility that the abundances are actually different. Unfortunately, it is very difficult to find an unambiguous signature of any particular abundance deviation (Davidson 1975; Kwan and Krolik 1981; Hamann and Ferland 1992). The final disability of this procedure lies at its very heart: we certainly do not expect that the distribution in physical conditions is a perfect delta function, and we would like to know about the spread in conditions, if possible correlating that spread with location and/or kinematics.

The next simplest procedure is to construct photoionization models as a function of line of sight velocity, *i.e.* with respect to the line profiles (Kallman *et al.* 1993). By doing so, one effectively projects the line emissivities onto a set of nested surfaces, the surfaces of constant line of sight velocity. The mean physical conditions inferred for each model then correspond to the conditions averaged over each of these surfaces. Unfortunately, this technique suffers from two nasty illnesses: we do not know the shapes of these isovelocity surfaces *a priori*, so we do not know where these inferred conditions apply; and worse, we do not even know if these surfaces exist. To the degree that the line emitting gas moves like a fluid, its velocity is well-defined at each point; however, if it moves collisionlessly (for example, if the gas is broken up into a large number of small clouds, or if the gas resides on

the surfaces of stars), its velocity distribution at any given point could be quite broad. When that is the case, the isovelocity surfaces become thick and start to overlap. Interpreting a profile-based photoionization model in such a situation is virtually hopeless.

3. Reverberation Mapping

Variability in AGN permits the application of a quite different method to this problem, the method of "reverberation mapping". While first proposed in concept many years ago (Bahcall, Kozlovsky, and Salpeter 1972), and first developed mathematically more than a decade ago (Blandford and McKee 1982), it was first attempted only in the last few years (Clavel *et al.* 1991; Maoz *et al.* 1991). The reason for this long delay, as we shall see, has to do with the very large quantity of data required, and the special planning which must go into the collection of this data if they are to be useful. Before discussing these issues, I digress to outline the basic idea behind the method.

Careful monitoring of AGN continuum emission (particularly for those of lower luminosity) shows that they are often quite variable in the optical and ultra violet. If the ionizing continuum follows the same light curve (as is suggested by the excellent correlation at zero lag between the optical and non-ionizing UV bands: Krolik *et al.* 1991), then the emission line response from any small volume within the broad emission line region should vary along with the ionizing continuum. Because it takes a finite time for the gas to adjust its equilibrium to the new value of the continuum flux, in principle there can be a small local lag; however, at the densities indicated by single-zone photoionization modelling, the local equilibration time is very short, perhaps $\sim$ 1 minute. By contrast, as we have already estimated, the light travel time across the region is far longer, at least days and possibly many weeks even in low luminosity AGN. Consequently, the line light curve we measure should be a delayed, and smoothed, replica of the continuum light curve. Because we understand (or think we do) how emission line gas responds to changes in the ionizing continuum, we can use monitoring data for both the continuum and the emission lines to invert this argument and give us information about the internal geometry of the emission line region. That is to say, we can project out a map of the emission line region, where the surfaces of projection are the surfaces of constant light travel-time with respect to us, defined by

$$r = c\tau(1 - \cos\theta), \tag{2}$$

where the polar axis with respect to which θ is defined is the direction towards us.

There are (at least) two major difficulties with this method: First, it requires a very large quantity of data, and the data must be obtained in the right way. To see just how much, and what the right way is, we write the relation between the

line light curve and the continuum light curve as

$$\delta F_l(t) = \int d\tau \Psi(\tau) \delta F_c(t-\tau), \tag{3}$$

where $\delta F_{l,c}$ are the fluctuations in the line and the continuum with respect to their mean values, and the response function $\Psi(\tau)$ is the marginal emissivity of the line with respect to changes in the continuum flux, averaged over the surface with delay τ. Assuming a linear relation between continuum fluctuations and line fluctuations is only a good approximation when $\delta F_l / \langle F_l \rangle \ll 1$, but this is commonly the case. Equation 3 is a convolution relation, and so has the formal solution

$$\Psi(\tau) = \int df\, e^{-2\pi i f \tau} \frac{\hat{\delta} F_l(f)}{\hat{\delta} F_c(f)}, \tag{4}$$

where the symbol $\hat{X}$ denotes the Fourier transform of X. Thus, if the response function has structure on the scale τ_o, for us to measure it the AGN must have significant variability on that timescale, and we must sample that variability on a timescale shorter than $\tau_o/2$. In addition, to obtain a statistically reliable measure of that variability, we should extend our observations for times considerably longer than τ_o. To do this well, and to be sensitive to a reasonably wide range of possible timescales, requires many many observations. Simulations, and experimental experience, show that $\simeq 50$ observations is a bare minimum. Moreover, to obtain the least biassed picture of the variability components on different timescales, the observations should preferably be evenly-spaced. As several recent campaigns have shown (Clavel *et al.* 1991; Peterson *et al.* 1991; Reichert *et al.* 1993) it is possible to do this, but a very large amount of labor is required.

The second difficulty is actually inverting the convolution equation. Despite the existence of the formal solution just shown, this is not so easy. Unfortunately, in most cases the Fourier solution cannot be used. The problem is that most AGN fluctuation power spectra are fairly "red", *i.e.* most of the power is found at low frequencies. As a result, measurement error, which has a white noise spectrum, overpowers the signal at high frequencies. At the same time, though, this lost high frequency information is necessary if we are to obtain maximal resolution in the response function.

Another possible approach is to discretize the convolution according to the actual sampling. The integral equation then has the appearance of a conventional linear equation, which one might think should be directly soluble. But this, too, does not work because the smoothing produced by the convolution creates a large ambiguity in the possible solutions (*i.e.* many different kinds of smoothing all create the same final result). This ambiguity is expressed mathematically by the fact that the matrix representing the integral equation kernel always has at least a few eigenvalues many orders of magnitude smaller than the typical value of elements in the matrix. If there is any noise in the data with projection onto the corresponding eigenvectors, the implicit matrix inversion of the linear solution

multiplies these noise components by numbers of order the inverse of the very small eigenvalues; these inverses are, of course, very large. In practise, this noise amplification is completely catastrophic.

Hitherto, this problem has been solved by a variety of model-fitting techniques, most prominently maximum entropy (Krolik *et al.* 1991; Maoz *et al.* 1991). While model-fitting is easy to make stable, it, of course, always entails some level of model-dependence in the answer. In the maximum entropy version, it is particularly hard to trace the impact of particular model assumptions on the solution. Maximum entropy also has the further drawback of requiring positive values of Ψ because the nonlinear function it maximizes to select otherwise equally acceptable solutions is undefined for negative arguments. While most lines respond positively to continuum fluctuations in most circumstances, this is not true in general (Gaskell and Sparke 1986; Sparke 1993), and it would be desirable to test for negative responses.

4. A New Inversion Method: Regularized Linear Inversion

Fortunately, there is a direct inversion method, called regularized linear inversion, which does not suffer from these defects, and is also computationally very efficient (Press *et al.* 1992). Christine Done and I have recently shown how to apply this method to the reverberation mapping problem (Done and Krolik 1993), and I summarize our results in the remainder of this review. The heart of this method is the recognition that in order to break the degeneracy of the inversion one must inject some *a priori* information. This is done by simultaneously minimizing both the deviation of the solution from the measured data and the deviation of the solution from one's *a priori* assumption. Obviously, this method entails model-dependence; its beauty is in the clear dependence of the solution on these model assumptions through a single tunable parameter.

More quantitatively, if we wished solely to find the solution which best reproduced the observed data, we would minimize the quantity

$$\chi^2 = (C \cdot \boldsymbol{\Psi} - \mathbf{L})^2 , \tag{5}$$

where $\mathbf{L}$ is the list of observed line flux fluctuations normalized by their uncertainty

$$L_i = \delta F_l(t_i)/\epsilon_i, \tag{6}$$

$\boldsymbol{\Psi}$ is the list of values of the response function at lags τ_j, and the matrix C is the discretized kernel of the integral equation similarly normalized

$$C_{ij} = \delta F_c(t_i - t_j)/\epsilon_i. \tag{7}$$

On the other hand, if our *a priori* condition is to look for the "smoothest" possible solution, this means that we are searching for solutions with small derivatives, and these can be represented by a differencing operator R. Thus, the total quantity

to be minimized is the sum $\chi^2 + \lambda \mathbf{\Psi} \cdot R^T \cdot R \cdot \mathbf{\Psi}$. The balance between satisfying the data and satisfying our prejudices is expressed by the tunable parameter λ. Because both quantities to minimize are quadratic forms in Ψ, the solution is found by solving a simple linear equation

$$\left(C^T \cdot C + \lambda R^T \cdot R\right) \cdot \mathbf{\Psi} = C^T \cdot \mathbf{L}. \tag{8}$$

Even though the solution of either minimization separately would be disastrously unstable because of the small eigenvalue problem, their simultaneous solution is quite stable because the probability that the eigenvectors of the two pieces coincide is extremely small. This method also possesses the virtues of very clear error propagation, and directly testable model-dependence through manipulation of λ.

5. Application to Real Data

There have now been four major monitoring programs with sampling good enough to attempt a solution for the response function: a purely ground-based program on NGC 4151 (Maoz *et al.* 1991); one on NGC 5548 combining IUE data with coordinated observations from the ground (Clavel *et al.* 1991; Peterson *et al.* 1991); another IUE plus ground program on NGC 3783 (Reichert *et al.* 1993; Stirpe *et al.* 1993); and, most recently, a return to NGC 5548 using HST, IUE, and more ground observations (Korista *et al.* 1994). Of these, the one most suitable for immediate analysis with linear regularization is the 1989 IUE campaign on NGC 5548. As can be seen from the forms just presented, the method works best with evenly-sampled data, so the ground-based NGC 4151 data is eliminated.

The HST data is not yet completely reduced (see Peterson's article in this volume for a preliminary look at the light curves), so it is eliminated. Finally, the NGC 3783 campaign was the victim of bad fortune: the utility of any monitoring data clearly depends on the ratio of real variance on the relevant timescales to noise variance, and this Seyfert galaxy happened not to vary terribly much on the right timescales.

To place this program in context, I will just remind you that NGC 5548 is a nearby type 1 Seyfert galaxy whose monochromatic luminosity νF_ν in the UV is $1.6 \times 10^{43} h^{-2}$ erg s^{-1}. Interestingly, *no* single-zone photoionization model provides a good match to its line strengths: at least two zones are required, one at $\simeq 10h^{-1}$ lt-d distance, the other at $\sim 200h^{-1}$ lt-d (Krolik *et al.* 1991), When maximum entropy analysis is applied to the seven UV line light curves (Krolik *et al.* 1991), seemingly robust solutions can be found for five of the lines, HeII 1640, NV 1240, Lα, CIV 1549, SiIV 1400, and CIII] 1909. The other two lines, MgII 2800 and OI 1304, did not vary enough relative to the measurement error. In the maximum entropy solution, the response functions of all but CIII] 1909 peak either at or very near zero lag, and trail off towards higher lags. The maximum lag at which significant response is found increases more or less with declining ionization level,

from HeII 1640, whose response function is small beyond $\simeq$ 15d, to SiIV 1400, which retains significant response out to 30d or more. However, in evaluating these results, it is important to bear in mind that the criterion of acceptable χ^2 was defined with respect to all line light curves simultaneously, so substantial deviations in individual line light curves could be, and are, present.

The picture as seen by linear regularization is surprisingly different. Taking the solutions one line at a time, one finds that only three lines yield solutions with acceptable χ^2: Lα, CIII] 1909, and SiIV 1400. For all three of these, the reduced χ^2 is $\simeq$ 1.5. Even with $\lambda = 0$, there is *no* solution yielding acceptable χ^2 for the other lines. The reason for this discrepancy between the two methods is partly that the χ^2 for some of the individual light curves found by maximum entropy is, in fact, rather large, and partly that the maximum entropy solution used treated the continuum light curve as a set of free parameters constrained, but not fixed, by the observations. This extra freedom makes the number of free parameters actually rather larger than the number of unknown variables.

In the case of Lα, the response function derived by regularized linear inversion has very small, and possibly negative, amplitude at zero lag, rising to a peak at around 8d. Beyond 16d the response amplitude is fairly small. In contrast, the maximum entropy solution peaks at zero, and falls smoothly to low levels beyond 20d.

The regularized inversion solution for CIII] 1909 is strikingly different from the maximum entropy solution. The new method finds a response amplitude which is large and negative at zero lag, but which rises sharply with increasing lag. It passes through zero around 4d, and peaks about 10d, where it rolls off gently. The maximum entropy solution for this same data was positive everywhere (by assumption), but had a clean peak between 20 and 30d.

The response function found by regularized linear inversion for SiIV 1400 qualitatively resembles that of CIII] 1909, but it is possible that an acceptable reduced χ^2 is achieved not because the deviation from the observed light curve is small, but because the error bars are very large. A similar comment applies to the maximum entropy solution for this line.

What are we to make of the lines for which no acceptable solution is found ? This class is best exemplified by CIV 1549, the line with the best effective signal/noise of all the lines. The solution for this line produces a light curve which tracks the observed light curve fairly well up until its final positive excursion. At that point, while the line flux increases sharply, the continuum immediately before has only a weak maximum. Consequently, it is very hard for *any* time-steady linear response function with a width shorter than the interval between major continuum fluctuations to reproduce the line light curve. In the maximum entropy solution, this problem was cured by introducing significant response at a lag of 180d. Because the stretch of data constraining the solution at such large lags is relatively short, we cut off the linear inversion response function at a maximum lag of 80d. Thus, use of the linear inversion technique brings clearly into focus that the simplest model

of line response fails to explain a significant element in the light curve. Just what is needed to fix it is unclear. Possibly response at very large lag, as suggested by the maximum entropy solution, is the correct answer. Possibly the response function changes on a timescale shorter than the duration of the monitoring, as suggested by Netzer and Maoz (1990). Possibly the line is responding to a continuum component varying in a way which is not simply proportional to the near-UV continuum. At this stage we do not have enough guidance from the data to choose the correct explanation.

6. Conclusions

Much effort has been expended over a very long period of time to try to understand the physical conditions and location of the broad emission line gas in AGN. The basic idea of photoionization is amply confirmed, but many details remain to be worked out.

The most active portion of this field at the moment is the application of variability monitoring techniques to the inference of the geometrical structure of the broad emission line region. A number of monitoring campaigns have generated very impressive data sets, but their interpretation is still problematic. The dramatic contrast between the response functions obtained by maximum entropy and regularized linear inversion, operating on identical data, demonstrates that even with these very large data sets, we are not in a position to unambiguously determine the response functions. In a sense, maximum entropy provides the "prettiest" possible solution, the one which is found by looking for that continuum light curve which, within the constraints of the measurement, allows the solution to very closely track the observed line light curve. On the other hand, regularized linear inversion, by only considering the "most likely" continuum light curves, presents the most conservative and stringent test of the simple time-steady linear response model.

Homing in on the correct response function is not, of course, the end of our job. Even with that in hand, it will be necessary to construct photoionization models whose values of $\partial F_l/\partial F_c$ averaged over the isodelay surfaces best match the response functions. Velocity-resolved reverberation mapping (as we hope to obtain from the recent HST campaign) will add another layer of complexity to the problem; however, if carried out successfully, it will also carry us to a much deeper level of understanding. In particular, it will only be with the conclusion of that program that we can answer some of the basic question raised at the beginning of this paper, such as the direction of flow within the broad line region.

7. References

Bahcall, J.N. and Kozlovsky, B.-Z. 1969, Astrophysical Journal**155**, 1077.
Bahcall, J.N., Kozlovsky, B.-Z., and Salpeter, E.E., 1972, Astrophysical Jour-

nal**171**, 467.
Blandford, R.D. and McKee, C.F., 1982, Astrophysical Journal**255**, 419.
Clavel, J. *et al.* 1991, Astrophysical Journal**366**, 64
Davidson, K. 1972, Astrophysical Journal**171**, 213.
Davidson, K. 1975, Astrophysical Journal**195**, 285
Done, C. and Krolik, J.H. 1993 in preparation
Dumont, A.M. and Collin-Souffrin, S. 1990, Astronomy and Astrophysics**229**, 313
Gaskell, C.M. and Sparke, L.S. 1986, Astrophysical Journal
Hamann, F. and Ferland, G. 1992, Astrophysical Journal**391**, L53
Kallman, T.R., Wilkes, B., Krolik, J.H., and Green, R.F. 1993, Astrophysical Journal**40 3**, 45
Kazanas, D., 1989, Astrophysical Journal**347**, 74
Korista, K. *et al.* 1994 in preparation
Krolik, J.H., 1988, Astrophysical Journal**325**, 148
Krolik, J.H. and Kallman, T.R. 1988, Astrophysical Journal**324**, 714
Krolik, J.H., Horne, K., Kallman, T.R., Malkan, M.A., Edelson, R.A., and Kriss, G.A. 1991, Astrophysical Journal**371**, 541
Kwan, J.Y. 1984, Astrophysical Journal**283**, 70
Kwan, J.Y. and Krolik, J.H. 1981, Astrophysical Journal**250**, 478.
MacAlpine, G.M. 1972, Astrophysical Journal**175**, 11.
Maoz, D. *et al.* 1991, Astrophysical Journal**367**, 493.
Netzer, H. and Maoz, D. 1990, Astrophysical Journal, Letters to the Editor**365**, L5.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992, it Numerical Recipes, 2d ed., (Cambridge University Press: Cambridge)
Peterson, B.M. *et al.* 1991, Astrophysical Journal**368**, 119
Rees, M.J., Netzer, H., and Ferland, G.A. 1989, Astrophysical Journal**347**, 640.
Reichert, G. *et al.* 1993, Astrophysical Journalin press
Scoville, N.Z. and Norman, C.A. 1988, Astrophysical Journal**332**, 163
Sparke, L. 1993, Astrophysical Journal**404**, 570
Stirpe, G. *et al.* 1993 in preparation