This paper presents a parameterized version of the stochastic information gradient (SIG) algorithm, in which the error distribution is modeled by generalized Gaussian density (GGD), with location, shape, and dispersion parameters. Compared with the kernel-based SIG (SIG-Kernel) algorithm, the GGD-based SIG (SIG-GGD) algorithm does not involve kernel width selection. If the error is zero-mean, the SIG-GGD algorithm will become the least mean $p$-power (LMP) algorithm with adaptive order and variable step-size. Due to its well matched density estimation and automatic switching capability, the proposed algorithm is favorably in line with existing algorithms.

Keywords: Minimum error entropy criterion (MEE); stochastic information gradient (SIG) algorithm; generalized Gaussian density (GGD); least mean $p$-power (LMP).

1. Introduction

Stochastic gradient (SG) algorithm under error criteria has been extensively studied and widely used in the last decades.\textsuperscript{1,2} The least mean square (LMS) algorithm, which is based on the minimum mean-square error (MSE) criterion, is perhaps the most popular algorithm because of low computation requirements and optimality for Gaussian cases. However, in the presence of non-Gaussian
noises, the LMS algorithm will result in significant performance degradation. To improve the learning performance (degree of stability, convergence speed, misadjustment error, etc.) for non-Gaussian situations, many stochastic gradient algorithms under non-mean-square error (Non-MSE) criteria are studied. Among these algorithms, the least mean $p$-power (LMP) algorithm deserves special attention due to its ease of mathematical tractability and satisfactory performance. In scenarios of adaptive system training, the LMP algorithm with order $p$ is given by

$$W(k+1) = W(k) - \eta \frac{\partial}{\partial W} (|e(k)|^p)$$

$$= W(k) - \eta p |e(k)|^{p-1} \text{sign}(e(k)) \frac{\partial}{\partial W} e(k),$$  \hspace{1cm} (1)

where $W(k) = [w_1(k), w_2(k), \ldots, w_M(k)]^T$ is the weight vector of the adaptive model at $k$ iteration, $\eta > 0$ is the step-size, $e(k)$ denotes the prediction error, and the gradient $\frac{\partial}{\partial W} e(k)$ depends on the topology of the adaptive system under consideration. Well-known special cases of the LMP algorithm include the least absolute difference (LAD) algorithm ($p = 1$), LMS algorithm ($p = 2$), and the least mean fourth (LMF) algorithm ($p = 4$).

More recently, researchers propose the minimum error entropy (MEE) criterion as an information theoretic alternative to traditional error criteria in supervised adaptive system training. As shown in Fig. 1, under MEE criterion, the weight vector of the adaptive model $h(x,W)$ is adjusted online so that the error’s entropy $H(e(k))$ is minimized. The error’s entropy measures the average uncertainty contained in the error signal. Its minimization forces the error samples to concentrate. Numerical examples suggested that the MEE criterion could be able to achieve a better error distribution, especially for higher-order statistics in the adaptive system training. Under MEE criterion, the optimal weight vector $W^*$ is
given by

\[ W^* = \arg \min_{W \in \mathbb{R}^M} H(e) \]
\[ = \arg \min_{W \in \mathbb{R}^M} \left\{ -\int_{\mathbb{R}} p_e(e) \log p_e(e) \, de \right\} \]
\[ = \arg \min_{W \in \mathbb{R}^M} E_e(-\log p_e(e)), \] \hspace{1cm} (2)

where \( p_e(\cdot) \) denotes the PDF of the error, and \( E_e(\cdot) \) denotes expectation operator. The optimal weight vector can be searched by stochastic information gradient (SIG) algorithm.\(^9\)

\[ W(k+1) = W(k) - \eta \frac{\partial \{- \log p_e(e(k))\}}{\partial W}. \] \hspace{1cm} (3)

The key problem of SIG algorithm is how to estimate the PDF of error \( e(k) \). The nonparametric kernel approach has been widely used due to its smoothing characteristics.\(^{13}\) By kernel density estimation (KDE), the estimated error distribution is given by

\[ \hat{p}_e(e(k)) = \frac{1}{L} \sum_{i=k-L+1}^{k} K_h(e(k) - e(i)), \] \hspace{1cm} (4)

where \( K_h(\cdot) \) represents the kernel function, \( h \) is the kernel width, and \( L \) denotes the sliding window length. By kernel approach, one is often confronted with the problem of how to choose a suitable value of the kernel width. An inappropriate choice of width will significantly deteriorate the behavior of the algorithm.\(^{8,10}\) Though the effects of the kernel width on the shape of the performance surface and the eigenvalues of the Hessian at and around the optimal solution have been carefully investigated,\(^{10}\) the choice of the kernel width is still a difficult task. In Ref. 24, an adaptive kernel width is proposed, however, the computation of the kernel width is based on a gradient search algorithm which is even as complex as the adaptive algorithm itself. Hence, certain parameterized density estimation, which does not involve the choice of kernel width, might be more practical. Further, if some prior knowledge about the structure of the density is available, the parameterized density estimation may achieve a better accuracy than nonparametric ones. These facts motivate us to develop a parameterized SIG algorithm, in which the error’s distribution is modeled by a certain parametric family of probability density functions.

In this paper, the generalized Gaussian density (GGD)\(^{14}\) is used to match the error’s distribution. The reasons for the choice of GGD densities are at least twofold: (1) The GGD densities belong to a family of maximum entropy densities, which have simple yet flexible functional forms and could approximate a large number of different signals, comprising a wide range of statistical distributions;\(^{15-18}\) (2) The
GGD densities have only three parameters, i.e., location, shape and dispersion, which can be easily estimated from the sample data.\textsuperscript{17,18}

The paper is organized as follows: In Sec. 2, the generalized Gaussian distribution and its parametric estimation are introduced. In Sec. 3, the GGD-based SIG algorithm is derived, and the connections with LMP algorithm are discussed. In Sec. 4, a series of Monte Carlo simulation experiments are performed to demonstrate the performance of the new algorithm. Finally, concluding remarks are presented in Sec. 5.

2. Generalized Gaussian PDF and Its Parametric Estimation

The generalized Gaussian density (GGD) with location $\mu$ and standard deviation $\sigma > 0$ is defined as

$$p_X(x) = \frac{\alpha}{2\beta \Gamma(\frac{1}{\alpha})} \exp\left(-\left(\frac{|x - \mu|}{\beta}\right)^\alpha\right), \quad (5)$$

in which

$$\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}, \quad (6)$$

where $\Gamma(\cdot)$ denotes the Gamma function\textsuperscript{19} given by

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx, \quad z > 0. \quad (7)$$

In Eq. (5), $\alpha$ describes the exponential rate of decay, and hence the shape of PDF, while $\beta$ models the dispersion of the distribution. Usually, $\alpha$ is called the shape parameter and $\beta$ is referred to as the dispersion or scale parameter. The GGD densities include Gaussian ($\alpha = 2$) and Laplacian ($\alpha = 1$) distributions as special cases. Figure 2 shows the GGD distributions for several shape parameters with the same variance ($\mu = 0$). From Fig. 2, it is evident that smaller values of the shape parameter correspond to heavier tails and therefore to sharper distributions. In the limiting cases, as $\alpha \to \infty$, the GGD becomes close to the uniform distribution, whereas for $\alpha \to 0^+$, it approaches an impulse function ($\delta$-distribution). The GGD densities are the maximum entropy (Maxent) distributions, which satisfy the following constrained optimization (see Appendix A):

$$\begin{aligned}
\max_p \left\{ -\int_{-\infty}^{+\infty} p(x) \log p(x) dx \right\} \\
\text{s.t.} \int_{-\infty}^{+\infty} |x - \mu|^\alpha p(x) dx = m_\alpha = \beta^\alpha/\alpha,
\end{aligned} \quad (8)$$

where $m_\alpha$ denotes the $\alpha$-order absolute central moment.
According to Jaynes’s maximum entropy principle (MEP),\textsuperscript{20} the Maxent distribution is “uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters.” Therefore, the GGD distribution would be the most unbiased distribution that agrees with a certain absolute central moment constraint.

Until now, a number of methods have been proposed for the parametric estimation of a GGD distribution. The state-of-the-art can be found in Refs. 14 and 18. Here we only introduce the moment-matching estimators (MMEs). A complete statistical description for the GGD-modeled distributions, by simply matching the underlying moments of the data set with those of the assumed distribution, is first obtained by Mallat.\textsuperscript{21} For a GGD distribution, the $r$-order absolute central moment is calculated as

$$m_r = E[|X - \mu|^r] = \int_{-\infty}^{+\infty} |x - \mu|^r p_X(x) dx = \beta^r \Gamma\left(\frac{r + 1}{\alpha}\right) \left/ \Gamma\left(\frac{1}{\alpha}\right)\right..$$  \hfill (9)

The ratio of the mean absolute deviation ($m_1$) to standard deviation ($\sqrt{m_2}$) is

$$M(\alpha) = \frac{m_1}{\sqrt{m_2}} = \frac{\Gamma\left(\frac{2}{\alpha}\right)}{\sqrt{\Gamma\left(\frac{3}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}\right)}}.\quad (10)$$

Fig. 2. Generalized Gaussian distribution for different shape parameters with unit variance.
which is a steadily increasing function of shape parameter $\alpha$. So an estimate for $\alpha$ can be expressed as

$$\hat{\alpha} = M^{-1}\left(\frac{\hat{m}_1}{\sqrt{\hat{m}_2}}\right),$$

(11)

where $M^{-1}(\cdot)$ denotes the inverse function of $M(\cdot)$, $\hat{m}_1$ and $\hat{m}_2$ represent the sample mean absolute deviation and the sample standard deviation, respectively. Given independent, identically distributed (i.i.d.) random samples $\{x_1, x_2, \ldots, x_N\}$, we have

\[
\begin{align*}
\hat{\mu} &= \frac{1}{N} \sum_{i=1}^{N} x_i, \\
\hat{m}_1 &= \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{\mu}|, \\
\hat{m}_2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2.
\end{align*}
\]

(12)

The shape parameter $\alpha$ can also be expressed in terms of the kurtosis of the GGD distribution. A similar ratio, which is equal to the reciprocal of the square root of kurtosis of the GGD distribution is given by

$$K(\alpha) = \frac{m_2}{\sqrt{m_4}} = \frac{\Gamma\left(\frac{3}{\alpha}\right)}{\sqrt{\Gamma\left(\frac{3}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}\right)}}.$$  

(13)

Therefore, another estimate of $\alpha$ could be

$$\hat{\alpha} = K^{-1}\left(\frac{\hat{m}_2}{\sqrt{\hat{m}_4}}\right),$$

(14)

where $K^{-1}(\cdot)$ denotes the inverse function of $K(\cdot)$, $\hat{m}_4$ is the sample fourth-order central moment, that is,

$$\hat{m}_4 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^4.$$  

(15)

To estimate the shape parameter $\alpha$, we have to solve the inverse functions $M^{-1}(\cdot)$ or $K^{-1}(\cdot)$, whose curves are depicted in Fig. 3. As both the inverse functions have no explicit forms, some interpolation method should be used in practical applications.

Plugging the estimate of $\alpha$ into (6) and substituting $\sigma$ for $\sqrt{\hat{m}_2}$, we may get the estimate of the dispersion parameter:

$$\hat{\beta} = \sqrt{\hat{m}_2 \frac{\Gamma(1/\hat{\alpha})}{\Gamma(3/\hat{\alpha})}}.$$  

(16)
3. GGD-Based Stochastic Information Gradient Algorithm

Using GGD density to approximate the error’s distribution, we have

\[ \hat{p}_e(e(k)) = \frac{\hat{\alpha}_k}{2\hat{\beta}_k \Gamma\left(\frac{1}{\alpha_k}\right)} \exp\left(-\frac{|e(k) - \hat{\mu}_k|^{\alpha_k}}{\hat{\beta}_k}\right), \] (17)

where \( \hat{\mu}_k, \hat{\alpha}_k, \) and \( \hat{\beta}_k \) denote the estimated location, shape, and dispersion parameters at \( k \) iteration. By moment-matching estimators, say the kurtosis-based approach, we have

\[
\begin{cases}
\hat{\mu}_k = \frac{1}{L} \sum_{i=k-L+1}^{k} e(i), \\
\hat{\alpha}_k = K^{-1} \left( \frac{1}{L} \sum_{i=k-L+1}^{k} (e(i) - \hat{\mu}_k)^2 \right) \\
\hat{\beta}_k = \left[ \frac{1}{L} \sum_{i=k-L+1}^{k} (e(i) - \hat{\mu}_k)^2 \right]^{\frac{1}{2}} \times \frac{\Gamma\left(\frac{1}{\alpha_k}\right)}{\Gamma\left(\frac{3}{\alpha_k}\right)}.
\end{cases}
\] (18)

Fig. 3. Curves of the inverse function \( M^{-1}(\cdot) \) and \( K^{-1}(\cdot) \).

3. GGD-Based Stochastic Information Gradient Algorithm

Using GGD density to approximate the error’s distribution, we have

\[ \hat{p}_e(e(k)) = \frac{\hat{\alpha}_k}{2\hat{\beta}_k \Gamma\left(\frac{1}{\alpha_k}\right)} \exp\left(-\frac{|e(k) - \hat{\mu}_k|^{\alpha_k}}{\hat{\beta}_k}\right), \] (17)

where \( \hat{\mu}_k, \hat{\alpha}_k, \) and \( \hat{\beta}_k \) denote the estimated location, shape, and dispersion parameters at \( k \) iteration. By moment-matching estimators, say the kurtosis-based approach, we have

\[
\begin{cases}
\hat{\mu}_k = \frac{1}{L} \sum_{i=k-L+1}^{k} e(i), \\
\hat{\alpha}_k = K^{-1} \left( \frac{1}{L} \sum_{i=k-L+1}^{k} (e(i) - \hat{\mu}_k)^2 \right) \\
\hat{\beta}_k = \left[ \frac{1}{L} \sum_{i=k-L+1}^{k} (e(i) - \hat{\mu}_k)^2 \right]^{\frac{1}{2}} \times \frac{\Gamma\left(\frac{1}{\alpha_k}\right)}{\Gamma\left(\frac{3}{\alpha_k}\right)}.
\end{cases}
\] (18)
where $L$ denotes the sliding window length. Thus, the GGD-based SIG algorithm can be derived as

$$W(k+1) = W(k) - \eta \frac{\partial}{\partial W} \left\{ -\log \frac{\hat{\alpha}_k}{2\beta_k \Gamma(\frac{1}{\alpha_k})} \exp \left( -\left( \frac{|e(k) - \hat{\mu}_k|}{\beta_k} \right)^{\alpha_k} \right) \right\}$$

$$= W(k) - \eta \frac{\partial}{\partial W} \left( \frac{|e(k) - \hat{\mu}_k|}{\beta_k} \right)^{\alpha_k}$$

$$= W(k) - \eta \frac{\partial}{\partial W} \left( \frac{|e(k) - \hat{\mu}_k|}{\beta_k} \right)^{\alpha_k} \text{sign}(e(k) - \hat{\mu}_k) \frac{\partial}{\partial W} e(k). \quad (19)$$

In this paper, to make a distinction, we denote “SIG-Kernel” and “SIG-GGD” the SIG algorithms based on kernel approach and GGD densities, respectively. Compared with the SIG-Kernel algorithm, the SIG-GGD just needs to estimate the three parameters of GGD density, without resorting to the choice of kernel width and the calculation of kernel function.

Comparing (1) and (19), we find that when $\hat{\mu}_k \approx 0$, the SIG-GGD algorithm can be regarded as the LMP algorithm with adaptive order $\alpha_k$ and variable step-size $\eta/\beta_k^{\alpha_k}$. In fact, under certain conditions, the SIG-GGD algorithm will converge to a specific LMP algorithm. Consider the finite impulse response (FIR) system identification case, in which the plant and the adaptive model are both FIR filters with the same order. In this case, the error signal $e(k)$ is given by

$$e(k) = W^T X(k) + n(k) - W^T(k) X(k)$$

$$= V^T(k) X(k) + n(k), \quad (20)$$

where $W^* = [w_{1}^*, w_{2}^*, \ldots, w_{M}^*]^T$ is the weight vector of unknown system, $n(k)$ is the interference noise, $X(k) = [x(k), x(k-1), \ldots, x(k-M+1)]^T$ is the input vector, and $V(k) = W^* - W(k)$ is the weight error vector. Then, near the convergence, we have $W(k) \approx W^*$, and hence

$$e(k) = (W^* - W(k))^T X(k) + n(k) \approx n(k). \quad (21)$$

If $\{n(k)\}$ is zero-mean white noise, we have $\hat{\mu}_k \approx 0$ and

$$\hat{\alpha}_k \approx K^{-1} \left( \frac{1}{L} \sum_{i=k-L+1}^{k} n^2(i) \right)$$

$$\hat{\beta}_k \approx \sqrt{\frac{1}{L} \sum_{i=k-L+1}^{k} n^2(i) \times \frac{\Gamma \left( \frac{1}{\alpha_k} \right)}{\Gamma \left( \frac{1}{\alpha_k} \right)}}. \quad (22)$$
Assuming $L$ as large enough, by the limit theorems of probability,

\[
\begin{align*}
\frac{1}{L} \sum_{i=k-L+1}^{k} n^2(i) & \to E[n^2(k)], \\
\frac{1}{L} \sum_{i=k-L+1}^{k} n^4(i) & \to E[n^4(k)].
\end{align*}
\]

(23)

Thus, the estimated parameters $\hat{\alpha}_k$ and $\hat{\beta}_k$ shall approach certain constants, i.e., the SIG-GGD algorithm shall converge to a certain LMP algorithm with fixed order and step size. For example, if $n(k)$ is zero-mean Gaussian distributed (with shape parameter $\alpha = 2$), we have $\hat{\alpha}_k \approx 2$ near the convergence, and consequently, the SIG-GGD algorithm converges to the LMS algorithm. Similarly, for Laplace noises, the SIG-GGD algorithm shall converge to the LAD algorithm. In Ref. 4, it has been shown that under slow adaptation, the LMS and LAD algorithms are, respectively, the optimum algorithms for the Gaussian and Laplace interfering noises. We may therefore conclude that the SIG-GGD algorithm has the ability to adjust its parameter pair $(\hat{\alpha}_k, \hat{\beta}_k)$ so as to switch to a certain optimum algorithm. This automatic switching property will be verified in the following section through a set of Monte Carlo simulations.

4. Simulation Results

To demonstrate the performance of the SIG-GGD algorithm, we perform a series of Monte-Carlo simulations for the FIR channel identification, in comparison with the SIG-Kernel, LAD, LMS, and LMF algorithms. Let the weight vector of unknown channel be $W^* = [0.1, 0.3, 0.5, 0.3, 0.1]^T$, and the input signal be white Gaussian noise with unit power. For the SIG-Kernel algorithm, the Gaussian kernel is used, and the kernel width is chosen according to the well-known Silverman’s rule of thumb.\(^{13}\) For the SIG-GGD algorithm, the kurtosis-based moment-match estimator is used (see Ref. 18). To solve the inverse function $K^{-1}(\cdot)$, we adopt the basic linear interpolation method, with step 0.05. In addition, to avoid “very large gradient,” we set the upper bound of $\hat{\alpha}_k$ equal to 4, that is, if $\hat{\alpha}_k > 4$, we set $\hat{\alpha}_k = 4$. In the experiments below, we focus on four special distributions of the disturbance noise, whose density functions are depicted in Fig. 4.

Figures 5–8 illustrate the convergence curves of the mean-square deviation (MSD) $E[V^T(k)V(k)]$ averaged over 100 independent Monte-Carlo runs for each disturbance noise. Figure 9 shows the averaged parameter $\alpha_k$ for Laplace, Gaussian, and Uniform noises. To compare the statistical results, we also summarize in Table 1 the sample mean and standard deviation of the weight $w_3$ ($w_3^* = 0.5$).

From the figures and table, we have the following observations:

1. The performances of the LAD, LMS, and LMF algorithms depend crucially on the distribution of the disturbance noise. The three algorithms may achieve the
Fig. 4. Four probability density functions of the disturbance noise.

Fig. 5. Convergence curves of mean square deviation (Laplace noise case).
Fig. 6. Convergence curves of mean square deviation (Gaussian noise case).

Fig. 7. Convergence curves of mean square deviation (Uniform noise case).
smallest misadjustment for a certain noise distribution (e.g., the LMS algorithm for Gaussian noise); however, for other noise distributions, their performance will deteriorate dramatically.

(2) The SIG-GGD and SIG-Kernel algorithms are both robust to the noise distribution. In the case of symmetric and unimodal noise (e.g., Laplace, Gaussian, Uniform), the SIG-GGD algorithm may achieve a smaller misadjustment than the SIG-Kernel algorithm. Though in the case of nonsymmetric and non-unimodal noise (e.g., MixNorm), the performance of SIG-GGD algorithm may be not as good as the SIG-Kernel algorithm, it is still better than most of the LMP algorithms.

(3) Near the convergence, the parameter $\alpha_k$ converges approximately to 1, 2, and 4 (note that $\hat{\alpha}_k$ is restricted to $\hat{\alpha}_k \leq 4$ artificially) when disturbed by Laplace,
Gaussian, and Uniform noise, which implies the SIG-GGD algorithm will switch approximately to the LAD, LMS, and LMF algorithm for, the Laplace, Gaussian, and Uniform noise, respectively. This agrees with the previous discussion in Sec. 3.

5. Conclusions

We have presented in this paper a new stochastic information gradient (SIG) algorithm, in which the error’s distribution is modeled by the generalized Gaussian density (GGD). Different from the kernel-based SIG algorithm, the GGD-based algorithm does not resort to the choice of kernel width, and just relies on estimating the location, shape, and dispersion parameters during the adaptation. The relationships between the proposed algorithm and the least mean $p$-power (LMP)
algorithm are investigated. This new algorithm has the ability to adjust the parameters so as to switch approximately to a certain optimal LMP algorithm. Simulation results suggest that, for the case in which the noise has a symmetric and unimodal distribution, the SIG-GGD algorithm may outperform the SIG-Kernel algorithm. In the case of nonsymmetric and nonunimodal disturbance, the performance of the SIG-GGD algorithm may be not as good as the SIG-Kernel algorithm, however, it still outperforms most of the LMP algorithms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 60904054), and was partially supported by NSF Grant ECCS 0856441, NSF IIS 0964197, and ONR N00014-10-1-0375.

Appendix A

To solve the optimization of (8), we create an unconstrained expression for the entropy using Lagrange multipliers:

\[
L = - \int_{-\infty}^{+\infty} p(x) \log p(x) dx - (\gamma_0 - 1) \left\{ \int_{-\infty}^{+\infty} p(x) dx - 1 \right\} \\
- \gamma_1 \left\{ \int_{-\infty}^{+\infty} |x - \mu|^\alpha p(x) dx - \beta^\alpha/\alpha \right\}.
\]

(A.1)

Using calculus of variations, we maximize \( L \) with respect to \( p(x) \), that is,

\[
\frac{\partial}{\partial p} \left\{ -p \log p - (\gamma_0 - 1)p - \gamma_1 |x - \mu|^\alpha \right\} = 0
\]

\[
\Rightarrow - \log p - 1 - (\gamma_0 - 1) - \gamma_1 |x - \mu|^\alpha = 0
\]

\[
\Rightarrow p = \exp\{ -\gamma_0 - \gamma_1 |x - \mu|^\alpha \},
\]

where \( \gamma_0 \) and \( \gamma_1 \) are determined by

\[
\begin{cases}
\int_{-\infty}^{+\infty} \exp\{ -\gamma_0 - \gamma_1 |x - \mu|^\alpha \} dx = 1, \\
\int_{-\infty}^{+\infty} |x - \mu|^\alpha \exp\{ -\gamma_0 - \gamma_1 |x - \mu|^\alpha \} dx = \beta^\alpha/\alpha.
\end{cases}
\]

(A.3)

Hence, it is easy to derive \( \gamma_0 \) and \( \gamma_1 \) as follows:

\[
\begin{align*}
\gamma_0 &= \log\{(2\beta\Gamma(1/\alpha))/\alpha\}, \\
\gamma_1 &= 1/\beta^\alpha.
\end{align*}
\]

(A.4)

Combining (A.2) and (A.4) yields the GGD distribution. This completes the proof.
References

