The submitted manuscript has been authored by a contractor of the U. S. Government under contract No. W-31-109-ENG-38. Accordingly, the U. S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U. S. Government purposes.

RECEIVED
OCT 21 1994
OSTI

UNBINNED MAXIMUM LIKELIHOOD FIT FOR THE CP CONSERVING COUPLINGS FOR W^+ PHOTON PRODUCTION AT CDF ¹

by
Kevin Lannon ²
St. Norbert College
De Pere, Wisconsin

Dr. Robert Wagner
High Energy Physics Division
Argonne National Laboratory

MASTER

Abstract

We present an unbinned maximum likelihood fit as an alternative to the currently used fit for the CP conserving couplings in W plus photon production studied at CDF. We show that a four parameter double exponential fits the E_T spectrum of the photon very well. We also show that the fit parameters can be related to and by a second order polynomial. Finally, we discuss various conclusions we have reasoned from our results to the fit so far.

¹Work performed at Argonne National Laboratory, a contract Laboratory of the United States Department of Energy.

²Participant in the summer 1994 Student Research Participation Program. This program is coordinated by the Division of Educational Programs.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

yes

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

I Introduction

Particle physics shows four possible couplings between the W boson and the photon in W plus photon production resulting from high energy proton, anti-proton collisions. Two couplings, $\Delta\kappa$ and λ , are CP conserving couplings. The other two, $\tilde{\kappa}$ and $\tilde{\lambda}$, are CP violating couplings. This summer, I have worked closely with Dr. Robert Wagner on developing an unbinned maximum likelihood fit to obtain the CP conserving couplings, $\Delta\kappa$ and λ , and their limits from the transverse energy (E_T) spectrum of the photons emitted during W plus photon production. This fitting method would be available as an alternative to the current binned maximum likelihood fit already used to determine limits on the couplings for the data collected from the CDF collider detector at Fermi National Laboratory in Batavia, Illinois. To develop this fit, we first have had to derive a functional form to describe the shape of the E_T spectrum of the photon plus the background, and that work still continues. Once we have a functional form and calculate uncertainties on the fit, the fit will be put into a FORTRAN routine, utilizing MINUIT, to be used on actual data taken from the CDF detector.

Currently, the CP conserving couplings are determined by a program developed by Dr. Thomas Muller from the University of California-Los Angeles. This method divides the E_T spectrum of the photon into the following bins: 7-11 GeV, 11-15 GeV, 15-19 GeV, 19-27 GeV, and greater than 27 GeV. The bins are assumed to have an approximately Poisson distribution with an expected number of events calculated by the following equation:

$$N_i = a_i + b_i\Delta\kappa + c_i\lambda + d_i\Delta\kappa^2 + e_i\lambda^2 + f_i\Delta\kappa\lambda + \text{background}$$

This program then calculates the $\Delta\kappa$ and λ that give the maximum likelihood, or greatest probability, of getting the same number of events in each of the five bins as the data. Uncertainties are accounted for through Gaussian smearing of the fit. This fit does not address the nearly 100% correlation among the five bins, but it doesn't have to because the results of the fit are almost entirely determined by the number of events in the "greater than 27 GeV" bin. However, as CDF begins to detect more events with high energy photons, the "greater than 27 GeV" bin will need to be broken up into two or more bins. With several bins at high E_T , the fit will no longer be completely determined by one bin, so bin correlations will become important. Note: for the rest of this paper, I will refer to this fitting method as the UCLA fit.

Our goal in trying to use an unbinned maximum likelihood fit to determine CP conserving couplings is two-fold. Primarily, we hope to solve the potential problem of bin correlations from the UCLA fit by eliminating bins altogether. Instead of looking at the numbers of events from that fall within certain E_T ranges, we will consider the individual E_T 's of each particle and treat the particles as one group. By treating the data set as a whole, we will eliminate any problems that might arise from correlations among the bins as well as any difficulties stemming from the sensitivity of our fit results to the choice of size and number for the bins. Our second goal in working on this new fitting method is that we hope to be able to use it to gain stricter limits on $\Delta\kappa$ and λ for CDF, thus gaining greater sensitivity to detecting "new physics." We will not be able to evaluate our success at this objective until we have found a working functional form and have begun to take fit uncertainties into account. At this point however, it appears that we might be able to develop a fitting method that is completely independent of the measurement of CDF luminosity, thus eliminating the uncertainty on that measurement from the uncertainty on our fit.

Once we have a successful unbinned maximum likelihood fit for the CP conserving couplings in W plus photon production, we will potentially be able to apply the techniques developed to fit

for the CP violating couplings in W plus photon production, both the CP conserving and CP violating couplings in Z plus photon production, and even fit other spectra, like the transverse mass spectrum of the W or the invariant mass spectrum of the Z .

II Scientific Background

This section has been provided for those who are unfamiliar with any of the terms, topics, or concepts presented in the paper.

The project I am working on deals with the analysis of data collected at the CDF collider detector at Fermi National Laboratory in Batavia, Illinois. The CDF detector monitors and records collisions that occur in Fermilab's Tevatron Accelerator. The Tevatron is a 6.5 km circular accelerator that maintains two particle beams in opposite directions, one of protons and one of antiprotons. These beams are each accelerated up to 900 GeV energy. When they collide, the proton and antiproton beams are brought together to produce a center of mass energy for the collision of 1800 GeV, which makes Fermilab's Tevatron the highest energy accelerator operating presently in the world. The ring where the beams circulate is divided into segments whose centers are labeled A0 through F0. The beams collide in two locations on the ring, at D0, which is where the D0 detector operates, and at B0 which is where CDF takes its data.

When protons and antiprotons collide at such high energies, a variety of interactions can occur and many types of particles can be produced. My project concerns itself primarily with the process of W plus photon production, although it will eventually deal with Z plus photon production as well. W 's and Z 's are subatomic particles known as weak gauge bosons. A gauge boson is a particle that mediates a force interaction, like that of gravity or magnetism. W 's and Z 's mediate the weak nuclear force. Photons, the gauge bosons for the electromagnetic force, can be viewed as massless packets of energy. W plus photon production is a process in which both a W particle and a photon result from the proton, antiproton collisions. The W decays into other particles which are picked up by the detector at CDF in addition to the photon, which can be detected directly.

The concept of background deals with other particles or groups of particles that the detector picks up which do not indicate W plus photon production, but which are indistinguishable from W plus photon production. One common source of background for the photon is a particle known as the π^0 . This particle decays into two photons. If the photons enter the detector close enough together, they will look exactly like the single photon produced in W plus photon production. Background is accounted for at CDF in W plus photon production by making an estimate of the level of background present in the data and then factoring in that a certain portion of the data sample comes not from W plus photon production, but actually from W plus "fake" photons. One way to do this in an unbinned maximum likelihood fit is to find a functional form for the E_T distribution of the background and add this function into the overall fit for CP conserving couplings.

In the scientific model of how W plus photon production occurs, the W can have four couplings to the photon, known as κ , λ , $\tilde{\kappa}$, $\tilde{\lambda}$. Two of these couplings, κ and λ , are charge-parity (CP) conserving couplings. The other two, $\tilde{\kappa}$ and $\tilde{\lambda}$, are CP violating couplings. Although the model allows for each of these couplings to take on essentially any numeric value, the current Standard Model of particle physics sets κ to one and λ , $\tilde{\kappa}$, and $\tilde{\lambda}$ to zero. For convenience,

scientists have defined the term $\Delta\kappa \equiv \kappa - 1$. This means for the Standard Model, the couplings are defined as follows: $\Delta\kappa = 0$, $\lambda = 0$, $\tilde{\kappa} = 0$, $\tilde{\lambda} = 0$. My project has concerned itself primarily with the CP conserving couplings: $\Delta\kappa$, λ .

This project involves performing an unbinned maximum likelihood fit. An unbinned fit is different from a binned fit, like the UCLA fit, because in an unbinned fit, the data is treated as a whole rather than divided up into segments or bins. This type of fit can be advantageous for several reasons. In the case of the UCLA fit, binning the data can cause problems because one then has to consider how the bins correlate with, or affect and relate to, to one another. An unbinned fit can avoid this problem. Also, occasionally the answers obtained from one fit can be sensitive to the choice of bin number or bin size. In some problems, choosing a certain number or size for the bins in the fit can cause the results to appear drastically different from the results with a different choice of bin number or size. Again, doing an unbinned fit sidesteps the issue of results being sensitive to bin size.

A maximum likelihood fit is a fit that comes up with an equation that describes the probability of an event or set of events occurring and then maximizes this equation with respect to a set of fit parameters to obtain estimators for the parameters. This is different from a chi-square, or least squares, fit, another fitting method used later in my project to help find the functional form for the photon E_T distribution. For a chi-square fit, one needs an equation that depends on the fit parameters which predicts the results of the experiment, usually based on probabilities. One then calculates the square of the difference between the actual results from the experiment and the predicted results, and weights these values by some estimate of the uncertainties or standard deviations of the results. The sum of these various weighted, squared differences for all the results of the experiment is known as the chi-square value of the fit, or simply the fit's chi-square. The best fit is obtained by minimizing the chi-square value with respect to the fit parameters. This method is extremely useful for fitting histograms.

To help me on my project, I relied primarily on two software packages. The first, MINUIT, is a program that finds the minimum of various functions supplied by the user. MINUIT works on a FORTRAN routine that calculates the value of the function for different values of the fit parameters. MINUIT varies these parameters until it locates the minimum of the function. MINUIT can be used to find the maximum of a function by multiplying the function by negative one and then finding the minimum. The other software package upon which I relied heavily is PAW, Physics Analysis Workstation. PAW is an extremely versatile software package that specializes in the creation and analysis of histograms, a kind of bar graph useful for displaying experimental results. PAW includes options that perform a chi-square or maximum likelihood fit of a function to a histogram as well as the creation of ntuples. An ntuple is a PAW format used to store many characteristics of a set of data points. Ntuples allow easy graphing of any of the data's characteristics into histograms. Both MINUIT and PAW were created by and are available through the CERN Laboratory in Geneva, Switzerland.

To find the functional form for the E_T spectrum, I needed another computer resource in addition to PAW and MINUIT. Because the samples of actual data taken from the collisions at CDF are so small, I needed to use the Baur and Fast Monte Carlo programs to generate larger samples of data. A Monte Carlo program is a program that uses random number generation to simulate complex processes. The Baur Monte Carlo, written by Uli Baur of Florida State University and Ed Berger of Argonne National Laboratory, simulates what happens during the actual proton and anti-proton collisions that result in W plus photon production. The Fast Monte Carlo, written by

Steve Errede of the University of Illinois, simulates the way the CDF detector monitors W plus photon production, including missed and misidentified events. The Fast Monte Carlo is designed to run on data generated from the Baur Monte Carlo.

III Results

Although my project is not yet concluded, we have obtained some useful results.

The transverse energy spectrum of the photon is fit nicely by a function that is the sum of two exponentials. This four parameter, double exponential takes the following form:

$$f(E_T) = \exp(P_1 + P_2 E_T) + \exp(P_3 + P_4 E_T)$$

The first term in the double exponential function controls the shape of the E_T spectrum at low energies. The second terms controls the shape at higher E_T 's. The transition from low E_T term to high E_T term, which usually occurs between 20 GeV and 50 GeV, is what causes the bend in the shape of the E_T spectrum of the photon.

This fit has been tested for the following $\Delta\kappa$, λ pairs over the E_T range of 7 GeV to 100 GeV with Monte Carlo data:

$\Delta\kappa$	λ	$\Delta\kappa$	λ	$\Delta\kappa$	λ	$\Delta\kappa$	λ	$\Delta\kappa$	λ
-3	-2	-1	-1	0	-1	1	-1	2	1
-3	0	-1	0	0	0*	1	0	3	-2
-3	2	-1	1	0	1	1	1	3	0
-2	-1	0	-2	0	2	2	-1	3	2
-2	1								

*Standard Model

Errede has suggested that we might be able to extend our fitting range to 120 GeV and fit to the sum of three exponentials, but we have not had yet tested this possibility.

We have also had some success parametrizing the double exponential in terms of $\Delta\kappa$ and λ . Each of the four parameters in the double exponential fits reasonably to a second degree polynomial for $\Delta\kappa$ and λ although more recent information suggests that this relationship between the double exponential parameters and the CP conserving couplings may be inadequate. Errede originally suggested trying a second degree polynomial because a polynomial of degree two was used to calculate the expected number of events per bin in the UCLA fit. Errede reasoned that there should be some mapping of the expected number of events per bin equations onto our function and he agreed with us that the parameter values looked like they could fit a second order polynomial. Our initial set of parameter values did not contradict a fit to a second degree polynomial, but there were too few values for a decisive fit. New parameter values strongly confirm our original fit for P_1 of the double exponential function, but they suggest that we might want to go to third order or higher for the other parameters. Fit ambiguities, however, make this conclusion uncertain, and we would like to avoid using polynomials of degrees higher than two because increasing the degree of

the polynomial will add coefficients, hence parameters, to our fit for coefficients. Currently, we use the following equation to relate parameter values to $\Delta\kappa$ and λ :

$$P_i = a_i + b_i\Delta\kappa + c_i\lambda + d_i\Delta\kappa^2 + e_i\lambda^2 + f_i\Delta\kappa\lambda$$

Finally, our fit had trouble distinguishing between certain $\Delta\kappa$, λ pairs. These “mirror” pairs occurred as corresponding points in opposite coordinates of $\Delta\kappa$, λ space, such as (0, 1) and (0, -1); (-3, 0) and (3, 0); (-3, -2) and (3, 2); as well as (3, -2) and (-3, 2). These “mirror” pairs produce photon E_T spectra that appear essentially identical. Consequently, our fit cannot tell the difference between them. This feature of the various E_T spectra makes sense if the parameters of the double exponential are described by paraboloid functions. A parabolic shape allows an entire ellipse of points that share the same parameter value and therefore are indistinguishable to that parameter. We are uncertain at this point if there are only pairs of points in $\Delta\kappa$, λ space which are indistinguishable or if there is actually a whole ellipse of indistinguishable points as one might expect from the parabolic nature of the equations that we use to relate the double exponential parameters to $\Delta\kappa$ and λ .

IV Methods

An unbinned maximum likelihood method (MLM) fit, in this case, is computed by maximizing the probability that the photon E_T 's observed resulted from a particular distribution, in terms of $\Delta\kappa$ and λ . To do this we first needed to obtain a probability density function (p.d.f. often written $f[E_T]$) for the E_T distribution, that is, a function that describes the probability that a photon detected during W plus photon production would have a certain E_T . This p.d.f. would also have to be parametrized for $\Delta\kappa$ and λ so that for a given E_T , the best values of $\Delta\kappa$ and λ could be calculated. The likelihood function ($\mathcal{L}$) is then obtained by multiplying the individual probability functions together. This yields the overall probability of getting a particular set of photons with the given transverse energies. The likelihood function can be written as follows:

$$\mathcal{L}(E_T ; \underline{P}) = \prod_{i=1}^N [f(E_T i; \underline{P})]$$

Where $\underline{P}$ is the vector of fit parameters. Each P_i is assumed to be a function of $\Delta\kappa$ and λ .

Often, taking the natural logarithm of the likelihood function simplifies calculations somewhat by changing the product to a summation. This form of the likelihood function, known as the log-likelihood function is given below:

$$\ln[\mathcal{L}(E_T ; \underline{P})] = \sum_{i=1}^N \ln[f(E_T i; \underline{P})]$$

The maximum likelihood fit is therefore obtained by finding the maximum of the log-likelihood function with respect to the fit parameters.

Our plan for coming up with an unbinned maximum likelihood fit of the photon E_T spectrum to determine $\Delta\kappa$ and λ revolves around three main steps: finding the functional form for the E_T distribution without background, parametrizing this functional form in terms of $\Delta\kappa$ and λ , and finally, going back to account for background and fit uncertainties. To find the functional form

for the E_T spectrum, we would first examine the histograms and choose a function that appeared to reasonably fit our data. To make this fitted function usable as a p.d.f., we would then have to normalize it. This involves multiplying in a normalization factor to the function, taking the integral of the function over the range of the fit, setting the integral equal to one, and solving the integral for the normalization factor in terms of the fit parameters. Often it would be possible to eliminate a parameter from the function by making one of the parameters part of the normalization factor. Once we had a normalized function, we would be able to do a MLM fit of the data. If this fit were good, we would attempt to parametrize the function in terms of $\Delta\kappa$ and λ . This would be accomplished by fitting the function to Monte Carlo data for different values of $\Delta\kappa$ and λ , paying attention to how the fit changed with respect to each. It was assumed that the parametrization would ultimately be accomplished by fitting each of the fit parameters P_i to a function of $\Delta\kappa$ and λ , $g_i(\Delta\kappa, \lambda)$ and then substituting each $g_i(\Delta\kappa, \lambda)$ back into the appropriate P_i during the MLM fit. Once we had a good parametrized fit for the function, we would then turn to accounting for background in the fit, perhaps by fitting the background spectrum to a function and including this function as a term in the MLM fit. Also, we would need to come up with a way to calculate uncertainties in the fit. Finally, we would write the entire MLM fitting method including background and uncertainties into a FORTRAN routine that would use MINUIT to obtain the fit. At this point we would be ready to use the MLM method to actually analyze data. Our project is currently involved in the parametrization portion of this plan.

We began our search for the functional form of the photon E_T spectrum by fitting the E_T 's of photons from twenty-five actual events observed at CDF during the 1993/1994 experimental run to an exponential function. This produced a fair fit, but the fit tended to drop off more quickly at high energies than the actual data. From this we inferred that a second exponential term needed to be added to the function to improve the fit. However, the new function, now the sum of two exponentials, proved difficult to normalize, and MINUIT was unable to find a good fit for the complicated and unwieldy normalized double exponential function. At this point, we decided that a sample of data with more than twenty-five points would allow a better chance of finding the functional form. Therefore, we obtained nine sets of Monte Carlo data generated with the Baur and Fast Monte Carlos at the University of Illinois-Champaign Urbana compliments of graduate student Mark Vondracek. The Monte Carlo data was for the following $\Delta\kappa, \lambda$ pairs in the form of PAW ntuples: (-3, -2), (-3, 2), (-3, 0), (0, -2), (0, 0), (0, 2), (3, 0), (3, -2), (3, 2). Having the data in the form of ntuples, which are easily graphed as histograms, allowed us to use PAW's function fitting routine to do a chi-square fit of the data to a double exponential. In the case of fitting histograms, a chi-square fit is more convenient and PAW readily fit all the data sets to a four parameter double exponential function.

Once we had a good fit of the double exponential function to all nine sets of data, we attempted to parametrize the double exponential in terms of $\Delta\kappa$ and λ . Observation to that point, combined with suggestion from Errede indicated that a second degree polynomial of $\Delta\kappa$ and λ should be used. We chose to define the four parameters of the double exponential as a second order polynomial of $\Delta\kappa$ and λ . We immediately ran into trouble because the nine histograms contained differing numbers of events. This caused the parametrization of the function to $\Delta\kappa$ and λ to be inconsistent because the number of events in each histogram was not explicitly taken into account anywhere in the function. We solved this problem by normalizing the histograms to unit area. This eliminated any complication that might arise in the parametrization in the function because of numbers of events. It also sidestepped the difficult problem of normalizing the function because fitting to a normalized histogram produces an essentially normalized function. Around this

time, we also switched to a global chi-square fit for the parametrization of the functional form of the E_T spectrum. This global fit substituted the functions of $\Delta\kappa$ and λ into the four parameters of the double exponential function to fit for the parametrization of the function directly from the Monte Carlo data, rather than indirectly from the values of the double exponential functions four parameters after they had been fit to the data. This eliminated some of the inaccuracies resulting from the indirect two step process we had been using before. This global fit, combined with additional Monte Carlo data sets generated with the Baur and Fast Monte Carlos we had just installed at Argonne gave us a fair functional form for the photon E_T spectrum in terms of $\Delta\kappa$ and λ .

However, there were still indications that our parametrized fit could be improved. Though the double exponential function fit almost perfectly to the E_T spectrum for every set of Monte Carlo data we had, when we fit the version parametrized for $\Delta\kappa$ and λ to the Monte Carlo data, the fit was not as good, and the values of $\Delta\kappa$ and λ returned from these parametrized fits were only somewhat close in many cases to the values used to generate the Monte Carlo data. Also, the chi-square per degree of freedom of the global chi-square fit was extremely high—ten to twenty rather than the expected value of one that a good fit would produce. In fact, the more histograms used for the fit, the bigger the chi-square per degree of freedom became. In generating more Monte Carlo data files, we got more intermediate values for $\Delta\kappa$ and λ (i.e. values between 0 and ± 3 for $\Delta\kappa$ and between 0 and ± 2 for λ), which allowed us to get a better idea of how the double exponential parameters changed with respect to $\Delta\kappa$ and λ . These results suggested that we might want to use higher order functions of $\Delta\kappa$ and λ for our parametrization of the double exponential function. This conclusion is not certain, however, because we noticed a tendency for certain extremely low histogram bins to distort the fit. Eventually, we discovered a bug in the global chi-square program that was responsible for much of the inflated chi-square per degree of freedom. Removing this bug reduced the chi-square per degree of freedom for the best fit from about 16.5 to about 2.2. Still, this value is over twice the expected chi-square per degree of freedom for a good fit, suggesting that the parametrization could still use improvement.

The functional form we are currently using to fit the photon E_T spectrum is the sum of two exponentials, given as follows:

$$f(E_T) = \exp(P_1 + P_2 E_T) + \exp(P_3 + P_4 E_T)$$

To all indications, this is the best function to use over the range from 7 GeV to 100 GeV. Dr. Steve Errede has suggested that if the range were extended to 120 GeV, then the best fit would be a triple exponential, but this possibility has not yet been explored as we currently do not have any actual data in that range. The current parametrization of this function is accomplished by using a second degree polynomial in terms of $\Delta\kappa$ and λ for each of the four double exponential parameters. This parametrization is given below:

$$P_i(\Delta\kappa, \lambda) = a_i + b_i \Delta\kappa + c_i \lambda + d_i \Delta\kappa^2 + e_i \lambda^2 + f_i \Delta\kappa \lambda$$

We currently use the global chi-square fit to find the twenty-four parametrization coefficients $a_i, b_i, c_i, d_i, e_i, f_i; i = 1, 2, 3, 4$.

All the histograms used in these fits were normalized to unit area. This was achieved by dividing the contents of each bin of the histogram by the total contents of that histogram, including overflows, as well as by the bin width. Our histograms had a bin width of 1 GeV so this factor was irrelevant in the normalization. This normalization technique produced a histogram that had

an area of less than one over the range from 7 GeV to 100 GeV because each histogram had a significant amount of overflow above 100 GeV. This result is not undesirable, however, because the function should be normalized out to infinity. Normalizing the histogram out to the highest energy photon generated by the Monte Carlo comes as close as possible with this method to normalizing the histogram out to infinite E_T . Fitting the function to normalized histograms then solves both the problem of differing numbers of events among the various histograms as well as the complicated problem of normalizing the functional form of the E_T spectrum.

The global chi-square fit to obtain the twenty-four parametrization coefficients, mentioned above, uses a global chi-square statistic for all the histograms of the Monte Carlo data to fit the photon E_T spectrum parametrized in terms of $\Delta\kappa$ and λ to all the data simultaneously. This global chi-square statistic is merely a summation of the chi-squares of the fit of each histogram separately. The chi-square for each histogram is calculated by adding up the squares of the differences between the actual histogram contents of each bin and the values of the bins predicted by the functional form of the photon E_T spectrum. The contribution of each bin to the chi-square is weighted by the inverse of the approximate error for that bin. The chi-squares for each histogram are added together to form the global chi-square statistic which is then minimized with respect to the twenty-four parametrization coefficients to get the best fit. This has the benefit of ensuring that all the elements of the fit are coordinated so that the "best fit" fits every histogram with respect to every parameter in a consistent and equal manner.

Because of the nature of the histograms used for this fit, the approximate errors on the bins had to be calculated in a special way. The Baur and Fast Monte Carlo's produce an event sample that consists of weighted events. That means that when the events are put into a histogram, each must be multiplied by a weight that reflects the probability of actually getting such an event given the model of W plus photon production being used. For an unweighted histogram, bin errors are estimated by the square root of the contents of the bin. Approximate errors for bins in a weighted histogram, however, should be calculated as the square root of the sum of the squares of the event weights for the histogram bin. This was the method used to calculate approximate bin errors for both the fit of the double exponential to the photon E_T spectrum as well as for the global chi-square fit for the parametrization coefficients.

As mentioned earlier, the initial version of the global chi-square program contained an error. Such a mistake would not normally deserve much mention, but this error had an unusual effect on the fit. In fact, the best fit we have so far was obtained using the faulty version of the global chi-square fit with Monte Carlo data for the following $\Delta\kappa$, λ pairs: (-3, 0), (-2, -1), (-2, 1), (-1, -1), (-1, 0), (-1, 1), (0, -2), (0, -1), (0, 0), (0, 1), (0, 2), (1, -1), (1, 0), (1, 1), (2, -1), (2, 1), (3, 0). The bug in the program resulted from the counter variable that added up the chi-square for a single histogram not getting reset between histograms. Because of this, the contribution of the first histogram in the fit was added in as many times as there were histograms in the fit, while the second got added in one less time than the total number of histograms in the fit and so on. For example, in a seventeen histogram fit, the contribution of the first histogram in the fit would be added in seventeen times, the contribution of the second would be added in sixteen times, the third would be added in fifteen times, et cetera. The standard model was always the first histogram in the fit, so it always was most favored by the faulty global chi-square fitting program. The fact that this favoring of the standard model improved the fit suggests that more standard model data might be needed to get a good fit.

As more Monte Carlo data was generated and more photon E_T histograms became available,

another unexpected problem with event weights became evident. Certain isolated bins of the histograms, especially at high E_T 's, merely by chance contain only events of low weight. This causes these bins to have low approximate errors and, therefore, high weights in both the four parameter double exponential fit and the global chi-square fit. These low bins tend to pull the fit down and distort it from what would be expected given the overall shape of the E_T spectrum. This can be seen most clearly by comparing "mirror" sets of histograms, namely those histograms that come from corresponding $\Delta\kappa$, λ pairs in opposite quadrants of $\Delta\kappa$, λ space. These histograms appear close enough to visual inspection to be indistinguishable and such is usually the case for the functions that fit them as well. However, if one of these "mirror" pairs of the histograms has one or more unusually low bins while the other does not, the fit will be significantly different even though the histograms appear nearly identical. To attempt to correct for the effects of these unusually low bins in our global chi-square fit, we attempted to remove from the fit bins whose contribution to the chi-square was unusually high, with the reasoning that bins with unusually low event weight totals, even though they are favored by the fit, would have a high enough multiplier, because they are weighted in the chi-square by the inverse of the sum of the squares of their event weights, to keep their contribution to the chi-square of the histogram high. After comparing the distribution of the chi-square contributions of the individual bins to a chi-square distribution of one, the distribution the bins should have followed to some degree for a good fit, we decided to select out bins with a chi-square contribution of greater than eleven. This method for selecting out low bins was less than ideal however, because 93 bins total were selected out of the fit, far too many, while only nineteen of the forty-four bins we identified visually as being probably low enough to effect the fit were selected out. This method of the global chi-square fitting was less accurate than the method that did not select out any bins from the fit.

Presently, we are left with several options as to how to proceed on the fit. One approach that will almost inevitably need to be addressed in the fitting method at some point is to address adequately the problem of low bins. One method would be to develop a more efficient way of identifying and removing bins with unusually low event weights in an objective fashion. Monte Carlo results, as well as statistical principals, suggest the possibly better solution of generating larger Monte Carlo data samples. This would have the effect of "smoothing" out the histograms, making them less jagged and less likely to have a bin whose event weights were all low. The only drawback of this option is that would require a significant amount of disk space. Another method for attempting to improve the fit would be to investigate using higher order parametrizations with respect to $\Delta\kappa$ and λ for the fit. Although the results currently indicate that this is necessary, such results could merely be distorted because of the unusually low bins present in the fit. Finally, the approach we are currently favoring and have begun work on is to use a larger Monte Carlo data sample to fix the ai parametrization coefficients. We have already generated a standard model data sample that is six times larger than the original sample Mark Vondracek generated at the U of I. We will then use this sample to perform a four parameter fit for the double exponential function. Because the way this parametrization for $\Delta\kappa$ and λ is defined, for the Standard Model the values of the four parameters from the double exponential fit should be the same as the ai parametrization coefficients. The other terms in the parametrization become zero. This would both cause the fit to favor the Standard Model which is preferred, and would also simplify the global chi-square fit for parametrization coefficients by removing four parameters form the fit. This course of action will also give us an idea as to what size Monte Carlo data sample would be necessary to smooth out the problems of unusually low bins.

V Conclusions

To review, a double exponential function seems, with little doubt, to be a good fit to the E_T spectrum of the photon for W plus photon production over the range from 7 GeV to 100 GeV. We have had some success relating the four parameters of the double exponential fit to $\Delta\kappa$ and λ through a second order polynomial, but current results suggest that we may need to use a higher order polynomial to get the best fit. At this time, we are unable to distinguish between the photon E_T spectra of a $\Delta\kappa$, λ pair from its corresponding pair in the opposite quadrant (i.e. $[-3, 2]$ and $[3, -1]$). We do not foresee any way to distinguish these “mirror” pairs by only fitting the photon E_T spectra. Finally, our current best functional form for the photon E_T distribution from W plus photon production, parametrized for $\Delta\kappa$ and λ , was obtained with our faulty version of the global chi-square fitting program using the data for the following $\Delta\kappa$, λ pairs: $(-3, 0)$, $(-2, -1)$, $(-2, 1)$, $(-1, -1)$, $(-1, 0)$, $(-1, 1)$, $(0, -2)$, $(0, -1)$, $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, -1)$, $(1, 0)$, $(1, 1)$, $(2, -1)$, $(2, 1)$, $(3, 0)$.

As it stands now, the best way to proceed in finding this fit seems to be the following: First, complete the fit of the larger data sample for the standard model so that the a_i coefficients may be fixed and eliminated as parameters from the global chi-square fit. Next, if disk space permits, data samples at least twice as large as the ones we have now should be generated to try to eliminate the problems of low bins. If there is not enough disk space to do this, or if this technique does not solve the low bin problem, then we would need to develop an efficient and objective way to select out the bins in the fit low enough to distort our results. Then we should incorporate this selection method in our four parameter fit to the double exponential and get the four parameter values for all Monte Carlo data samples we currently have. These results should be studied to see if changes to our parametrization method for $\Delta\kappa$ and λ are warranted. Then, using any new parametrization, we should perform the global chi-square fit for parametrization coefficients, again using whatever method is necessary to correct for unusually low bins. If this approach yields a good enough fit we could then proceed to considerations of background and uncertainties in the fit.

In my personal opinion, this project is well on its way towards obtaining a functional form for the photon E_T spectrum for W plus photon production. In fact, the current best fit we have is a fairly decent one, and might even be acceptable for the range in which we need to fit for $\Delta\kappa$ and λ . However, we still have a number of options to explore for improving the fit, so there is no reason not to continue trying to improve the fit.

A Appendix: Program Notes

The following section details the various noncommercial programs and kumacs used during this project:

Kumac name: fit_gam_et.kumac

Description: This Kumac contains commands for PAW to create normalized histograms and fit them to a four parameter double exponential with correct errors.

Location: /Net/sgil/sd4b/cdf/rgwcdf/wgam_mc

Input: All wgam_*_hbook4_*.hb4 files containing ntuples and histograms to be fit.

Output: To screen (or metafile): for each histogram fit, this program creates a histogram with all the events in one bin, a histogram of the sum of the bin weights squared, and a normalized E_T histograms with the fit plotted on it.

Directions: Edit the kumac so that it opens the correct `wgam_*_hbook4_*.hb4` files to be fit. Make sure that the "dk" and "l" vectors contain values of $\Delta\kappa$ and λ that correspond to the values of the files to be fit in the same order that the files are assigned logical unit numbers. (Note: The "lunnum" variable should equal the number of the first logical unit used in the kumac.) Make sure the `savpar` and `saverr` vectors are the correct lengths if used. To run, enter PAW and type: `exec fit_gam_et.`

Kumac Name: `fit_gam_et_norm.kumac`

Description: This kumac contains PAW commands to fit normalized E_T histograms to a double exponential parametrized for $\Delta\kappa$ and λ . Note except where noted, this program is identical to the one above.

Location: same as above.

Programs involved: FORTRAN routine: `two_exp4.cdf` (not really cdf code) — contains function to be fit. (Note: This does not have to be compiled.)

Input: Same as above.

Output: Same as above except fit is in terms of two parameters, $\Delta\kappa$ and λ .

Directions: Same as above except type: `exec fit_gam_et_norm.`

Kumac Name: `get_contents.kumac`

Description: This kumac contains the PAW commands to normalize the E_T histograms and get correct errors. Also writes these values to disk.

Location: `/Net/sgil/sd4b/cdf/rgwcdf/wgam_mc`

Input: Same as 2 kumacs above.

Output: To screen (or metafile): Histogram with all contents in one bin and histogram of Bob.
To disk: `hist_*.dat` histogram contents and errors.

Directions: Edit kumac to make sure that the files to be normalized are opened. Make sure "dk" and "l" vectors contain $\Delta\kappa$ and λ values that correspond to the values of the files to be normalized in the order the files are assigned logical unit numbers. Make sure the kumac loops as many times as there are histograms. To run, enter PAW and type: `exec get_contents.`

Program Name: global_chisqr_fit.exe

Description: This program performs a global chi-square fit to the specified histograms to get out values for the parametrization coefficients. It also excludes bins from the fit whose contributions are greater than 11. (Note: This is written to use FORTRAN callable MINUIT.

Location: /Net/sgil/sd4b/cdf/rgwcdf/wgam_mc

Programs involved: Source code: (located in same area) glob_main.f, global_chisqr_fit.f

CERN libraries: packlib.a, kernlib.a

Makefile: (located in the same area) global_fit.mak

Input: all hist_*.dat files for the histograms to be in the global fit.
parameters.inp — starting parameter values.

Output: To screen (unless redirected to a file) MINUIT output and list of bins removed from fit as it progresses.

To disk: coefficients.dat — file of final coefficients.

chisqr_bins.dat — file of chi-square contributions of remaining bins in fit.

Directions: Edit global_chisqr_fit.f so that the "file" array contains the correct files to be included in the fit. Change parameter "num_hist" so that it reflects the correct number of histograms in the fit. Make sure the elements in "dk" and "l" arrays correspond to the correct values of $\Delta\kappa$ and λ used for the files in the "file" array in the same order. Edit glob_main.f so that the array "mult" has the appropriate number of elements (number of histograms, 100) and is initialized correctly (number of histograms * 100) in the data statement. When the program is ready, run as normal (optionally, redirect standard output to an output file [global.out]).

Note: An older version of this program in MINUIT data driven form that does not remove bins from the fit is located on the VAX in the following area: ANHEP3::\$9\$DRA1:[LANNON.MAXLIKE]gl. The source code is min_main.for and global_chisqr_fit.for. The program is made by the global_fit_make.com file and is run by the global_fit_run.com file. The input file is global_inp.dat and the output goes to global_out.lis. This version does require the hist_*.dat files to run.

Program Name: Two_exp4.cdf (Note — not actually CDF code.)

Description: This routine is used by fit_gam_et_norm.kumac to fit PAW histograms to a double exponential parametrized for $\Delta\kappa$ and λ .

Location: /Net/sgil/sd4b/cdf/rgwcdf/wgam_mc

Programs involved: Kumac: fit_gam_et_norm.kumac — calls two_exp4.cdf

Input: coefficients.dat — file with global fit coefficients

Output: Only to PAW and kumac — user never sees directly.

Directions: This routine is called automatically from `fit_gam_et_norm.kumac`. It can also be called with the PAW command:

`h/fit (histogram ID) TWO_EXP4.CDF ! 2 par`

NOTE: For some reason, `TWO_EXP4.CDF` must be typed in all capitals for the `h/fit` command to work even though the filename is lowercase.

Monte Carlos: (Note: these programs were written before this project by people not directly involved in the project.)

Program Name: `wgam_baur.exe` (Baur Monte Carlo)

Authors: Uli Baur (Florida State University) and Ed Berger (ANL).

Description: This Monte Carlo generates $W\gamma$ event data to be analyzed by the Fast Monte Carlo.

Location: `/Net/sgil/sd4b/cdf/rgwcdf/wgam_mc/exe`

Programs involved: Source Code: all *.f programs from `/usr/people/rgwcdf/wgam_mc/wgam_baur`.

Note: eventually, up to date source code should be moved to `/Net/sgil/sd4b/cdf/rgwcdf/wgam_mc/w`

Makefile: Stored on sgi8 as `usr/people/rgwcdf/wgam_mc/wgam_baur/makefile`

Input: by hand or from the appropriate `wpgam_*.inp` file.

Output: To screen (or redirected to a file): standard output.

To disk: `wpgam*.cmp` — binary data on events generated.

`wgam_*.sbg*.dat` — generation data file.

Program Name: `wgam_fmc.exe` (Fast Monte Carlo)

Authors: Steve Errede (University of Illinois).

Description: This Monte Carlo analyzes output from the Baur Monte Carlo, among others.

Location: `/Net/sgil/sd4b/cdf/rgwcdf/wgam_mc/exe`

Programs involved: Source Code: all *.f files on sgi8 in `/usr/people/rgwcdf/wgam_mc/wgam_fmc`.

Note: Source code should eventually be moved to `/Net/sgil/sd4b/cdf/rgwcdf/wgam_mc/fmc`.

Makefile: Stored on sgi8 as `/usr/people/rgwcdf/wgam_mc/wgam_fmc/makefile`

Input: By hand or from appropriate `wgam_*.inp_fmc` file, plus appropriate `wpgam*.cmp` and `wgam_*.sbg*.dat` files.

Output: To screen (unless redirected to a file): standard output.

To disk: `wgam*.hb4` — PAW ntuples and histograms `wgam_*.sbc*.dat`, `wgam_*.nevnts*.dat` — cuts data files.

Directions: To generate a Monte Carlo data set from scratch, edit the `wgam_mc.csh` shell script that runs the program, changing the file names to reflect the characteristics of this new run. Edit the `wpgam*.inp` file for the Baur Monte Carlo to change the random seed number, and

correctly indicate which anomalous couplings if any are required. Edit the `wgam*.inp.fmc` file to contain the correct input file names from the Baur Monte Carlo run. Run the `wgam.mc.csh` to generate the files. If the Fast Monte Carlo fails after the Baur Monte Carlo has completed, the Fast Monte Carlo may be restarted as follows: Delete any files generated by the Fast Monte Carlo before it crashed. Edit the `wgam.fmc.csh` shell script as the `wgam.mc.csh` script was originally edited. Run the `wgam.fmc.csh` shell script. Note: The input and output file-names listed above are the ones to which the shell script renames the files that the Baur and Fast Monte Carlo generates. To see the names the program gives to these files, examine the shell scripts to run the program. The shell scripts rename the files along the following general conventions:

`wgam prefix_particle_file type_structure function_couplings.scale factor.file extension.` e.g. `wpgam_ele_sbg_3.31_0.1_0.0_110.dat` For some files, naming conventions may vary slightly. The couplings part of the file name is replaced with "sm" for the Standard Model files.

MLM Fitting Program

Although we have not yet had to write the FORTRAN routine to fit the actual data to our final functional form using an unbinned MLM fit, several programs originally written to use an MLM fit to find the functional form could provide a good template for the final program. These programs are all on the VAX in the `ANHEP3::$9$DRA1:[LANNON.MAXLIKE]` area. Probably, the best one of these programs to use as a template is `gin.et_2exp2.for`, a FORTRAN routine meant to be used with data driven MINUIT. The main program for this routine is `2exp2_main.for`. The relevant *.com files are `2exp2_make.com` and `2exp2_run.com`. This program uses `2exp_inp.dat` and `wgm.et.dat` files for input.