Closed-Form Expressions of Some Stochastic Adapting Equations for Non-Linear Adaptive Activation Function Neurons

Simone Fiori*

October 21, 2002

Abstract

In recent works we introduced non-linear adaptive activation function (FAN) artificial neuron models which learn their activation functions in an unsupervised way by information-theoretic adapting rules. We also applied networks of these neurons to some blind signal processing problems, such as independent component analysis and blind deconvolution. The aim of the present paper is to study some fundamental aspects of FAN units learning by investigating the properties of the associated learning differential equation systems.

Keywords: Adaptive activation function neurons; entropy optimization; unsupervised learning; non-linear dynamical systems; phase-portrait.

1 Introduction

As mentioned in early and recent works [2, 3], the use that the brain can make of large amounts of information that come unaccompanied by any explicit instruction may be explained in terms of 'unsupervised learning'. Auto-organization, or unsupervised learning, denotes the activity of spontaneous learning to perform sensory processing, perception and inference.

According to everyday observations, the external environment exciting our senses has a big deal of statistical structure, that is, it is perceived as a large

*The author is with the Faculty of Engineering of the Perugia University – Loc. Pentima bassa, 21, I-05100 Terni (Italy). Email: sfr@unipg.it

set of spatio-temporal patterns to be identified and classified by suitable neural mechanisms [4]. As a consequence, the richness of information impinging our senses could be identified in the statistical structure of the associated patterns, whose suitable description may be given by using the concepts owing to information theory.

These observations have inspired a large amount of research work in the engineering community, aimed at reproducing the basic mechanisms and functions of brain into signal/data processing algorithms capable of extracting information from raw data and untreated signals. As specific examples, adaptive-activation-function and high-order neuron models have been proven worth studying in recent years (see e.g. [15, 21]).

The analysis of the behavior of non-linear adaptive (non-stationary) artificial neurons is a challenging research topic in the neural network field, which may require analyzing non-linear differential equations of neuron's parameters. The problem complicates when the external excitations of the neuron are not deterministic but stochastic and the aim is to find a statistical description of the system's response and of neuronal system features [24]. The formal techniques available in the literature for studying such systems benefited from cross-fertilization among the scientific communities related to information theory, signal processing and circuit theory.

Recently, several researchers have focused their attention on this class of stochastic non-linear adapting theories, with applications to blind separation of sources by the independent component analysis [5, 6, 27], linear estimation and time-series prediction [23], probability density estimation/shaping [22, 27, 28], non-linear system modeling [8, 25], self-organizing classification [26], adaptive control [16, 29] and blind system deconvolution [5, 12].

As a special case, the aim of the different techniques applied to univariate probability density function estimation by single non-linear unit is to find a non-linear transformation of an input random process (with unknown statistics) that optimizes a statistical criterion function of non-linear unit's response. Then, the found transformation approximates the cumulative distribution function of the input random process, and the first derivative of the transformation approximates the probability density function of the input signal [23, 26].

In recent contributions, we presented some results related to the use of flexible activation function non-linear neuronal units, termed FANs, trained in an

stochastic way by means of an entropy-based criterion: In [11, 13] we proposed some general structures and adapting frameworks for FAN non-linear unit, while papers [10, 12] have been devoted to the application of these artificial neuronal units to blind signal processing tasks, such as blind source separation by the independent component analysis and blind system deconvolution; in these works we also compared the proposed structures to other flexible topologies known in the scientific literature, as e.g. the mixture-of-kernel one, showing that the new approach may exhibit better estimation/approximation ability at a lower complexity burden. Also, the recent paper [13] surveys the problem of PDF approximation and estimation from incomplete data, and discusses the suitable FAN neural models in relationship to classical probability density function estimation techniques.

The aim of our preceding work was to introduce the new artificial FAN structures and adapting theories and to assess their features through numerical experiments on real-world data; however, due to the strong non-linearity of the involved learning equations, we did not present any theoretical considerations about their mathematical structure and properties. In the present brief paper we recall the basic adapting formulas and present the closed-form expressions of them for some special cases; our main goal is to discuss their features in an analytical way, in order to gain a deeper insight into the behavior of the non-linear differential equations governing information-theoretic FAN non-linear neuron adapting and to better explain the previous numerical results. The present paper also extends the results on probability-density-function matching neurons recently presented in the note [14].

2 Learning maximum entropy-gap activation function in a FAN unit

In this paper, the following input-output description for a FAN non-linear unit is assumed:

$$y = \Psi(x; \mathbf{a}) = \mathrm{sgm}[\varphi(x; \mathbf{a})] \ , \tag{1}$$

where $\mathrm{sgm}(\cdot)$ is a saturating function, bounded above and below, continuous and strictly increasing; $\varphi(x; \mathbf{a})$ is a monotonic polynomial in x depending upon parameters in $\mathbf{a} = [a_0 \ a_1 \ a_2 \ \cdots \ a_n]^T$. The variable x denotes the external

excitation to the neuron, while the variable y denotes the neuron's response. If x is supposed to be a stationary continuous-time random process $x = x(t) \in \mathcal{X}$ with *probability density function* (PDF) $p_x(x)$, then y will be a random process $y = y(t) \in \mathcal{Y}$ too, with a PDF denoted as $p_y(y; \mathbf{a})$. Note that the neuron's response distribution has a functional dependence on neuron's activation function structural parameters a_k. Also, in order for the learning theory to be consistent, we need to make the hypothesis that $p_x(x)$ is endowed with bounded moments up to some order.

The differential entropy of the random processes $x(t)$ and $y(t)$ defines as:

$$H_x \stackrel{\text{def}}{=} -\int_{\mathcal{X}} p_x(\xi) \log p_x(\xi) d\xi \ , \ H_y(\mathbf{a}) \stackrel{\text{def}}{=} -\int_{\mathcal{Y}} p_y(\eta; \mathbf{a}) \log p_y(\eta; \mathbf{a}) d\eta \ . \tag{2}$$

The excitation entropy and the response entropy relate by means of the PDF transformation formula to give the relationship:

$$H_y(\mathbf{a}) = H_x + \int_{\mathcal{X}} p_x(\xi) \log[\Psi'(\xi; \mathbf{a})] d\xi \ , \tag{3}$$

which describes how the neuron modifies the information content of the excitation into the entropy content pertaining to the response signal.

Our aim is to make the non-linear unit adapt the set of parameters in $\mathbf{a}$ so that its non-linear transference approximates the cumulative distribution function of the input signal. Note that estimating the cumulative distribution function, the probability density function, the score function or the quantile function, is fully equivalent [10, 11, 23, 27]; however, the cumulative distribution function is the only one, among these, exhibiting bounded, saturating shape, thus it is worth identifying it as the sigmoidal activation function of a neural unit.

A valid criterion for cumulative distribution function approximation is the *entropy gap* between the input excitation and the non-linear unit's response $G_h \stackrel{\text{def}}{=} H_y - H_x$, that maximizes when, and only when, the non-linear unit's response is uniformly distributed, which, in turn, happens when, and only when, the non-linear unit's transfer function approaches the cumulative distribution of the input signal; this is, of course, because the set $\mathcal{Y}$ is by definition a limited interval due to the boundness of function $\text{sgm}(\cdot)$. In the present case the entropy gap writes:

$$G_h(\mathbf{a}) = \int_{\mathcal{X}} p_x(\xi) \log\{\text{sgm}'[\varphi(\xi; \mathbf{a})] \varphi'(\xi; \mathbf{a})\} d\xi \ . \tag{4}$$

Now we wish to find a vector of parameters $\mathbf{a}$, hence a configuration of the FAN non-linear unit, that maximizes the entropy-gap. To this aim, a set of continuous-time adapting equations derived by the gradient steepest ascent method can be used here. Such equations write in the form:

$$\frac{d\mathbf{a}}{dt} = \mathbf{R}(\mathbf{a}) \frac{\partial G_h}{\partial \mathbf{a}} \ , \tag{5}$$

where $\mathbf{R}(\mathbf{a})$ is a positive-definite gain matrix. In the present case the above gradient particularizes into:

$$\frac{\partial G_h}{\partial \mathbf{a}} = \int_{\mathcal{X}} p_x(\xi) \left[\frac{\mathrm{sgm}''[\varphi(\xi;\mathbf{a})]}{\mathrm{sgm}'[\varphi(\xi;\mathbf{a})]} \frac{\partial \varphi(\xi;\mathbf{a})}{\partial \mathbf{a}} + \frac{1}{\varphi'(\xi;\mathbf{a})} \frac{\partial \varphi'(\xi;\mathbf{a})}{\partial \mathbf{a}} \right] d\xi \ . \tag{6}$$

An interesting theoretical question is the relationship between the entropy gap and the mutual information criterion. For the sake of generality, let us consider a noisy artificial neuron model described by $y = \Psi(x) + n$, where n denotes a random noise having arbitrary statistics and being statistically independent by the stimulus x. The mutual information between neuron's input and output signals is given by $I_{x,y} = H_x + H_y - H_{x,y}$, where $H_{x,y}$ denotes the joint neuron's input/output differential entropy. Because of the independence of the noise by the stimulus and of the monotonicity of neuron's transference, it is readily noted that $H_{x,y} = H_x + H_n$, thus $I_{x,y} = H_y - H_n$ and, consequently, $G_h = I_{x,y} - H_x + H_n$. This expression tells that in a noisy neuron the entropy gap can be uselessly grown by adding stronger noise, while the noiseless neuron (1) that maximizes its input-output entropy gap also maximizes its capability of embodying input-output mutual information. These considerations are closely related to the ones emerged in classical works by Linsker *et al.* [17, 18].

Another important question concerns the provability of convergence of the general learning equations (5). This question can be addressed by the help of a Lyapunov function $V(\mathbf{a}(t))$ that enjoys $V(\mathbf{a}(t)) \geq 0$ and $V'(\mathbf{a}(t)) \leq 0$, with $V'(\mathbf{a}^\star) = 0$, with $\mathbf{a}^\star$ denoting the equilibrium point of the learning system. In system theory it is proven that the existence of a Lyapunov function for a dynamical system's equilibrium state ensures its asymptotic stability. In the present case it is worth assuming $V(\mathbf{a}) = -G_h(\mathbf{a}) \geq 0$. It is then readily found that:

$$\frac{dV(\mathbf{a})}{dt} = \left(\frac{\partial V(\mathbf{a})}{\partial \mathbf{a}} \right)^T \left(\frac{d\mathbf{a}}{dt} \right) = - \left(\frac{\partial G_h(\mathbf{a})}{\partial \mathbf{a}} \right)^T \mathbf{R}(\mathbf{a}) \left(\frac{\partial G_h(\mathbf{a})}{\partial \mathbf{a}} \right) \ .$$

The quantity in the right-hand side is always non-positive because of the assumed positive-definiteness of the gain matrix $\mathbf{R}$, proving that $-G_h$ is a Lyapunov function for the FAN learning system.

Because of the strong non-linearity of the adapting equations (5)-(6), it is difficult to carry out any theoretical consideration about its mathematical structure and properties. In the following we aim at discussing some special cases that give rise to tractable mathematics.

3 Closed-form adapting differential equations

In the following we describe some cases-study of FAN adapting. It is necessary to make a preliminary choice about the non-linear structure of the neural unit. Here we consider the topology:

$$\begin{aligned} \mathrm{sgm}(u) &= \frac{1}{2} + \frac{1}{2}\mathrm{erf}(u) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^u \exp(-v^2) dv \ , & (7) \\ \varphi(x;\mathbf{a}) &= a_0 + a_1^2 x + a_3^2 x^3 + a_5^2 x^5 + \cdots + a_{2r+1}^2 x^{2r+1} \ , & (8) \end{aligned}$$

with $2r+1$ being the polynomial order. The particular expression of the polynomial (8) ensures $\varphi'(x;\mathbf{a}) \geq 0$; it is also worth noting that the chosen function $\mathrm{sgm}(u)$ is such that $\mathrm{sgm}''(u)/\mathrm{sgm}'(u) = -2u$, which remarkably simplifies the integral in (6). Also, in this case the entropy gap writes:

$$G_h(\mathbf{a}) = \int_{\mathcal{X}} p_x(x) \log \left\{ \frac{1}{\sqrt{\pi}} \exp\left[-\varphi^2(\xi;\mathbf{a})\right] \varphi'(\xi;\mathbf{a}) \right\} d\xi \ . \quad (9)$$

The polynomial form (8) makes the non-linear unit resemble a pseudo-Volterra filter (without temporal dynamics), which possesses the known ability to capture some high-order features of the input signal [20]. Also, an interesting parallelism arises in the field of theoretical neurobiology where pseudo-polynomial and pseudo-Volterra-type neural transfer functions have been recently introduced to explain the non-linear behavior of the biological neurons (see e.g. the *clusteron* model [21] and the recent review in [15]; also see the review paper [9] focusing on the usefulness of connectionist models in information processing sciences).

It is also interesting to note that the non-linear function (7) has been chosen as a saturating (sigmoidal) transference which is highly appealing from a biological point of view, as many physiological phenomena have been described through such kind of non-linearities [19].

3.1 First case-study

The first case-study concerns the analytical investigation of the estimation of a 'bi-constant' probability density function defined as:

$$p_x(x) = \begin{cases} \frac{3}{4} & , \quad -1 \leq x < -\frac{1}{3} \ , \\ 0 & , \quad -\frac{1}{3} \leq x < \frac{1}{3} \ , \\ \frac{3}{4} & , \quad \frac{1}{3} \leq x < 1 \ . \end{cases} \tag{10}$$

This has been chosen as an interesting case-study and numerically tackled in [11]. We consider this PDF again here because it is piece-wise constant, therefore it makes the integrals involved in equations (6) and (9) be tractable.

As neuron's activation function polynomial part, we choose a third-order polynomial, that is $\varphi(x;\mathbf{a}) = a_0 + a_1^2 x + a_3^2 x^3$. As the PDF to be approximated is symmetrical about $x = 0$, we directly set $a_0 = 0$, so that the image of $G_h = G_h(a_1, a_3)$ is a surface in the ordinary space, and its gradient field lies in the plane (a_1-a_3). The derivatives of the entropy gap in this case write:

$$\begin{aligned} \frac{1}{2a_1}\frac{\partial G_h}{\partial a_1} &= -2a_1^2\mu_2 - 2a_3^2\mu_4 + \\ &\quad \frac{\sqrt{3}}{2a_1 a_3}\left[\arctan\left(\frac{\sqrt{3}a_3}{a_1}\right) - \arctan\left(\frac{a_3}{\sqrt{3}a_1}\right)\right] \ , \qquad (11) \\ \frac{1}{2a_3}\frac{\partial G_h}{\partial a_3} &= -2a_1^2\mu_4 - 2a_3^2\mu_6 + \frac{1}{a_3^2} + \\ &\quad \frac{\sqrt{3}a_1}{2a_3^3}\left[\arctan\left(\frac{\sqrt{3}a_3}{a_1}\right) - \arctan\left(\frac{a_3}{\sqrt{3}a_1}\right)\right] \ , \qquad (12) \end{aligned}$$

where the quantities μ_2, μ_4, and μ_6 represent the first three even moments of the random signal x, which have values:

$$\mu_2 = \frac{13}{27} \ , \ \mu_4 = \frac{121}{405} \ , \ \mu_6 = \frac{1093}{5103} \ .$$

On the basis of these findings and assuming $\mathbf{R}(\mathbf{a}) = \mathbf{I}_2$ as metric tensor[1], the adapting equations for a_1 and a_3 may be rewritten in compact notation as follows:

$$\begin{aligned} \frac{da_1}{dt} &= -4a_1^3\mu_2 - 4a_3^2 a_1\mu_4 + \frac{\sqrt{3}}{a_3}\arctan\left(\frac{2}{\sqrt{3}}\frac{a_1 a_3}{a_1^2 + a_3^2}\right) \ , \qquad (13) \\ \frac{da_3}{dt} &= -4a_1^2 a_3\mu_4 - 4a_3^3\mu_6 + \frac{2}{a_3} + \frac{\sqrt{3}}{a_3^2}\arctan\left(\frac{2}{\sqrt{3}}\frac{a_1 a_3}{a_1^2 + a_3^2}\right) \ . \qquad (14) \end{aligned}$$

[1]This choice corresponds to assuming a Euclidean metric in the parameter space.

Some interesting observations may be carried out about the above differential system. The differential equation for a_1 clearly has an equilibrium point in $a_1 = 0$; also, the last term in the right-hand side is of limited magnitude order as its absolute value is bounded from above by $\left|\frac{\pi}{a_3}\right|$, while the sum of the first two terms is proportional to $-a_1(a_1^2\mu_2+a_3^2\mu_4)$, that has always the opposite sign of a_1. It can thus be envisaged that $a_1 = 0$ is a stable equilibrium point for the first adapting equation. Conversely, the term $\frac{2}{a_3}$ in the differential equation for a_3 keeps $a_3 \neq 0$; moreover, in the hypothesis that $a_1(t) \to 0$ as t increases, the equilibrium point of the second adapting equation, computed via the condition $\frac{da_3}{dt} = 0$, results to be $a_3 = \pm\sqrt[4]{\frac{1}{2\mu_6}}$.

The considerations about the stability of the above equilibrium points, already supported by the general Lyapunov analysis, are also supported by the phase-portrait of the differential adapting system, depicted in the Figure 1. The phase-portrait analysis of systems is a common and very general technique allowing the study of stability of complex non-linear dynamics when standard linear-analysis tools are inapplicable. Figure 1 shows that the line $a_3 = 0$ is clearly a barrier line, while the derivative field vanishes in correspondence to the mentioned equilibrium points, which are clearly attractors of the system. Because of the presence of the barrier line in $a_3 = 0$, it may be concluded that the sign of $a_3(t)$ never changes, that is, $\text{sign}[a_3(t)] = \text{sign}[a_3(0)]$ for any t.

Remarkably, due to the piece-wise constant structure of the chosen $p_x(x)$, in this case it is possible to express the entropy gap in closed form. It reads:

$$\begin{aligned} G_h(a_1, a_3) \quad = \quad & -\log(\sqrt{\pi}) - 2 - \left(\frac{a_1^2}{3} + \frac{13a_1^4}{54} + \frac{5a_3^2}{27} + \frac{121a_1^2a_3^2}{405} + \frac{1093a_3^4}{10206}\right) + \\ & -\frac{\sqrt{3}a_1}{a_3}\arctan\left(\frac{2}{\sqrt{3}}\frac{a_1a_3}{a_1^2+a_3^2}\right) + \frac{3}{2}\log(a_1^2+3_3^2) - \\ & \frac{1}{2}\log\left(a_1^2+\frac{a_3^2}{3}\right) \ . \end{aligned} \tag{15}$$

A simulated dynamics of the adapting equations is reported in the Figure 2 along with the corresponding dynamics of the Lyapunov function $-G_h(a_1(t), a_3(t))$. The asymptotic value of a_3 in the simulation is about 1.2361, that coincides to the optimal (computed) value of a_3 corresponding to the given μ_6. These results are completely consistent with the numerical simulation results already obtained in [11].

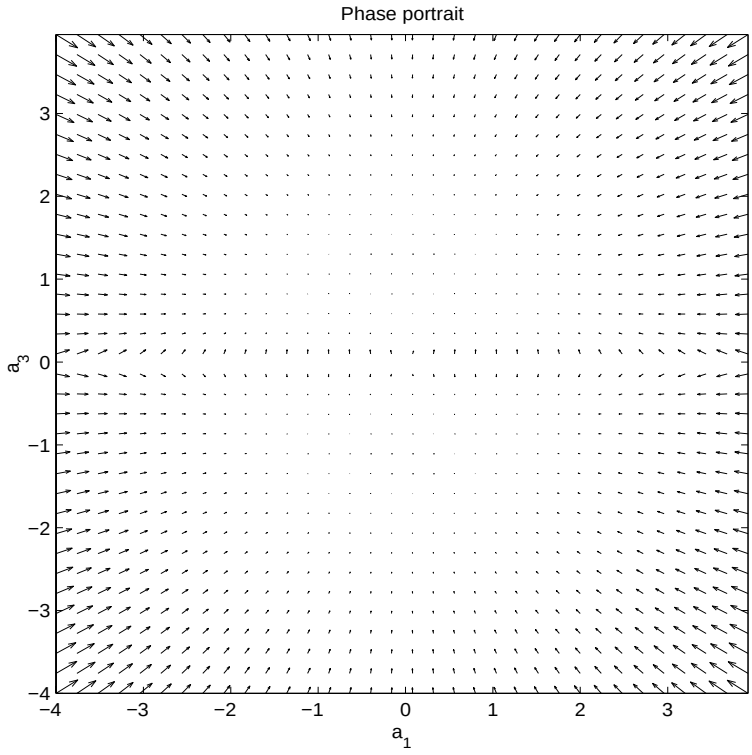

Figure 1: Phase portrait of the adapting system of Case 1.

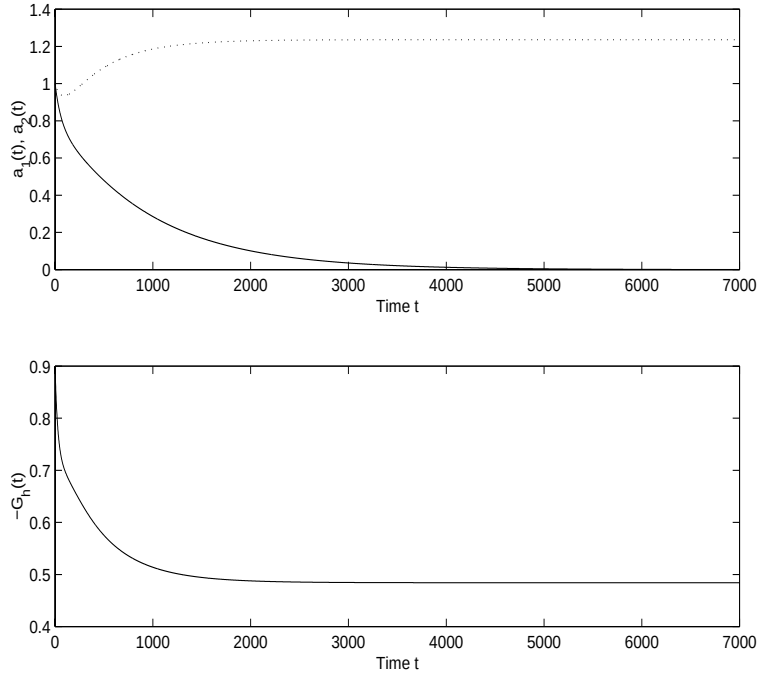

Figure 2: A simulation of the adapting system of Case 1: Top dynamics of $a_1(t)$ and $a_3(t)$; Bottom: Dynamics of $-G_h(t)$.

3.2 Second case-study

An interesting case-study, where the dynamical behavior and stable state of the adapting differential equations can be determined analytically, arises in the following situation:

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \ , \ \varphi(x;\mathbf{a}) = a_0 + a_1^2 x \ . \tag{16}$$

In this case, the partial derivatives of the entropy gap write:

$$\begin{aligned} \frac{1}{2}\frac{\partial G_h}{\partial a_0} &= -(a_0 + \mu a_1^2) \ , \\ \frac{1}{2}\frac{\partial G_h}{\partial a_1} &= \frac{1}{a_1} - 2[a_0 a_1 \mu + a_1^3(\sigma^2 + \mu^2)] \ , \end{aligned}$$

while the estimated PDF reads:

$$\Psi'(x;\mathbf{a}) = \frac{a_1^2}{\sqrt{\pi}} e^{-(a_0 + a_1^2 x)^2} \ . \tag{17}$$

The entropy-gap optima are readily found to be:

$$\mathbf{a}^\star = [-\frac{\mu}{\sqrt{2}\sigma} \ \pm \frac{1}{\sqrt[4]{2\sigma^2}}]^{\mathrm{T}} \ . \tag{18}$$

Plugging these optimal values into expression (17) gives the exact Gaussian probability density function with right mean and variance.

In order to get some qualitative information about the behavior of adapting equations for FAN non-linear unit in the present case, let us assume $\mathbf{R}(\mathbf{a}) = \frac{\alpha}{2}\mathbf{I}_2$ in (6), with $\alpha > 0$. The adapting differential equation for $a_0(t)$ may be solved analytically and has solution:

$$a_0(t) = Ae^{-\alpha t} - \mu a_1^2(t) \ , \tag{19}$$

where A depends from the initial conditions. Consequently, the variable a_0 tends to $-\mu a_1^2$ with a speed controlled by the magnitude of α.

By replacing this expression for $a_0(t)$ in the adapting differential equation of $a_1(t)$, the latter writes:

$$\frac{1}{\alpha}\frac{da_1}{dt} = \frac{1 - 2\mu A e^{-\alpha t} a_1^2 - 2\sigma^2 a_1^4}{a_1} \ ;$$

with the variable-change $c(t) = a_1^2(t) + \frac{\mu A}{2\sigma^2} e^{-\alpha t}$, the above differential equation recasts into the more friendly equation:

$$\frac{1}{2\alpha}\frac{dc}{dt} = 1 - 2\sigma^2 c^2 + \frac{\mu^2 A^2}{2\sigma^2} e^{-2\alpha t} - \frac{\alpha \mu A}{2\sigma^2} e^{-\alpha t} \ .$$

If α is sufficiently large, the last two terms vanish rapidly to zero, and the above differential equation may be approximately solved yielding:

$$a_1^2(t) \approx c(t) \approx -\frac{1}{\sqrt{2}\sigma}\frac{1 - Be^{4\sqrt{2}\alpha\sigma t}}{1 + Be^{4\sqrt{2}\alpha\sigma t}} \ , \tag{20}$$

where the constant B depends upon the initial conditions. The variable $a_1^2(t)$ tends asymptotically to $\frac{1}{\sqrt{2}\sigma}$, as expected; the convergence speed is proportional to the magnitude of α and to the standard deviation of the Gaussian excitation.

3.3 Third case-study

In the previous case we have tackled the problem of approximating the PDF of a Gaussian random signal by means of a FAN non-linear unit endowed with a first-order polynomial obtained by letting $r = 0$. This value of the polynomial order indicator was optimal, in the sense that the approximating function $\Psi'(x; \mathbf{a})$ may represent exactly the signal's PDF provided that the coefficients in $\mathbf{a}$ are properly learnt.

Here we wish to investigate the effect of a bad choice of the polynomial order again in presence of Gaussian excitation: Namely, we choose $r = 1$, that gives rise to a third-order polynomial, and wonder if the system is still able to find a suitable combination of the parameters that ensures a good approximation of the true PDF of incoming stimulus

Formally, we consider the following case:

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{x^2}{2\sigma^2}\right] \ , \ \varphi(x; \mathbf{a}) = a_1^2 x + a_3^2 x^3 \ . \tag{21}$$

Note that the mean-value μ of the excitation has been chosen null and, consequently, the bias parameter a_0 has been directly set to 0; such choice allows treating again a two-variable differential system and, what is important as well, allows computing the entropy-gap integrals in closed form.

In this case the expression of the PDF approximating function writes:

$$\Psi'(x; \mathbf{a}) = \frac{a_1^2 + 3a_3^2 x^2}{\sqrt{\pi}}\exp[-(a_1^2 x + a_3^2 x^3)^2] \ . \tag{22}$$

The derivatives of the entropy-gap read:

$$\begin{aligned}
\frac{1}{2a_1}\frac{\partial G_h}{\partial a_1} &= \int_{\mathcal{R}} p_x(\xi)\left[-2(a_1^2\xi + a_3^2\xi^3)\xi + \frac{1}{a_1^2 + 3a_3^2\xi^2}\right] d\xi \ , \\
\frac{1}{2a_3}\frac{\partial G_h}{\partial a_3} &= \int_{\mathcal{R}} p_x(\xi)\left[-2(a_1^2\xi + a_3^2\xi^3)\xi^3 + \frac{3\xi^2}{a_1^2 + 3a_3^2\xi^2}\right] d\xi \ .
\end{aligned}$$

By making use of the known integrals and the expression of the first three even moments μ_2, μ_4 and μ_6 of a Gaussian random signal reported in the Appendix, the adapting system for the present case writes:

$$\begin{aligned}\frac{da_1}{dt} &= -4a_1(a_1^2\mu_2 + a_3^2\mu_4) + \\ &\quad 2a_1\sqrt{\frac{\pi}{6a_1^2a_3^2\sigma^2}}\exp\left(\frac{a_1^2}{6a_3^2\sigma^2}\right)\operatorname{erfc}\left(\sqrt{\frac{a_1^2}{6a_3^2\sigma^2}}\right) , \qquad (23)\\ \frac{da_3}{dt} &= -4a_1(a_1^2\mu_4 + a_3^2\mu_6) + \frac{2}{a_3} - \\ &\quad 2a_3\sqrt{\frac{\pi a_1^2}{6a_3^6\sigma^2}}\exp\left(\frac{a_1^2}{6a_3^2\sigma^2}\right)\operatorname{erfc}\left(\sqrt{\frac{a_1^2}{6a_3^2\sigma^2}}\right) ; \qquad (24)\end{aligned}$$

note that it has been assumed again $\mathbf{R}(\mathbf{a}) = \mathbf{I}_2$.

It is interesting to note that the above two equations have some repeating terms; also, they do not actually depend on a_1 and a_3, but on $b_1 \stackrel{\text{def}}{=} a_1^2$ and $b_3 \stackrel{\text{def}}{=} a_3^2$. On the basis of these observations, the adapting system recasts into:

$$F(u) \stackrel{\text{def}}{=} \sqrt{\pi u}\exp(u)\operatorname{erfc}(\sqrt{u}) , \qquad (25)$$

$$\frac{1}{4}\frac{db_1}{dt} = -2b_1(b_1\mu_2 + b_3\mu_4) + F\left(\frac{b_1}{6b_3\sigma^2}\right) , \qquad (26)$$

$$\frac{1}{4}\frac{db_3}{dt} = -2b_3(b_1\mu_4 + b_3\mu_6) + 1 - F\left(\frac{b_1}{6b_3\sigma^2}\right) , \qquad (27)$$

where it is understood that the new variables b_1 and b_3 may assume only non-negative values. The function $F(u)$ plays a central role in the behavior of the differential adapting system, thus it deserves a particularized study. For small values of u it behaves as $\sqrt{\pi u}$, while for large values of the argument it tends asymptotically to 1 (see the Appendix).

These considerations lead to the conclusion that both $b_1 = 0$ and $b_3 = 0$ might be equilibrium lines for the differential adapting system; in particular, note that for $b_3 \to 0$ it holds $1 - F\left(\frac{b_1}{6b_3\sigma^2}\right) \to 0$, therefore the last two terms in the equation (27) tend to disappear and the remaining part of the differential equation forces b_3 to zero.

Again, the result of general Lyapunov analysis is confirmed by these observations as well as by the appearance of the phase-portrait of the above system depicted in the Figure 3. The phase-portrait shows that there exists a point $(b_1, b_3) = (b_1^\star, 0)$ that is an attractor for the system, around which the entropy-gap field is much strong; in particular, the stable value of b_1 differs from zero

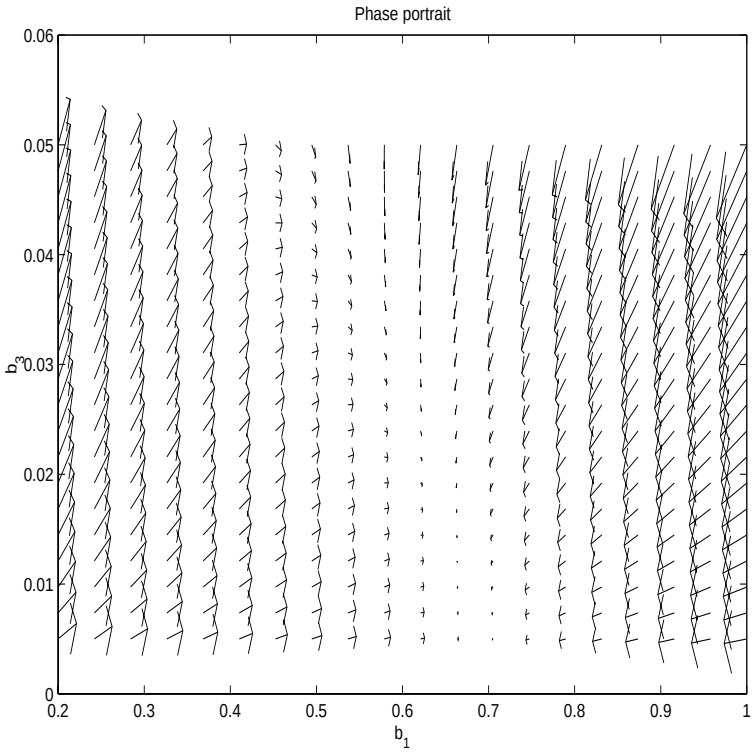

Figure 3: Phase portrait of the adapting system of Case 3 (referred to the adjunct variables b_1-b_3).

and may be found exactly by vanishing the right-hand side of the adapting differential equation for b_1, which gives $-4b_1^2\mu_2 + 2 = 0$. This ultimately means $b_1^\star = \frac{1}{\sqrt{2\sigma^2}}$.

In the present case, the closed-form computation of the entropy gap is not straightforward. However, a closed-form expression may be found by making use of a hyper-geometric expansion, which leads to the following expression:

$$\begin{aligned} G_h(a_1, a_3) &= -\log(\sqrt{\pi}) - (\mu_2 a_1^4 + 2\mu_4 a_1^2 a_3^2 + \mu_6 a_3^4) - \log(a_1^2) \\ &+ \pi \operatorname{erfi}\left(\sqrt{\frac{a_1^2}{6a_3^2\sigma^2}}\right) - \gamma - \frac{a_1^2}{3a_3^2\sigma^2}\, {}_2F_2\left(1, 1; \frac{3}{2}, 2; \frac{a_1^2}{6a_3^2\sigma^2}\right) \\ &- \log\left(\sqrt{\frac{2a_1^2}{3a_3^2\sigma^2}}\right) . \end{aligned} \tag{28}$$

The details of the above complicated expression are reported in the Appendix.

Even in the present case, where the degree of the approximating polynomial has been over-estimated, the theoretical solutions of the differential adapting equations are correct, and coincide to the solutions already found in the previous section. This is a remarkable conclusion, which confirms the robustness of the structure with respect to model over-sizing, already observed through numerical simulations.

The Figure 4 shows the simulated dynamics of the adapting differential equations for $\sigma = 1$: As expected, a_3 tends to zero (in the simulation 0.0497), while a_1 tends to $1/\sqrt[4]{2}$ (in the simulation 0.8363).

To conclude this analysis, it deserves to consider an important problem in artificial neural network and machine-learning studies, namely generalization. As this paper is devoted to analytical study of closed-form expressions of entropy gap and learning equation, we did not make any distinction between e.g. the true entropy and the empirically sampled entropy arising by the empirical distribution function. As a matter of fact, however, in the applications the PDF of the incoming signal is supposed to be unknown (see for instance [10, 12]), therefore the adapting equations cannot be implemented in closed form; their stochastic versions obtain by replacing each term of the form $E_x[f(x)] \stackrel{\text{def}}{=} \int_{\mathcal{X}} p_x(\xi) f(\xi) d\xi$ with its instantaneous (stochastic) approximation $f(x)$ [13] to simulate complete on-line memoryless adaptation.

In order to briefly investigate this problem, an example of the behavior of the stochastic adapting equations is also reported in the Figure 4; it refers to a

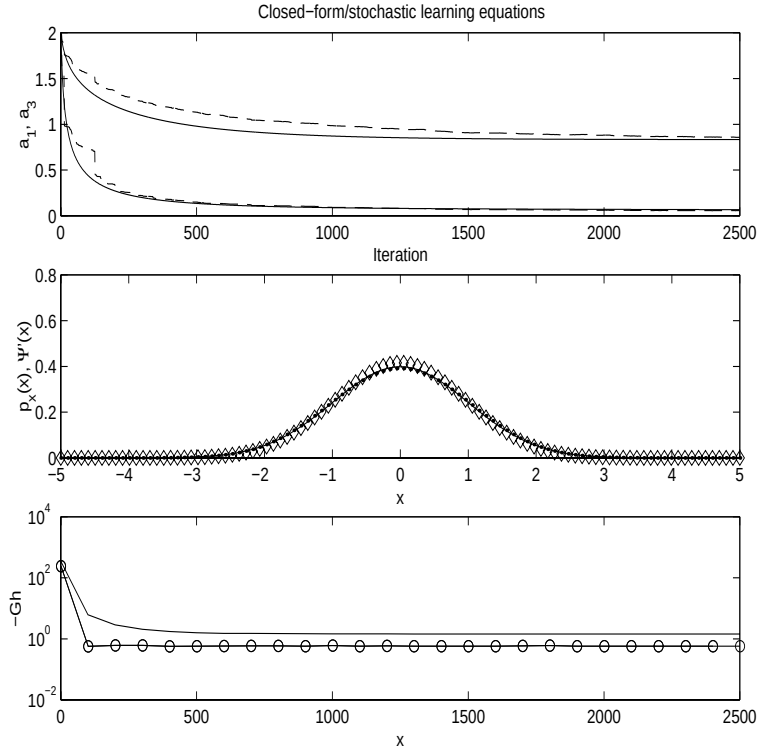

Figure 4: A simulation of the closed-form and stochastic adapting system of Case 3. Top: Functions $a_1(t)$ and $a_3(t)$ in the closed-form case (solid line) and in the stochastic case (dashed line); Middle: Probability density function of the stimulus (solid line) and approximated PDF in the closed-form case (pointed line) and in the stochastic case (diamond-style line); Bottom: Entropy gap computed by the closed-form equation (28) (solid) and by sample-mean (solid-circle).

learning case from incomplete data, which generate the generalization problem; it is important to note that, in order to carry-out such generalization analysis, few input samples were employed $(2, 500)$. In this simulation the courses of variables a_1 and a_3 confirm that the behavior of the stochastic learning equations is consistent with the predicted theoretical behavior. Also, the exact value of the entropy-gap is compared with its sample-mean estimate (obtained by replacing the integral of general formula (4) with sample-average over 50 input samples): Their values are in good agreement, too. Other numerical examples tackling the case of PDF-learning for incomplete real-world data are available in [10, 12, 13].

3.4 Discussion

The previous sections provided a short study on the analytical forms of updating equations for a particular class of neuron model. Such specific model consists of a polynomial expansion of the input to the neuron followed by a monotonic saturating function. The objective is then to select the model parameters so that the entropy gap between the original and transformed random variable is maximized for purposes such as probability density/cumulative distribution function matching. By employing Lyapanov equations the stability of the emerging continuous-time parameter estimation equations has been studied.

It is worth noting that in the original problem the model is essentially a stochastic system, while in the present treatment the stochastic differential equations are given a mean-field-type investigation. While there exist some ways to deal with the stochastic dynamics (see for instance the literature on on-line learning such as [24]), in the present contribution the normal differential equations are dealt with by reducing the analysis to nonlinear dynamics. Also, the properties of the studied dynamical systems referred to the cases-of-study treated, that are helpful for quantitative analysis.

In the present section we wish to briefly discuss some aspects of the presented theory and of the achieved results in view of future research efforts in this field, on the basis of the previous remarks. In particular, the following points can be worth considering:

- **Excitations distributions.** The statistical distributions considered in the cases of study, namely a mixture of two uniform and a Gaussian, are quite typical and help extracting some important properties, such as

the behavior of the neuron in presence of over-estimated model size. In presence of such well-behaved distributions, the closed-form computation of the mean-fields, which, in a general analysis, appears to be extremely difficult, is feasible and leads to closed form expressions for the mean-field learning equations. In a real-world context, such as the ones considered e.g. in [10], the probability density function of the excitations is in general more involved, but it can often be modeled through a closed-form density or a mixture of them, thus the analysis proposed in the present paper may be generalized to other cases. The only complexity relies in case-by-case computation of the learning quantities.

- **Opportunity of mean-field analysis.** On-line learning algorithms are iterative techniques for parameter estimation that distinguish from standard iterative estimation techniques in that each updating-step is based on a single datum. Each successive datum is drawn at random from an underlying distribution, thus the sequence of parameter estimates is given by a Markov process. Most machine learning algorithms possess a on-line version: For very large and redundant datasets, these algorithms can offer a significant time saving in comparison to the batch algorithms that use all the data for each parameter re-estimation step. An extremely interesting and challenging way to characterize the on-line learning is to investigate the evolution of the probability density functions of the neural-system parameters over time by the help of the Kolmogorov master equation (i.e. through its truncated Kramers-Moyal expansion that leads e.g. to the Fokker-Plank equation). However, it should be noted that in the present context (probability density function estimation of incoming stimuli) both data redundancy and on-line-learning-related advantages are not determinant factors, because normally the problem of data scarcity should be tackled and the PDF-estimation operation may be performed off-line because it is often a batch process (see e.g. [1]). In this case, therefore, the opportunity of the presented mean-field analysis stems naturally from the nature of the problem itself.

- **On-line/batch comparison.** To further corroborate the conclusion about the opportunity of the mean-field analysis, the numerical results presented in the Figure 4 are worth considering again in view of assessing

the quality of the prediction of the stochastic behavior of the considered neural system. Figure 4 shows the close similarity of the stochastic behavior of the neuron's tunable parameters and their mean-field behavior during learning confirming that, as far as stochastic learningvis considered, its behavior can be accurately described through a mean-field analysis.

In conclusion, the above considerations suggest that a mean-field analysis of the stochastic learning equations in some typical cases constitutes a sensible choice in the present context and the generality of the mathematical tools employed to achieve such results guarantee their easy extendibility to other cases of interest.

4 Conclusions

The aim of the present brief paper was to analytically illustrate the behavior of FAN non-linear neuronal units in presence of some special stochastic excitations that give rise to tractable mathematics, following the reserach-line initiated with the contribution [14]. First, a general Lyapunov-stability analysis has been carried out, which ensures the convergence of the maximum-entropy-gap learning process toward stable equilibrium states. Then, a case-by-case specific analysis has been conducted in order to gain a deeper insight into the specific features of the considered excitation/FAN-structure pairs. Due to the non-linearity of the involved adapting equations, the differential adapting systems are difficult to analyze: They have been studied through non-linear system analysis tools, such as phase-portraits analysis, in order to predict the behavior of the non-linear unit's parameters during the adapting phase. The numerical simulation of the adapting equations also illustrated the result of the theoretical considerations. It is believed that the provided results may be of future use in the study of this particular form of neuron type model.

5 Acknowledgment

This manuscript is the previously unpublished version of the oral presentation given during WIRN'01 conference, when I received the 2001 "E.R. Caianiello Award" as a young investigator in the neural network field. The award has been conferred to me by Prof. M. Marinaro (University of Salerno) and Prof. G. Orlandi (University of Rome "La Sapienza"): I would take the opportunity

to gratefully thank them and their co-workers as well as Prof. R. Tagliaferri (University of Salerno) for his kind encouragement.

References

[1] J. Biagiotti, S. Fiori, L. Torre, M.A. López-Manchado and J.M. Kenny, *Mechanical Properties of Polypropylene Matrix Composites Reinforced with Natural Fibers: A Statistical Approach*, Polymer Composites. Accepted for publication

[2] H.B. Barlow, *Unsupervised Learning*, Neural Computation, Vo. 1, pp. 295 – 311, 1989

[3] H.B. Barlow, *Guest editorial*, Perception, Vol. 27, pp. 885 – 888, 1998

[4] H.B. Barlow, *Redundancy Reduction Revisited*, Network: Computation in Neural Systems, Vol. 12, pp. 241 – 253, 2001

[5] A.J. Bell and T.J. Sejnowski, *An Information Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation, Vol. 7, No. 6, pp. 1129 – 1159, 1996

[6] A. Cichocki and S.-i. Amari, *Adaptive Blind Signal and Image Processing*, J. Wiley & Sons, Chichester, England, 2003

[7] B. Dwork, *Generalized Hypergeometric Functions*, Oxford, England: Clarendon Press, 1990

[8] D.M. Etter and Y.-F. Cheng, *System modeling using an adaptive delay filter*, IEEE Trans. on Circuits and Systems, Vol. CAS-34, pp. 770 – 774, July 1987

[9] J.A. Feldman and D.H. Ballard, *Connectionist models and their perspectives*, Computer Science, Vol. 6, pp. 205 – 254, 1982

[10] S. Fiori, *Blind Signal Processing by the Adaptive Activation Function Neurons*, Neural Networks, Vol. 13, No. 6, pp. 597 – 611, Aug. 2000

[11] S. Fiori and P. Bucciarelli, *Probability Density Estimation Using Adaptive Activation Function Neurons*, Neural Processing Letters, Vol. 13, No. 1, pp. 31 – 42, Feb. 2001

[12] S. Fiori, *A Contribution to (Neuromorphic) Blind Deconvolution by Flexible Approximated Bayesian Estimation*, Signal Processing, Vol. 81, No. 10, pp. 2131 – 2153, Sept. 2001

[13] S. Fiori, *Probability Density Function Learning by Unsupervised Neurons*, International Journal of Neural Systems, Vol. 11, No. 5, 399 – 417, Oct. 2001

[14] S. Fiori, *Notes on Bell-Sejnowski PDF-Matching Neuron*, Neural Computation,Vol. 14, No. 12, pp. 2847 – 2855, Dec. 2002

[15] M.W. Spratling and G.M. Hayes, *Learning Synaptic Clusters for Nonlinear Dendritic Processing*, Neural Processing Letters, Vol. 11, No. 1, pp. 17 – 27, Feb. 2000

[16] Y.H. Kim, F.L. Lewis and D.M. Dawson, *Hamilton-Jacobi-Bellman Optimal Design of Functional Link Neural Network Controller for Robot Manipulator*, Proc. of the 36^{th} IEEE Conference on Decision and Control, Vol. 2, pp. 1038 – 1043, 1997

[17] S.B. Laughlin, *A Simple Coding Procedure Enhances a Neuron's Information Capacity*, Zeitschrift fur Naturforschung, Vol. 36, pp. 910 – 912, 1981

[18] R. Linsker, *Local Synaptic Rules Suffice to Maximize Mutual Information in a Linear Network*, Neural Computation, Vol. 4, pp. 691 – 702, 1992

[19] M.C. Mackey and L. Glass, *Oscillation and Chaos in Physiological Control Systems*, Science, pp. 287 – 289, July 1977

[20] V.J. Mathews, *Adaptive polynomial filtering*, IEEE Signal Processing Magazine, pp. 10 – 26, 1991

[21] B.W. Mel, *Information Processing in Dendritic Trees*, Neural Computation, Vol. 6, pp. 1031 – 1085, 1994

[22] P. Moerland, *Mixture of Experts Estimate A-Posteriori Probabilities*, International Conference on Artificial Neural Networks (ICANN'97), pp. 499 – 505, 1997

[23] Z. Roth and Y. Baram, *Multidimensional Density Shaping by Sigmoids*, IEEE Trans. on Neural Networks, Vol. 7, No. 5, pp. 1291 – 1298, Sept. 1996

[24] D. Saad (Ed.), *On-line Learning in Neural Networks*, Cambridge University Press, 1998

[25] I.W. Sandberg, *Notes on uniform approximation of time-varying systems on finite time intervals*, IEEE Trans. Circuits and Systems – Part I, Vol. CAS-45, pp. 863 – 865, Aug. 1998

[26] A. Sudjianto and M.H. Hassoun, *Nonlinear Hebbian Rule: A Statistical Interpretation*, Proc. of International Conference on Neural Networks (ICNN'94), Vol. 2, pp. 1247 – 1252, 1994

[27] A. Taleb and C. Jutten, *Entropy Optimization - Application to Source Separation*, Artificial Neural Networks, pp. 529 – 534, Springer-Verlag, 1997

[28] Y. Yang and A.R. Barron, *An Asymptotic Property of Model Selection Criteria*, IEEE Trans. on Information Theory, Vol. 44, No. 1, pp. 95 – 116, Jan. 1998

[29] Z. Yu and C. Yefang, *Adaptive Control of Dynamical Systems Using Functional-Link Net*, Proc. of the IEEE Int. Conf. on Industrial Technology (ICIT'96), pp. 821 – 823, 1996

A Appendix

A.1 Appendix: Known integrals for case 3

The following integrals are useful for completing the calculi for Case 3:

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathcal{R}}\frac{e^{-\frac{x^2}{2\sigma^2}}}{\kappa^2+x^2}dx &= \sqrt{\frac{\pi}{2\kappa^2\sigma^2}}\,e^{\frac{\kappa^2}{2\sigma^2}}\mathrm{erfc}\left(\sqrt{\frac{\kappa^2}{2\sigma^2}}\right)\ ,\\ \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathcal{R}}\frac{x^2e^{-\frac{x^2}{2\sigma^2}}}{\kappa^2+x^2}dx &= 1-\sqrt{\frac{\pi\kappa^2}{2\sigma^2}}\,e^{\frac{\kappa^2}{2\sigma^2}}\mathrm{erfc}\left(\sqrt{\frac{\kappa^2}{2\sigma^2}}\right)\ .\end{aligned}$$

The symbol $\mathrm{erfc}(\cdot)$ denotes the complementary 'error function'.

Also, the first three even moments of a Gaussian distribution are necessary. The second-order moment μ_2 coincides to the variance σ^2; the others two write:

$$\mu_4=\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathcal{R}}x^4e^{-\frac{x^2}{2\sigma^2}}dx=4\sigma^4\ ,\ \mu_6=\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathcal{R}}x^6e^{-\frac{x^2}{2\sigma^2}}dx=15\sigma^6\ .$$

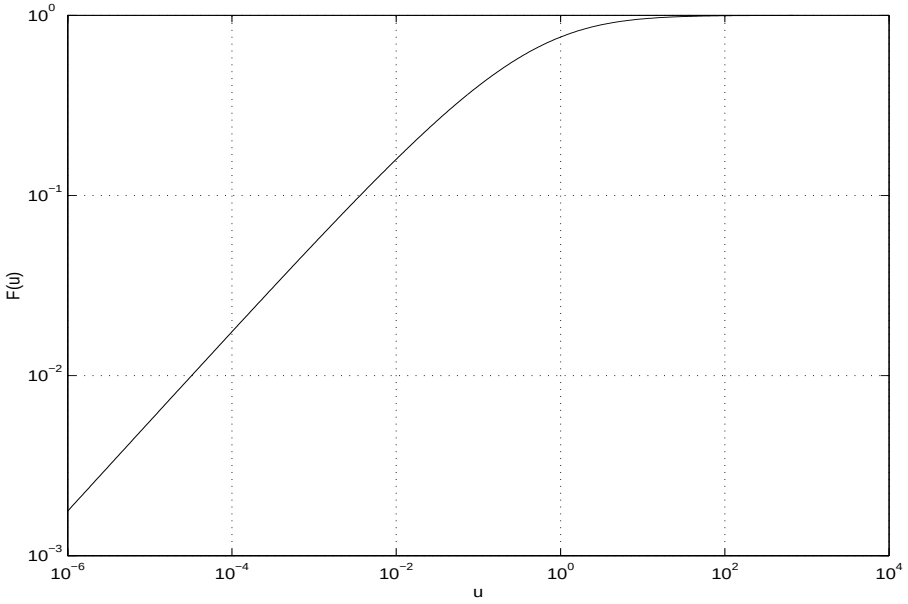

Figure 5: Sketch of the function $F(\cdot)$ of case 3.

A.2 Appendix: Study of the asymptotic behavior of function $F(u)$

The behavior of function $F(u)$ for large u's may be analyzed by making the variable change $z = \frac{1}{\sqrt{u}}$ and by rewriting the function as:

$$F\left(\frac{1}{z^2}\right) = \frac{\sqrt{\pi}\text{erfc}\left(\frac{1}{z}\right)}{z\exp\left(-\frac{1}{z^2}\right)} \ ,$$

that is an indeterminate form for $z \to 0$. The L'Hôspital theorem allows concluding that in the limit $z \to 0$ the above ratio behaves as $\frac{2}{z^2+2}$, therefore for $z \sim 0$ the $F\left(\frac{1}{z^2}\right) \sim 1$.

A sketch of the function is shown in Figure 5 in logarithmic scale for graphical convenience.

A.3 Appendix: Details of the hypergeometric expansion result

Apart from the polynomial part, expression (28) contains a complicated function which, with the convention that $u \stackrel{\text{def}}{=} \frac{a_1^2}{6a_3^2\sigma^2}$, reads:

$$N(u) = \pi\,\text{erfi}(\sqrt{u}) - \gamma - 2u\,{}_2F_2(1,1;3/2,2;u) - \log(4u)\ .$$

Here we have accessed three special mathematical quantities: The 'imaginary error function' erfi(u) defined as $-i\,\text{erf}(iu)$ ($i = \sqrt{-1}$) which takes on real values, symbol γ denotes the Euler-Mascheroni constant (having value 0.5772156...), while symbol ${}_pF_q(a_1,...,a_p;b_1,...,b_q;u)$ denotes the generalized hypergeometric function, which allows expressing various mathematical constants, all the elementary functions as well as many special functions [7]. The generalized hypergeometric function is defined as:

$${}_pF_q(a_1,...,a_p;b_1,...,b_q;u) = \sum_{k=0}^{\infty} \frac{\prod_{j=1}^{p}(a_j)_k\,u^k}{\prod_{j=1}^{p}(b_j)_k\,k!}\ ,$$

where $(a)_k$ is the Pochhammer symbol, defined by $(a)_k = \prod_{m=1}^{k}(a+m-1)$.

For learning system convergence analysis purposes, we are mainly interested in the behavior of $N(u)$ for large values of u (also see Appendix A.2). Exact asymptotic analysis shows that $N(u)$ approaches 0 for large arguments (i.e. when a_3 tends to zero), as expected.