LA-UR-25-20107
Approved for public release; distribution is unlimited.

Title: Maximum Likelihood Estimation: Some Basics

Author(s): Adams, Terry R.

Intended for: Report

Issued: 2025-01-06

Maximum Likelihood Estimation: Some Basics

Terry R. Adams

X-Computational Physics Division
Lagrangian Codes and Applications Group (XCP-1)
P.O. Box 1663, Los Alamos National Laboratory, Los Alamos, New Mexico 87545 USA
24 December 2024

Abstract

The maximum likelihood estimation is a general estimation procedure. It is often compared to estimation procedures like the ordinary least squares regression or generalized method of moments, to name a few. We discuss some basics about the maximum likelihood estimation, its advantages and disadvantages, and provide an example application to a gamma distribution function.

Keywords: Maximum Likelihood Estimation, Likelihood Function, Log-Likelihood Function, Gamma Distribution Function

Email address: `drbubba@lanl.gov` (Terry R. Adams)

Contents

1. Introduction

"The principle of maximum likelihood can be stated as follows: Find an estimate for (parameter) θ such that it maximizes the likelihood of observing those data that were actually observed."
- S. R. Eliason (Eliason, 1993, p. 7)

A simple description for the maximum likelihood estimation (MLE)[1]: a method of estimating the parameters of an *assumed* probability distribution function (PDF)[2] for a given set of observed data. The MLE is a general estimation procedure that is an alternative approach for estimating distribution parameters. The MLE is good for general probability distributions, continuous PDFs, non-negative PDFs, and works best when the PDF is known.

The following sections will be as concise as a reasonable explanation will allow. Our hope is that it will pique the readers interest and they will follow up with several references provided. In Sec. (2) we will describe the likelihood function (LF) and how to find MLE of a distribution's parameters. To, hopefully, solidify the MLE idea we will present an example of the MLE applied on a two-parameter gamma PDF (GDF2) in Sec. (3). Lastly, in Sec. (3.2), we will generate sets of random variables by sampling for them from a GDF2 distribution and, then we will apply the MLE method on these sets of generated data. The maximum likelihood estimators found should confirm the two originally known parameters used to generate the random variables.

2. Maximum Likelihood Estimation: Procedure

The outline and description in this following section are based on presentations in Eliason (1993), Shenton and Bowman (1997), and Martin (1971).

2.1. Likelihood Function (LF)

Given a simple and general probability density function, written as

$$f(x|\theta), \tag{1}$$

where x is the random variable (unknown) and θ is a parameter of the PDF (known). The PDF provides the probability of selecting an x from a given distribution defined by the parameter, θ. Mathematically the likelihood functions looks very similar to the PDF. The LF is a joint distribution function (JDF) which is constructed by a product of all the individual PDFs, Eq. (1), for independent random variables:

$$L(\theta) \equiv L(\theta|x_i) \equiv L(\theta|x_1, \cdots, x_N) = \prod_i^N f(x_i|\theta) \tag{2}$$

where $\prod$ is the product operator, the θ parameter of a distribution function (unknown), and the x_i are the observed data (known) which are independent and identically distributed (iid) random variables. The LF provides the likelihood of the distribution parameter occurring given the observed data, x_i. *In other words: we want to fit a distribution to find the optimal parameter (θ) of the distribution from the observed data,* $x_i = \{x_1, x_2, ..., x_N\}$.

[1] We will use MLE for maximum likelihood estimation or estimator when appropriate, depending on the context.

[2] The PDF is for continuous random variables or probability function (PF) for discrete random variables. We will refer to PDFs since we will being using a Gamma PDF in this report.

It is important to emphasize a difference between the PDF and LF:

- A PDF expresses the probability of observing the data *given* the underlying distribution parameters. It assumes that the data are *unknown*.
- The LF expresses the likelihood of the distribution parameter values occurring *given* the observed data. It assumes that the parameters are *unknown*.

2.2. Maximum Likelihood Estimation (MLE)

We want to find the maximum value of the LF or the maxiumum likelihood estimator. Several approaches might immediately come to mind. For instance, one could plot the LF as a function of the parameter, θ, and "pick-off" the maximum value in the plot. Or in more precise way, one might take the derivative of the LF as a function of θ and set that equal to zero then solve for the θ. Depending on the LF form, it is often messy to deal with the taking the derivative of a large product of functions.

Thus it is a common practice to take the natural logrithm of the likelihood function before taking the LF derivative. There are several reasons this is a desirable operation. Taking the logrithm often makes things easier to mathematical manipulate the LF, especially due to the property of logrithms where *the logrithm of a product is equal to the summation of the logrithm*. Another useful reason is numerically it is easier to handle values that are nominally close to zero, after taking the logrithm of those small values. It is important to appreciate that finding the maximum of the log-likelihood function (LLF) will give equalivent results as finding the maximum of the LF.

Applying the natural logrithm to Eq. (2), the LFF can be expressed like

$$\lambda(\theta) \equiv \ln[L(\theta)] = \ln\left[\prod_i^N f(x_i|\theta)\right] = \sum_i^N \ln[f(x_i|\theta)]\,. \tag{3}$$

Now that we have the LLF, we can find its maximum by taking the first derivative of the LLF with respect to θ and setting it to zero:

$$\partial_\theta[\lambda(\theta)] = \partial_\theta\left[\sum_i^N \ln[f(x_i|\theta)]\right] = 0, \tag{4}$$

where $\partial_\theta \equiv \frac{\partial}{\partial\theta}$. Now we can solve for a θ estimation in Eq. (4) which we will call a θ-estimator denoted by, $\hat{\theta}$. Unless an analytic solution is found, then a numerically iterative solution has to be employed.

To round out this section we introduce, without explanation[3], a commonly used form for the variance of the esimator, $\hat{\theta}$,

$$V(\hat{\theta}) = \left[-\partial_\theta^2[\lambda(\theta)]\right]^{-1}\Big|_{\theta=\hat{\theta}}. \tag{5}$$

This equation involves taking the second-derivative of the LLF and will be used later to find the variance of the estimator, $\hat{\theta}$.

[3]This equation and a derivation is found in Martin (1971, p. 78). Similar equations can be found in Shenton and Bowman (1997) and Eliason (1993)

2.3. MLE: Advantages

- If the model (like the chosen PDF) is correctly assumed, the MLE provides the *most efficient*[4] estimator.
- The MLE results in unbiased estimates for larger samples. MLE is sometimes often called an asymptotic estimator.
- The MLE provides a consistent, but flexible, approach which makes it appropriate for a large range of applications. This is especially true when assumptions of other models are inappropriate.

2.4. MLE: Disadvantages

- MLE estimates can be biased in small sample sizes.
- The MLE relies on the assumption of a model, and the derivation of the LF which is not easily found.
- MLE can be sensitive to the initial starting values for the numerical, iterative solve for the estimator.

3. Gamma Distribution Function Two-Parameters (GDF2)

In this section we will apply the MLE approach, from Sec. (2), to the two-parameter gamma distribution function (Eric et al., 2021) (Eliason, 1993, p. 46) (Shenton and Bowman, 1997, p. 149) which can be written as

$$f(x|\mu,\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} x^{(\nu-1)} \exp\left(-\frac{\nu}{\mu}x\right), \tag{6}$$

where x is the random variable and the two parameters of the function is the mean, μ, and ν which is inversely proportional to the variance and describes the width of the distribution. And, lastly, $\Gamma(\nu)$ is the gamma function (Eric et al., 2021). Figure 1 illustrates the Eq. (6) curves for various ν values. Notice that as the value of ν gets larger that the width of the curve's shape gets narrower, thus the variance gets smaller.

[4]Efficiency is one measure of an estimator. The *most efficient* estimator is consistent and whose sampling variance for large samples is less than that of any other such estimator. (Martin, 1971)

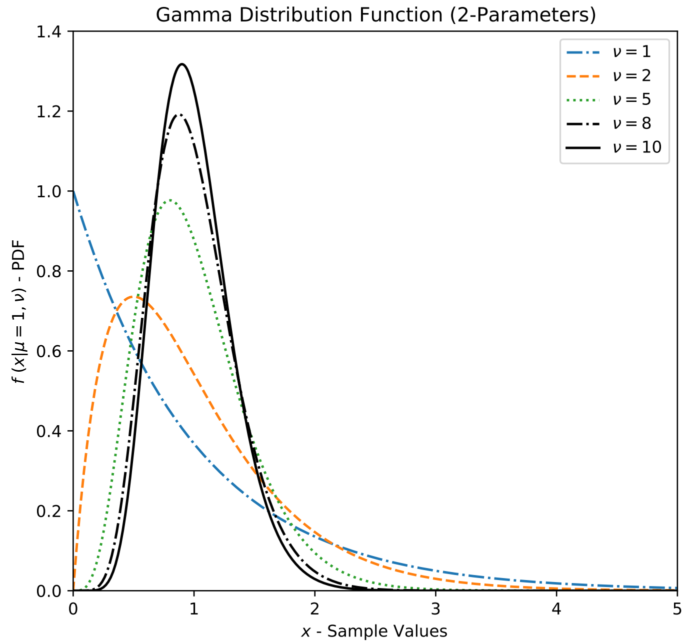

Figure 1: GDF2 curves for $\mu = 1$ and several demonstrative ν parameters. See $f(x|\mu, \nu)$ of Eq. (6).

3.1. Deriving the MLE of the GDF2

We can create a LF for our GDF2 of Eq. (6) using Eq. (2) which gives the form

$$L(\mu, \nu) \equiv L(\mu, \nu | x_1, \cdots, x_N) = \prod_{i}^{N} \left[\frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu} \right)^{\nu} x_i^{(\nu-1)} \exp\left(-\frac{\nu}{\mu} x_i \right) \right]. \tag{7}$$

With some algebraic manipulation we find the LF can be expressed as

$$\therefore L(\mu, \nu) = \left(\frac{1}{\Gamma(\nu)} \right)^{N} \cdot \left(\frac{\nu}{\mu} \right)^{N\nu} \cdot \prod_{i=1}^{N} \left(x_i^{(\nu-1)} \right) \cdot \exp\left(-\frac{\nu}{\mu} \sum_{i=1}^{N} x_i \right). \tag{8}$$

At this point, we might try to find the maximum of the Eq. (8). Though as described in Sec. (2.2) and Eq. (3), we will first find the the log-likelihood function (LLF) which takes the form

$$\lambda(\mu, \nu) \equiv \ln\left[L(\mu, \nu) \right] = -N \ln\left(\Gamma(\nu) \right) + \nu N \ln\left(\frac{\nu}{\mu} \right) + (\nu - 1) \sum_{i=1}^{N} \ln\left(x_i \right) - \frac{\nu}{\mu} \sum_{i=1}^{N} x_i \,. \tag{9}$$

The LLF equation (9) is consistent with a derivation by Eliason (1993, p. 46). For simple LLF functions we could just plot the function and "pick-off" the maximum from the curve plot. For example, shown in Fig. (2) is a plot of Eq. (9) as a function of $\hat{\nu}$.[5]

In general, it is more precise and useful to find the maximum of the LLF function of Eq. (9) by taking the derivative of this equation with respect to the targeted estimator parameter. We could start by trying to

[5] This curve was created by first generating random samples, x_i, from a GDF2, $f(\mu = 1, \nu = 2)$. Then assuming that $\hat{\mu} = 1$ and using N generated samples for x_i in Eq. (9), we generated Fig. (2).

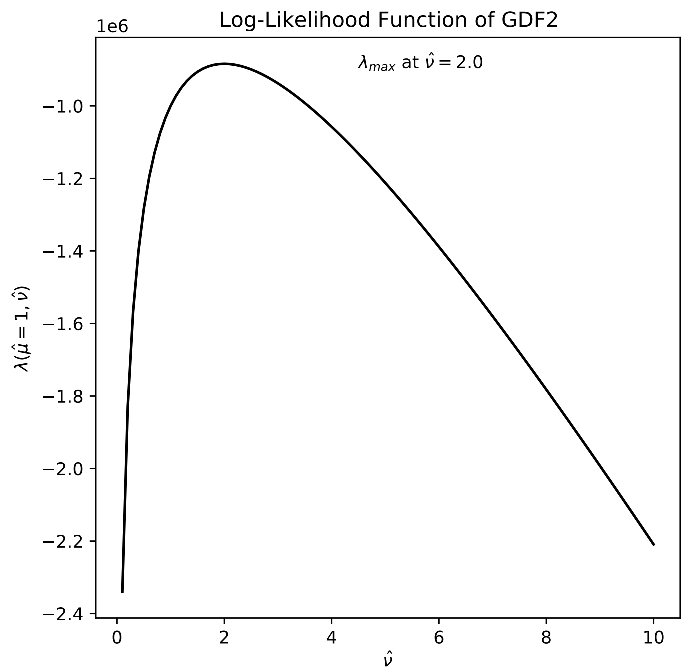

Figure 2: Plot of the LLF, $\lambda(\hat{\mu} = 1, \hat{\nu})$, in Eq.(9) as a function of $\hat{\mu}$. Notice that at $\hat{\nu} = 2$ the maximum is observed in this particular case, confirming this was the original ν used to generate the random variables, x_i.

estimate the mean, μ. Turns out this is a trivial case, it can be found analytically. By using Eqs. (4) and (5) we will find the form of $\hat{\mu}$ and $\hat{\sigma}_{\hat{\mu}}$, respectively, to be,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}(x_i - \hat{\mu})^2}, \tag{10}$$

where $\hat{\mu}$ is the mean (arithmetic) estimator and $\hat{\sigma}_{\hat{\mu}}$ is the standard deviation of the $\hat{\mu}$ estimator of the GDF2.

Since we can easily find an estimate of μ which we will call the μ-estimator and denote it by $\hat{\mu}$, then we will look for an estimate of ν by taking the derivative of Eq. (9) with respect to ν to find the MLE for this parameter

$$\partial_\nu \left[\lambda(\hat{\mu}, \nu)\right] = -N\psi(\nu) + N\ln(\nu) + \sum_{i=1}^{N} \ln\left(\frac{x_i}{\hat{\mu}}\right). \tag{11}$$

By setting this equation equal to zero

$$-\psi(\nu) + \ln(\nu) + \frac{1}{N}\sum_{i=1}^{N} \ln\left(\frac{x_i}{\hat{\mu}}\right) = 0, \tag{12}$$

where $\psi(\nu) \equiv \partial_\nu\left[\ln\left(\Gamma(\nu)\right)\right]$ which is known as the digamma function.[6] Then we can minimize this equation so as to find the estimator, $\hat{\nu}$, since at this point we know the set of x_i and the $\hat{\mu}$ estimator. Often a numerical

[6] As an aside, the summation term on the left-hand side in Eq. (12) can be rewritten to easily show that it is equal to the natural log of the geometric mean divided by the arithmetic mean.

and iterative method needs to be used to find this estimator. Note that Eq (12) is consistent with a derivation by Shenton and Bowman (1997, p. 150).

With knowledge of the estimator, $\hat{\nu}$, we now find its variance using $\hat{\nu}$ and substituting Eq. (9) in Eq. (5):

$$V(\hat{\nu}) = \left[-\partial_\nu^2 \left[\lambda(\hat{\mu}, \nu)\right]\right]^{-1} \Big|_{\nu=\hat{\nu}}. \tag{13}$$

We can leverage the knowledge of the first derivative of λ from Eq. (11) and rewrite the previous equation as

$$V(\hat{\nu}) = \left[N\partial_\nu\left[\psi(\nu)\right] - N\partial_\nu\left[\ln(\nu)\right] - \partial_\nu\left[\sum_{i=1}^{N} \ln\left(\frac{x_i}{\hat{\mu}}\right)\right]\right]^{-1} \Big|_{\nu=\hat{\nu}}. \tag{14}$$

This will reduce and provide a variance of the estimator, $\hat{\nu}$:

$$\therefore V(\hat{\nu}) = \left[N\left(\psi'(\hat{\nu}) - \frac{1}{\hat{\nu}}\right)\right]^{-1}, \tag{15}$$

where $\psi'(\nu) \equiv \partial_\nu \psi(\nu) = \partial_\nu^2 \left[\ln(\Gamma(\nu))\right]$ which is known as the trigamma function. Note that the deviation of $\hat{\nu}$ is, $\hat{\sigma}_{\hat{\nu}} = \sqrt{V(\hat{\nu})}$.

3.2. MLE applied to a GDF2 Sampled Data Sets

To get a feel for using the MLE of the GDF2 as seen in the previous section (Sec. (3.1)), we performed several simple computational experiments to see how well the MLE equations and our Python scripts perform. We will start by using a sampling routine for the GDF2 which is supplied in a Python module `random` and its two-parameter function `gammavariate` to sample N variants, $x_i = \{x_1, x_2, \ldots, x_N\}$, (see Appendix A).

In turn, once we sample for the iid values for x_i, we will use them as the experiments to find the two parameters of the GDF2 using the MLE approach in Sec. (3.1) which specifically means using Eq. (10) and an iterative process on Eq. (12) for $\hat{\mu}$ and $\hat{\nu}$, respectively.

The $\hat{\mu}$ estimator can be found analytically and is straight forward. For finding the $\hat{\nu}$ estimator we used the Python module `scipy` and in this case the `optimize` function using a Nelder-Mead method (Nelder and Mead, 1965; Gao and Han, 2012). For a snippet of the Python code, see Appendix B.

Our results are presented in Table 1 for parameter settings of $\mu = 403.43$ and $\nu = \{1, 5, 10\}$ used in the GDF2 routine to generate sample sizes of $N = \{500, 10^3, 2(10^3), 10^4, 10^6, 10^8\}$, for the random variables, x_i. Presented is the resulting $\hat{\mu}$ and $\hat{\nu}$ estimators fitted to a GDF2 using the previously generated variables x_i. The $\sigma_{\hat{\mu}}$ and $\sigma_{\hat{\nu}}$ are the standard deviations of the estimators $\hat{\mu}$ and $\hat{\nu}$, respectively.

There are several observations that can be seen. First look at the convergences of $\hat{\mu}$ and $\hat{\nu}$ for the same μ and ν as a function of N. The accuracy should be increasing as a function of $(1/\sqrt{N})$, as should the $\sigma_{\hat{\mu}}$ and $\sigma_{\hat{\nu}}$. They should converge to the original μ and ν that were used to generate the samples.

You also can notice that for a fixed N, say $N = 500$, that as ν changes from smaller to larger, there is less deviation from the mean in $\hat{\mu}$. This is consistent with a larger ν indicating that decreasing variance is seen in $\hat{\mu}$.

Also, the $\sigma_{\hat{\mu}}$ and $\sigma_{\hat{\nu}}$ results place the corresponding $\hat{\mu}$ and $\hat{\nu}$ within one to two standard deviations of expected results of μ and ν. This indicates that we seem to be sampling the GDF2 well and that the MLE of the GDF2 equations and Python source code are consistent and appear to be doing their job correctly.

Table 1: Example of MLE Applied to GDF2: First, the $\mu = 403.43$ and $\nu = \{1, 5, 10\}$ are the GDF2 parameters used to generate, N, random variables, x_i. Second, the $\hat{\mu}$ and $\hat{\nu}$ are the MLE estimators fitted to the GDF2 using the previously generated variables x_i. The $\sigma_{\hat{\mu}}$ and $\sigma_{\hat{\nu}}$ are the standard deviations of the estimators $\hat{\mu}$ and $\hat{\nu}$, respectively.

N	μ	$\hat{\mu}$	$\sigma_{\hat{\mu}}$	ν	$\hat{\nu}$	$\sigma_{\hat{\nu}}$
500	**403.43**	415.31	18.103	**1.0**	1.0454	0.0584
	”	405.35	8.226	**5.0**	4.8532	0.2970
	”	405.70	5.779	**10.0**	9.7246	0.6048
$1x10^{03}$	**403.43**	414.08	12.851	**1.0**	1.0262	0.0405
	”	399.31	5.761	**5.0**	4.7896	0.2072
	”	401.37	4.091	**10.0**	9.5357	0.4192
$2x10^{03}$	**403.43**	415.86	9.346	**1.0**	0.9881	0.0275
	”	399.56	4.061	**5.0**	4.8289	0.1477
	”	399.11	2.861	**10.0**	9.6245	0.2992
$1x10^{04}$	**403.43**	402.78	3.956	**1.0**	1.0185	0.0127
	”	402.45	1.800	**5.0**	5.0028	0.0685
	”	401.79	1.258	**10.0**	10.1280	0.1409
$1x10^{06}$	**403.43**	402.93	0.402	**1.0**	0.9997	0.0012
	”	403.64	0.181	**5.0**	4.9936	0.0068
	”	403.58	0.128	**10.0**	9.9805	0.0139
$1x10^{08}$	**403.43**	403.43	0.040	**1.0**	0.9999	0.0001
	”	403.42	0.018	**5.0**	5.0009	0.0007
	”	403.43	0.013	**10.0**	10.0000	0.0014

For some “eye-candy” for the reader, we have plotted simulations that can seen the figures in Appendix C for results similiar to Table 1, but using $\nu = 2$.

4. Summary

We have presented some basics about the maximum likelihood estimation method along with its application to the two-parameter gamma distribution function. We demonstrated their use through some simple python simulations. The results indicate that the equations and code give appropriate results.

For more about the likelihood function and maximum likelihood estimation then please see the attached references. Specifically for more details please see books by Eliason (1993); Martin (1971); Shenton and Bowman (1997) and for some shorter reading material like introductions and tutorials then please see articles by Cam (1990); Eric et al. (2021); Myung (2003).

Note: Not discussed in this paper is the MLE and small samples size (which is often defined as at, or less, than 100 samples). More on small sample sizes for the MLE can be found in Shenton and Bowman (1997) and Zacks and Even (1966).

5. Acknowledgements

Thanks to C. J. Solomon introducing me to the concept of *likelihood* and the method of *maximum likelihood* and the use of his original python code, and to Dru Renner for supporting the work that leveraged this approach. Lastly, much gratitude to Art Forster for being so generous with his time, and helping me clarify my fuzzy knowledge of probability and statistics.

References

Cam, L.L., 1990. "Maximum Likelihood: An Introduction". International Statistical Review / Revue Internationale de Statistique' 58, 153–171. URL: http://www.jstor.org/stable/1403464.

Eliason, S.R., 1993. Maximum Likelihood Estimation: Logic and Practice. International Educational and Professional Publisher, Newbury Park, California, USA.

Eric, U., Olusola, O.M.O., Eze, F.C., 2021. "A Study of Properties and Applications of Gamma Distribution". African Journal of Mathematics and Statistics Studies 4, 52–65. doi:10.52589/AJMSS-MR0DQ1DG.

Gao, F., Han, L., 2012. "Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters". Computational Optimization Applications 51, 259–277. URL: https://doi.org/10.1007/s10589-010-9329-3.

Martin, B.R., 1971. Statistics for Physicists. Academic Press, London, England.

Myung, I.J., 2003. "Tutorial on maximum likelihood estimation". Journal of Mathematical Psychology 47, 90–100. URL: https://www.sciencedirect.com/science/article/pii/S0022249602000287, doi:https://doi.org/10.1016/S0022-2496(02)00028-7.

Nelder, J.A., Mead, R., 1965. "A Simplex Method for Function Minimization". The Computer Journal 7, 308–313. URL: https://doi.org/10.1093/comjnl/7.4.308, doi:10.1093/comjnl/7.4.308.

Shenton, L.R., Bowman, K.O., 1997. Maximum Likelihood Estimation in Small Samples. MacMillan Publishing Co., Inc., New York, NY, USA.

Zacks, S., Even, M., 1966. "The Efficiencies in Small Samples of the Maximum Likelihood and Best Unbiased Estimators of Reliability Functions". Journal of the American Statistical Association 61, 1033–1051. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1966.10482193, doi:10.1080/01621459.1966.10482193.

Appendix A. Example of the GDF2 Sampling Code in Python

```
------------------------------------------------------------------------------------------
# A Simple Sampling Routine
# Random: https://docs.python.org/3/library/random.html
# Gamma Distribution:  https://en.wikipedia.org/wiki/Gamma_distribution
import random
import numpy

def setup_SamplingFromGDF2(mu, nu, N):
     initialize_seed = 12345679
     random.seed( initialize_seed )
     k     = nu
     theta = mu/k

     x = []
     for dummy in range(N):
         x.append( random.gammavariate(k, theta) )

     return numpy.array(x)
------------------------------------------------------------------------------------------
```

Appendix B. Python Code for Optimizing for ν

```
------------------------------------------------------------------------------------------
import numpy
import scipy.optimize as scipy_optimize
from scipy.special import polygamma

def mu_offset(x):
    mu     = numpy.average(x)
    offset = numpy.average( numpy.log(x/mu) )
    return  mu, offset

#  MLE method:
#     In parenthesis of the return function is equation (12) the derivative with
#        respect to nu of the log-likelihood of the two-parameter GDF set equal to 0.
#     The digamma function is polygamma(0,nu) function.
def objective_function(nu, offset):
    return (numpy.log(nu) - polygamma(0, nu) + offset)**2

def GammaDistribution_nu_fit(x):
    mu, offset = mu_offset(x)
    results    = scipy_optimize.minimize(objective_function, x0=[1.0], args=(offset),
                                         method='Nelder-Mead')
    nu         = results.x[0]
    return mu, nu
----------------------------------------------------------------------------------------
```

Appendix C. Plots for the MLE applied to GDF2 Sampled Data Sets

We have performed the simulations for $\nu = 2$ and the cases of $N = \{0.5 \cdot 10^3, 10^3, 10^4, 10^6, 10^8, 10^{10}\}$ shown in Figures (C.3, C.4, C.5, C.6, C.7, C.8), respectively, for the results of these simulations.

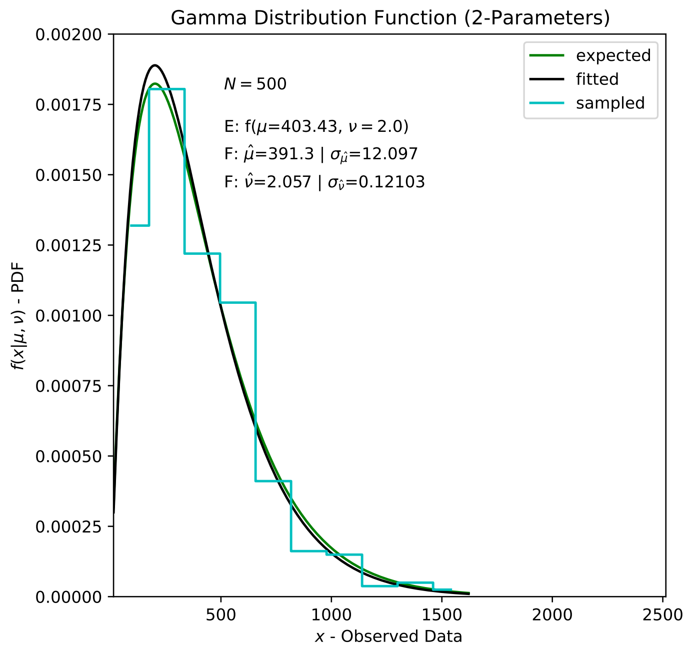

Figure C.3: Plot experiment for $N = 0.5 \cdot 10^3$ samples for x_i

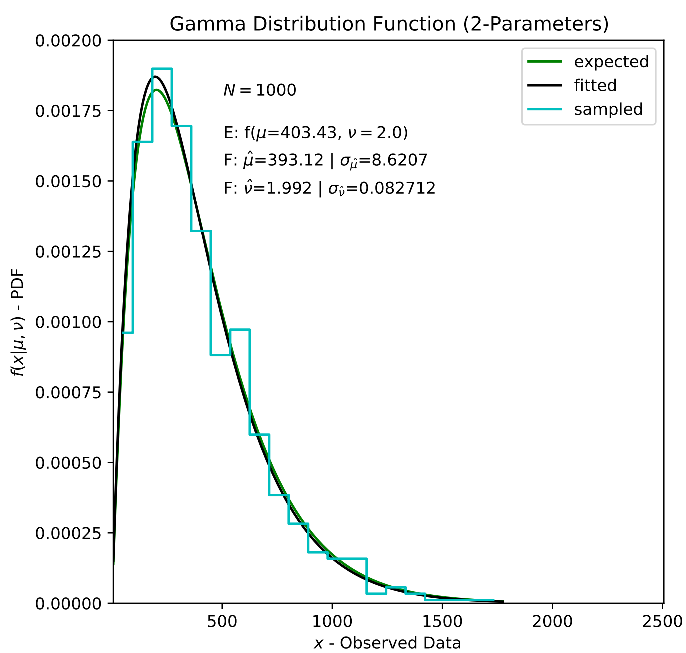

Figure C.4: Plot experiment $N = 10^3$ samples for x_i

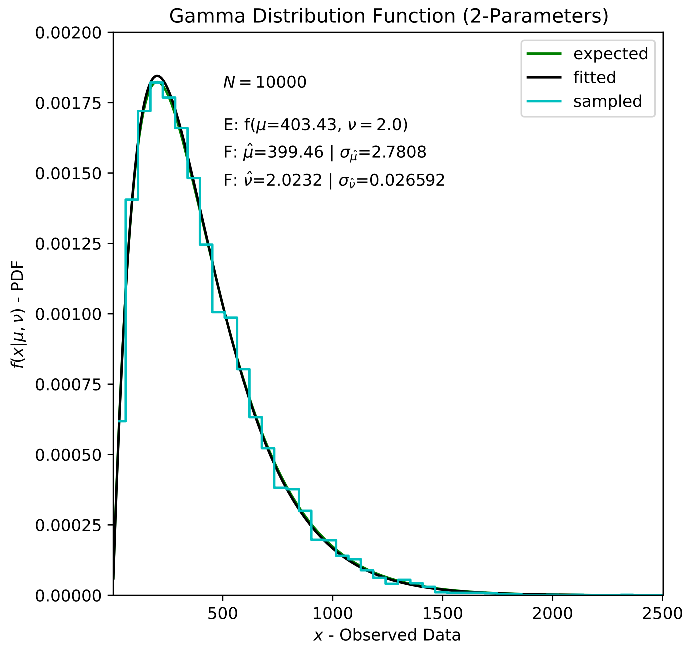

Figure C.5: Plot experiment $N = 10^4$ samples for x_i

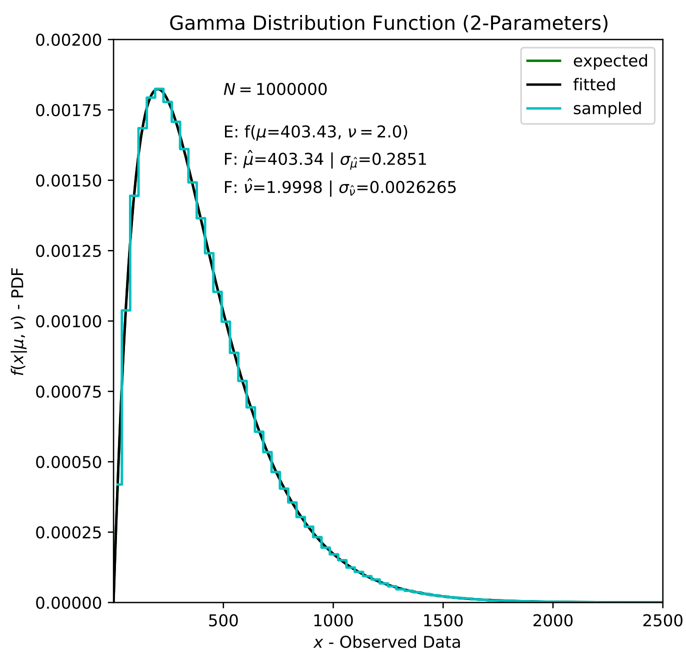

Figure C.6: Plot experiment $N = 10^6$ samples for x_i

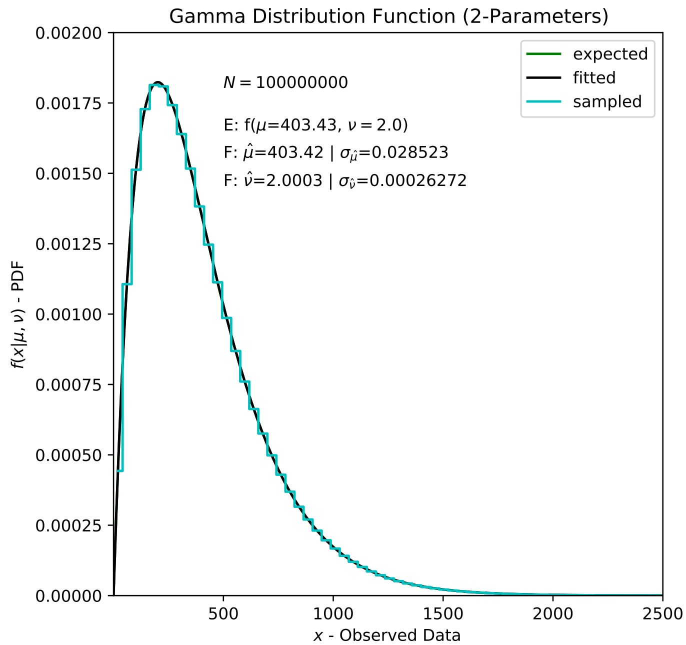

Figure C.7: Plot experiment $N = 10^8$ samples for x_i

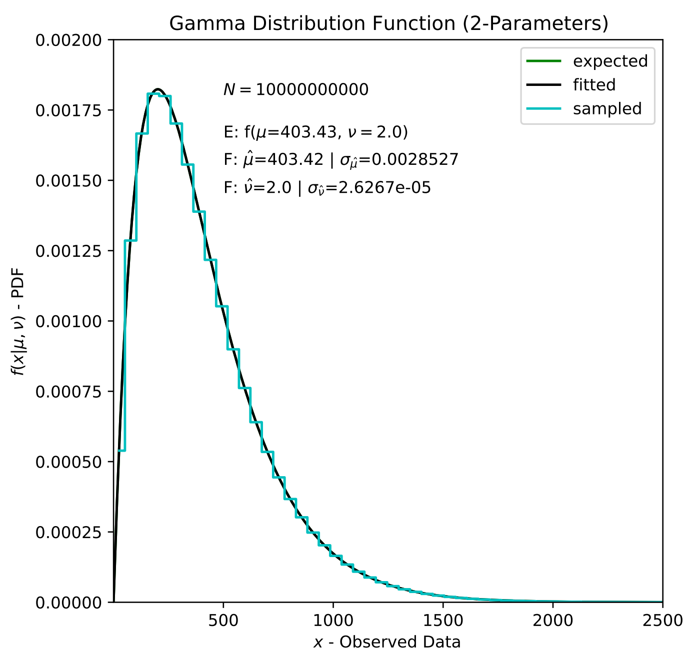

Figure C.8: Plot experiment $N = 10^{10}$ samples for x_i