Mon. Not. R. Astron. Soc. **000**, 1–3 (2015) Printed 17 February 2016 (MN LaTeX style file v2.2)

FermiFAST: A Fast Algorithm for Finding Point Sources in the Fermi Data Stream

Jeremy S. Heyl[1]*, Asha Ashathaman and Alistair Barton
[1] *Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, BC V6T 1Z1, Canada*

Accepted —. Received —; in original form —

ABSTRACT
This paper presents new and efficient algorithms for finding point sources in the photon event data stream from the Fermi Gamma-Ray Space Telescope.

Key words: methods: data analysis — methods: observational — techniques: image processing — astrometry

1 INTRODUCTION

2 THE PHOTON DATABASE

The key to the speed of this algorithm is the database that contains the position of the observed photons on the sky. Each photon is stored in a four-dimensional $k-d$ tree (Bentley 1975). We use the particularly memory efficient implementation of Lang (2009) (used in astrometry.net). The coordinates are actually stored as shorts instead of floats to save additional memory. The typical coordinates range from -1 to $+1$, so using shorts yields an angular precision of about six arcseconds much finer than that of the Fermi point-spread function (PSF). This memory efficient implementation allows us to store all of the photons detected by Fermi above 100 MeV and within a zenith angle of 100 degrees in memory simulataneously.

The first three dimensions contain the position of the photon on the celestial sphere as shown in the upper portion of Fig. 1. Storing the direction of the photon momentum in this manner removes the coordinate singularity of the spherical coordinates. Additionally it makes integrating over the celestial sphere straightforward because $\int d\Omega = \int 2\pi \delta p d\delta p$ where $\delta \mathbf{p} = \mathbf{p}' - \mathbf{p}$ is the three dimensional vector between two points on the sphere. The fourth coordinate that we denote by w depends on the point-spread function for the photon in question. In particular $w = \pm\sqrt{R_{\rm max}^2 - R_i^2}$ where R_i is the radius of the ninety-fifth percentile at the energy, entrance angle and front or back conversion for the photon. For convenience we use positive values of w for front-converted photons and negative values of w for back-converted photons. Furthermore, $R_{\rm max}$ is the ninety-fifth percentile for the photon with the poorest angular resolution. This is depicted in the lower panel of Fig. 1.

Once the $k-d$ tree is created, it is efficient to find all of the entries within the database within a given Cartesian distance of a particular point. In our case we query the database for all of the photons within a distance $R_{\rm max}$ of a particular point on the celestial sphere and use $w = 0$ for the fourth coordinate. The particular choice of w for the observed photons means that the query will yield all of the photons that

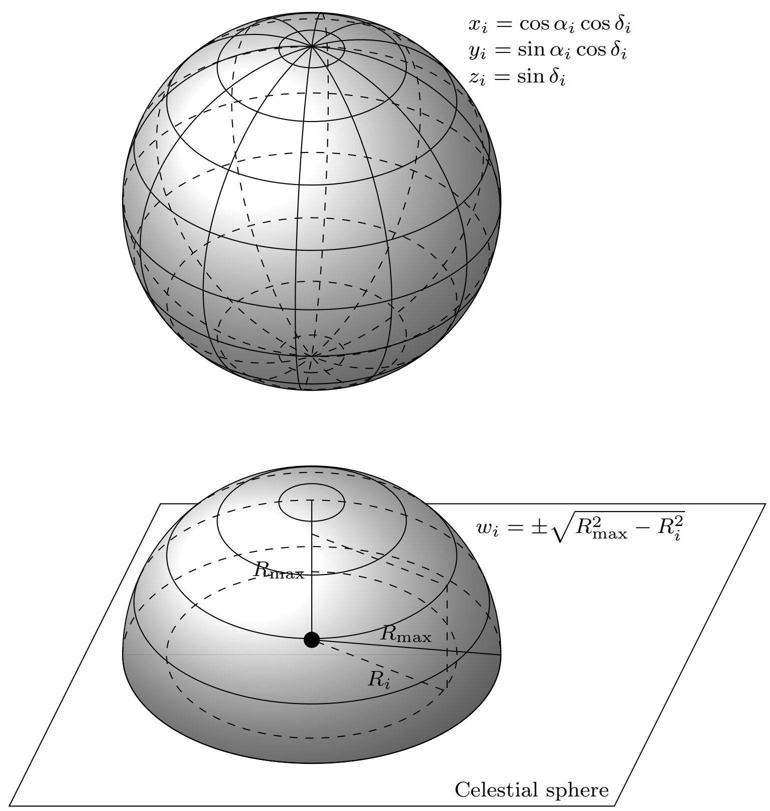

Figure 1. The location of a given photon event on the celestial sphere and in the additional dimension. R_i is the ninety-fifth percentile radius for the photon in question and $R_{\rm max}$ is the largest ninety-fifth percentile radius for the photons in the sample.

* Email: heyl@phas.ubc.ca; Canada Research Chair

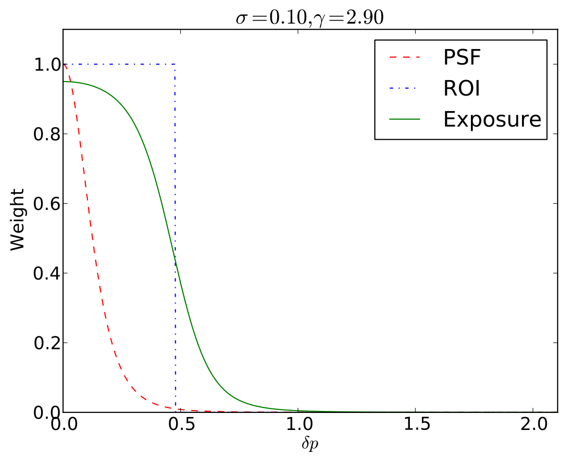

Figure 2. The Exposure Map

are within the ninety-fifth percentile of the PSF. In other words, if there is a point source located at that particular point the query will return on average 95% of the photons from that source. Of course, it will also return photons from the background and other nearby sources. This means that the region of interest for a particular prospective source is energy dependent. The form of the exposure map also is energy dependent as shown in Fig. 2. It peaks at 0.95 times the exposure time in the direction of the potential source and slowly drops to the ninety-fifth percentile of the PSF in radius and then drops according to the power-law of the tail component of the linear combinations of King function that are used to characterize the Fermi PSF.

Altough the construction of the tree is not done in parallel, the queries are performed in parallel using the tree held in shared memory. For example if one uses all the photons above 100 MeV from weeks 9 through 316 with a standard zenith angle cut (89,684,009 photons) requires one gigabyte to store the tree and about ten minutes to construct. The 200,000 location queries and likelihood calculations require 140,000 seconds (700 ms each), so the speed-up through parallisation can be dramatic. On the other hand, if one restricts to photons above 1 GeV (13,193,171), it only takes 90 seconds to construct the tree. At higher energies, it makes sense to make more location queries because the PSF is smaller. In this example 786,426 queries require 5,600 seconds of CPU time (7 ms each) or only six minutes on sixteen cores.

3 SOURCE LIKELIHOOD

Using the $k-d$ tree to determine the photons that would like within the 95% enclosure region of the point-spread function, we calculate several statistics of the observed photons to assess the likelihood of a source being at a particular position on the sky. We first determine two statistics whose distributions are known. First, if all of the photons within the region of interest (all photons that lie within the 95th-percentile cone of the potential source) indeed come from a uniform background, the ratio of the solid angle enclosed in a circle centred on the potential source and running through the observed to the total solid angle within the region of interest for that particular photon should be uniformly distributed between zero and one. We denote the mean of this ratio over the observed photons $\bar{r}^2$. This is a Bates distribution with mean of $1/2$ and variance of $1/(12N_{photons})$. Second, if all of the photons within the region of interest indeed come from a point source at the centre of the region of interest, the ratio of the percentile of a given photon within the cumulative PSF distribution for that particular photon to 0.95 should be uniformly distributed between zero and one. We denote this statistic by $\bar{f}$.

If we assume that the observed photons originate from a linear combination of these two possiblities, we can determine the ratio of the two contributions from these statistics. In particular the fraction of photons that come from the point source would be

$$A_f = \frac{\frac{1}{2} - \bar{r}^2}{\bar{f} - \bar{r}^2} \tag{1}$$

and we can estimate the significance of the value of A_f by

$$S(r^2) = \left(\bar{r}^2 - \frac{1}{2}\right)\sqrt{12N_{\rm photons}} \tag{2}$$

and the probabilty of getting a value of $S(r^2)$ larger than x by chance is

$$P\left[S(r^2) > x\right] = \frac{1}{2}{\rm erfc}\left(\frac{x}{\sqrt{2}}\right) \approx \exp\left(-\frac{x^2}{2}\right) \tag{3}$$

if we take the limit of many photons in the region of interest where the Bates distribution tends to the normal distribution.

These basic statistics are summarized in Tab. 1, and Tab. 2 lists these statistics for the ten most significant sources detected. These basic statistics really just compare two numbers about the distribution of the photons within the region of interest. We can use the detailed knowledge of the point spread function to develop a more comprehensive test of the distribution of photons. In particular we define the unbinned likelihood

$$\log L = \sum_{\rm photons} \log\left[A_{\rm PSF}\frac{{\rm PSF}_i\Omega_{{\rm max},i}}{0.95} + (1 - A_{\rm PSF})\right] \tag{4}$$

where we have dropped $N_{\rm pred}$ from the usual definition because we have defined the model in such a way that $N_{\rm pred} = N_{\rm photons}$ automatically and $dN_{\rm pred}/dA_{\rm PSF} = 0$. Furthermore, for $A_{\rm PSF} = 0$, $\log L = 0$ and because we are fitting a single variable, $\log L$ is distributed as a chi-squared distribution with a single degree of freedom and the probability of getting a value of $\log L$ larger than x by chance is

$$P(\log L > x) = \sqrt{\pi}{\rm erfc}\left(\sqrt{\frac{x}{2}}\right) \approx \exp\left(-\frac{x}{2}\right). \tag{5}$$

From Tab. 2 we can see that the values of A_f are similar to those of $A_{\rm PSF}$ at least for highly significant sources.

The first pass is to determine the value of $\log L$ on a HEALPix grid of potential sources. For example for the photons above 1 GeV we used `NSIDE` $= 256$ or 786,432 grid points. Of these 786,432 points, 18,425 have $P(\log L > x) < e^{-12.5} \approx 4\times10^{-6}$. Next we take this list of potentially significant sources and find the local maxima of $\log L$; this reduces the number to 1,226 unique potential sources. However, 426 have $A_{\rm PSF} < 0$, indicating not a source but a point-like

Table 1. Basic statistics calculated for the photon distribution around a potential source

Statistic	Symbol	Abbreviation	Definition
Number of photons	$N_{\rm photons}$	`N`	Number of photons that lie within the 95% percentile
Mean Solid Angle Ratio	$\bar{r}^2$	`MEANR2`	The mean of the ratio of solid angle enclosed between observed photon position and source location and the solid angle enclosed with the 95% percentile of the PSF
Mean Percentile Ratio	$\bar{f}$	`MEANFRAC`	The mean of the ratio of PSF percentile to 95%
Significance of `MEANR2`	$S(r^2)$	`SIGR2`	How many standard deviations is `MEANR2` away from 0.5
Significance of `MEANFRAC`	$S(f)$	`SIGFRAC`	How many standard deviations is `MEANFRAC` away from 0.5
Fraction of photons from point source	A_f	`AFRAC`	If one assumes that the photons come from the sum of a uniform background and a point source, what fraction come from the point source? $A_f = (0.5 - \bar{r}^2)/(\bar{f} - \bar{r}^2)$

deficit, leaving 800 sources. Each of these potential source positions is used more precisely to determine the local maxima of $\log L$ and get a more precise position for each source.

4 RESULTS

ACKNOWLEDGMENTS

The software used in this paper is available at `http://ubc-astrophysics.github.io`. We used the VizieR Service, the NASA ADS service, the Fermi Science Support Center, the astrometry.net $k-d$ tree library, the HEALPix and HEALPy libraries and arXiv.org. This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Foundation for Innovation, the British Columbia Knowledge Development Fund and the Bertha and Louis Weinstein Research Fund at the University of British Columbia.

REFERENCES

Bentley J. L., 1975, Commun. ACM, 18, 509
Lang D., 2009, PhD thesis, University of Toronto

Table 2. The results for the ten most significant peaks in the 1 GeV map

Source	RA	Dec	$N_{\rm photons}$	$\bar{r}^2$	$\bar{f}$	$S(r^2)$	$S(f)$	A_f	$TS_{\rm PSF}$	$A_{\rm PSF}$	$\ln P(TS)$
Vela	128.84	−45.18	172081	0.167	0.507	−478.9	9.7	0.98	156105.48	0.96	−78058.94
Geminga	98.48	17.77	90450	0.162	0.503	−352.3	3.4	0.99	84019.36	0.98	−42015.58
Crab	83.64	22.02	26547	0.189	0.528	−175.4	15.9	0.92	21557.26	0.90	−10783.85
PSR J1709-4429	257.42	−44.48	33105	0.258	0.589	−152.8	56.0	0.73	17373.25	0.70	−8691.73
PSR J1836+5925	279.06	59.43	16605	0.163	0.491	−150.3	−4.0	1.03	15372.34	0.97	−7691.22
3C 454.3	343.50	16.15	14021	0.169	0.511	−135.7	4.6	0.97	12422.41	0.96	−6216.15
PSR J0007.0+7303	1.76	73.05	13610	0.212	0.540	−116.2	16.3	0.88	9440.20	0.83	−4724.90
PSR J2021.5+4026	305.39	40.45	30942	0.324	0.665	−107.0	100.4	0.52	8803.10	0.51	−4406.32
PSR J1057-5226	164.49	−52.46	8222	0.227	0.560	−85.8	18.9	0.82	5301.47	0.79	−2655.25
PSR J2021+3651	305.26	36.85	22744	0.352	0.699	−77.5	103.8	0.43	4612.93	0.42	−2310.91

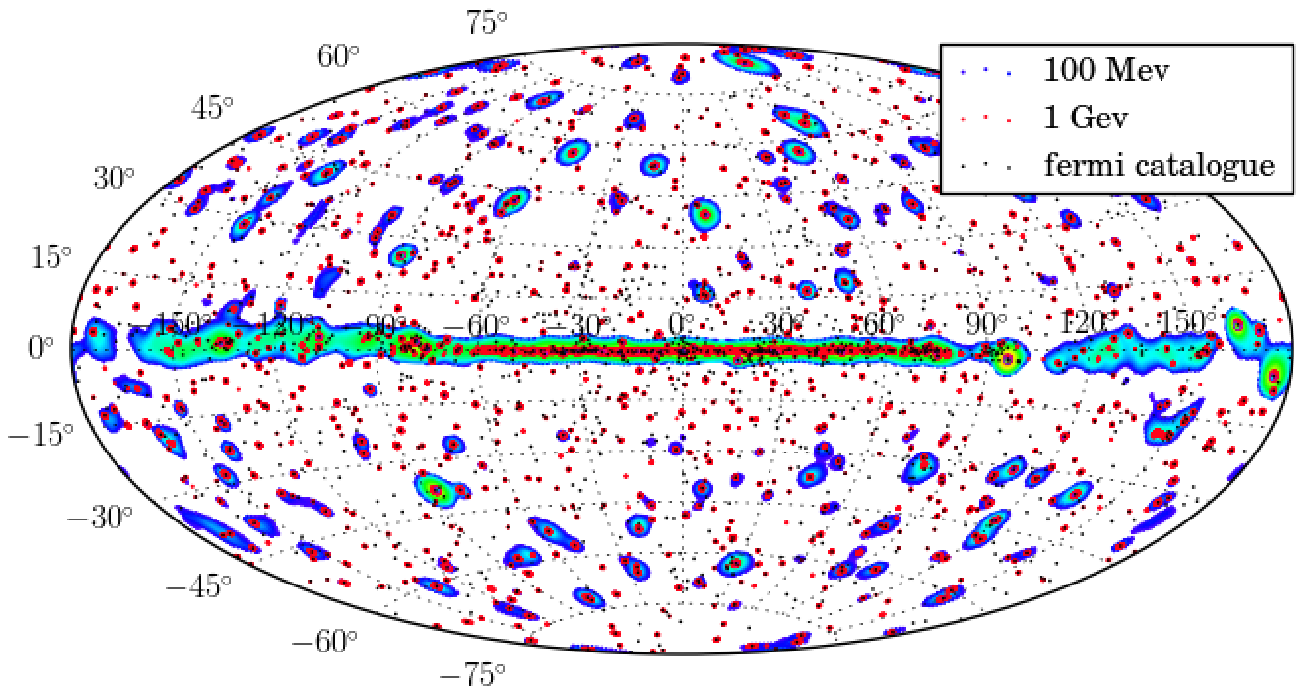

Figure 3. Likelihood map with regions where $TS > 25$, i.e. a significant detection of a source. Because the point spead function is broader at 100 MeV than that at 1 GeV, the TS map at 100 MeV is fuzzier.

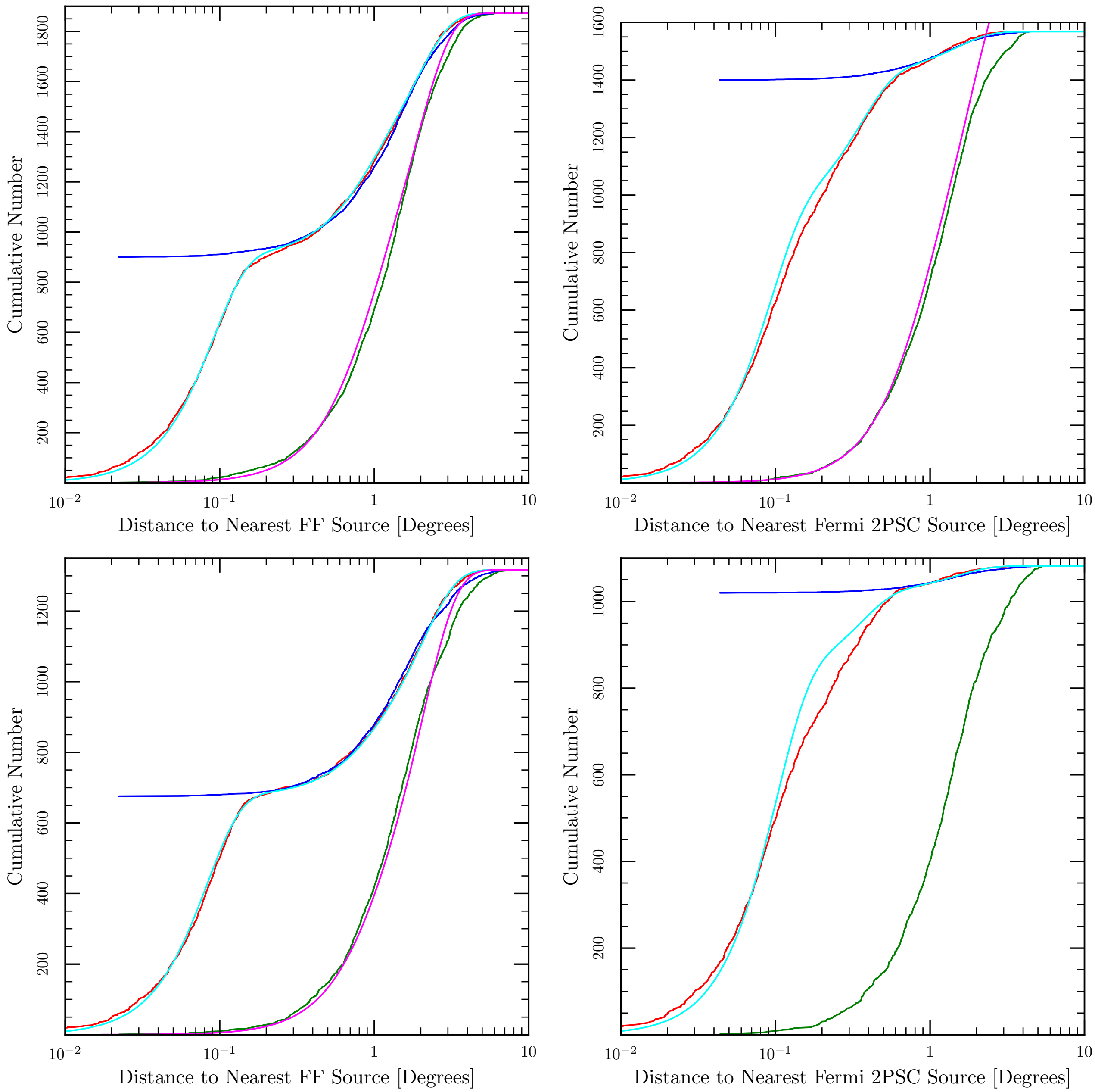

Figure 4. The upper panels give the results for all sources and the lower panels have $|b| > 10°$. Left: The distance from a Fermi 2PSC source to the nearest Fermi FAST source. This demonstrates that Fermi FAST finds about half of the Fermi 2PSC sources. Right: The distance from a Fermi FAST source to the nearest Fermi 2PSC source. This demonstrates that ninety percent of the Fermi FAST sources are associated with sources in the FERMI 2PSC. The red curve give the observed cumulative distribution of nearest distances. The green curves give the cumulative distribution that one would expect if there were no associated between the Fermi FAST and Fermi 2PSC sources. This is calculated by performing the same analysis as the red curves but with the Galactic coordinates inverted. The blue curve yields the detection rate in the case of the left panels and the false positive rate for the right panels. The cyan and magneta curves are Rayleigh distributions that are fit to the observed distributions. The typical positional error between associated 2PSC and FAST sources is four arcminutes.