2002s-46

Incorporating Second-Order Functional Knowledge for Better Option Pricing

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, René Garcia

Série Scientifique
Scientific Series

CIRANO
Centre interuniversitaire de recherche en analyse des organisations

Montréal
Mai 2002

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

- École des Hautes Études Commerciales
- École Polytechnique de Montréal
- Université Concordia
- Université de Montréal
- Université du Québec à Montréal
- Université Laval
- Université McGill
- Ministère des Finances du Québec
- MRST
- Alcan inc.
- AXA Canada
- Banque du Canada
- Banque Laurentienne du Canada
- Banque Nationale du Canada
- Banque Royale du Canada
- Bell Canada
- Bombardier
- Bourse de Montréal
- Développement des ressources humaines Canada (DRHC)
- Fédération des caisses Desjardins du Québec
- Hydro-Québec
- Industrie Canada
- Pratt & Whitney Canada Inc.
- Raymond Chabot Grant Thornton
- Ville de Montréal

ISSN 1198-8177

Incorporating Second-Order Functional Knowledge for Better Option Pricing

Charles Dugas[*], *Yoshua Bengio*[†], *François Bélisle*[‡], *Claude Nadeau*[§], *and René Garcia*[**]

Résumé / Abstract

Incorporer une connaissance a priori pour une tache particulière aux algorithmes d'apprentissage peut grandement améliorer leur performance en généralisation. Dans cet article, nous étudions un cas où nous savons que la fonction à apprendre est non-décroissante pour ses deux arguments, et convexe pour l'un d'entre eux. Pour ce cas particulier, nous proposons une classe de fonctions similaires aux réseaux de neurones multi-couches mais (1) avec les propriétés mentionnées plus haut, et (2) est un approximateur universel de fonctions continues avec ces propriétés et avec d'autres. Nous appliquons cette nouvelle classe de fonctions au problème de la modélisation du prix des options d'achat. Nos expériences montrent une amélioration pour la régression sur ces prix d'options d'achat lorsque nous utilisons la nouvelle classe de fonctions qui incorporent les contraintes a priori.

Incorporating prior knowledge of a particular task into the architecture of a learning algorithm can greatly improve generalization performance. We study here a case where we know that the function to be learned is non-decreasing in its two arguments and convex in one of them. For this purpose we propose a class of functions similar to multi-layer neural networks but (1) that has those properties, (2) is a universal approximator of continuous functions with these and other properties. We apply this new class of functions to the task of modeling the price of call options. Experiments show improvements on regressing the price of call options using the new types of function classes that incorporate the a priori constraints.

Mots-clés : Connaissance a priori, algorithme d'apprentissage, approximateur universel, options d'achat.

Keywords: *Prior knowledge, learning algorithm, universal approximator, call options.*

[*] CIRANO and Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Email: dugas@iro.umontreal.ca

[†] CIRANO and Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Tel: +1 (514) 343-6804, email: bengioy@iro.umontreal.ca

[‡] Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Email: belisle@iro.umontreal.ca

[§] Health Canada, Tunney's Pasture, PL 0913A, Ottawa, On, Canada K1A 0K9. Email: Claude_Nadeau@hc-sc.gc.ca

[**] CIRANO and Département de sciences économiques, Université de Montréal. Email: garciar@cirano.qc.ca

1. Introduction

Incorporating *a priori* knowledge of a particular task into a learning algorithm helps reduce the necessary complexity of the learner and generally improves performance, if the incorporated knowledge is relevant to the task and brings enough information about the unknown generating process of the data. In this paper we consider prior knowledge on the positivity of some first and second derivatives of the function to be learned. In particular such constraints have applications to modeling the price of European stock options. Based on the Black-Scholes formula, the price of a call stock option is monotonically increasing in both the "moneyness" and time to maturity of the option, and it is convex in the "moneyness". Section 3 better explains these terms and stock options. For a function $f(x_1, x_2)$ of two real-valued arguments, this corresponds to the following properties:

$$f \geq 0, \qquad \frac{\partial f}{\partial x_1} \geq 0, \qquad \frac{\partial f}{\partial x_2} \geq 0, \qquad \frac{\partial^2 f}{\partial x_1^2} \geq 0 \tag{1}$$

The mathematical results of this paper (section 2) are the following: first we introduce a class of one-argument functions that is positive, non-decreasing and convex in its argument. Second, we use this new class of functions as a building block to design another class of functions that is a universal approximator for functions with positive outputs. Third, once again using the first class of functions, we design a third class that is a universal approximator to functions of two or more arguments, with the set of arguments partitioned in two groups: those arguments for which the second derivative is known positive and those arguments for which we have no prior knowledge on the second derivative. All arguments have the property that their first derivative is positive. The universality property of the third class rests on additional constraints on cross-derivatives, which we illustrate below for the case of two arguments:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} \geq 0, \quad \frac{\partial^3 f}{\partial x_1^2 \partial x_2} \geq 0 \tag{2}$$

Comparative experiments on these new classes of functions were performed on stock option prices, showing improvements when using these new classes rather than ordinary feedforward neural networks. The improvements appear to be non-stationary but the new class of functions shows the most stable behavior in predicting future prices. The detailed results are presented in section 5.

2. Theory

Definition
A class of functions $\hat{\mathcal{F}}$ from $\mathbb{R}^n$ to $\mathbb{R}$ is a **universal approximator** for a class of functions $\mathcal{F}$ from $\mathbb{R}^n$ to $\mathbb{R}$ if for any $f \in \mathcal{F}$, any compact domain $D \subset \mathbb{R}^n$, and any positive ϵ, one can find a $\hat{f} \in \hat{\mathcal{F}}$ with $\sup_{x \in D} |f(x) - \hat{f(x)}| \leq \epsilon$.

It has already been shown that the class of artificial neural networks with one hidden layer

$$\hat{\mathcal{N}} = \{f(x) = b_0 + \sum_{i=1}^{H} w_i h(b_i + \sum_j v_{ij} x_j)\} \tag{3}$$

e.g. with a sigmoid activation function $h(s) = 1/(1+e^{-s})$, is a universal approximator of continuous functions (Cybenko (1988, 1989), Hornik et al. (1989), Barron (1993)). The number of hidden units H of the neural network is a hyper-parameter that controls the accuracy of the approximation and it should be chosen to balance the trade-off between accuracy (bias of the class of functions) and variance (due to the finite sample used to estimate the parameters of the model), see also (Moody (1994)). Because of this trade-off, in the finite sample case, it may be advantageous to consider a "smaller" class of functions that is appropriate to the task.

Since the sigmoid h is monotonically increasing, it is easy to force the first derivatives with respect to x to be positive by forcing the weights to be positive, for example with the exponential function:

$$\hat{\mathcal{N}}_+ = \{f(x) = b_0 + \sum_{i=1}^{H} e^{w_i} h(b_i + \sum_j e^{v_{ij}} x_j)\} \tag{4}$$

because $h'(s) = h(s)(1-h(s)) > 0$.

Since the sigmoid h has a positive first derivative, its primitive, which we call *softplus*, is convex:

$$\boxed{\zeta(s) = \log(1+e^s)} \tag{5}$$

i.e., $d\zeta(s)/ds = h(s) = 1/(1+e^{-s})$.

2.1 Universality for functions with strictly positive outputs

Using the softplus function introduced above, we define a new class of functions, all of which have strictly positive outputs:

$$\hat{\mathcal{N}}_{>0} = \{f(x) = \zeta(g(x)), g(x) \in \hat{\mathcal{N}}\} \tag{6}$$

Theorem Within the set of continuous functions from $\mathbb{R}^n$ to $\mathbb{R}_+ = \{x : x \in \mathbb{R}, x > 0\}$, the class $\hat{\mathcal{N}}_{>0}$ is a universal approximator.

Proof Consider $f(x)$, a function with strictly positive outputs which we want to approximate arbitrarily well. Consider $g(x) = \zeta^{-1}(f(x)) = \ln(e^{f(x)} - 1)$, the inverse softplus transform of $f(x)$. Choose $\hat{g}(x)$ from $\hat{\mathcal{N}}$ such that $\sup_{x\in D} |g(x) - \hat{g(x)}| \le \epsilon$. The existence of $\hat{g}(x)$ is ensured by the universality property of $\hat{\mathcal{N}}$. Set $\hat{f}(x) = \zeta(\hat{g}(x)) = \ln(1+e^{\hat{g}(x)})$. Consider any particular x and define $a = \min(\hat{g}(x), g(x))$ and $b = \max(\hat{g}(x), g(x))$. Since $b - a \le \epsilon$, then,

$$\begin{aligned}
|\hat{f}(x) - f(x)| &= \ln(1+e^b) - \ln(1+e^a) \\
&\le \ln\left(1 + (e^\epsilon - 1)e^a/(1+e^a)\right) \\
&< \epsilon
\end{aligned}$$

and the proof is complete. Thus, the use of the softplus function to transform the output of a regular one hidden layer artificial neural network ensures the positivity of the final output without hindering the universality property.

2.2 The class ${}_{c,n}\hat{\mathcal{N}}_{++}$

In this section, we use the softplus function, in order to define a new class of functions with positive outputs, positive first derivatives w.r.t. all input variables and positive second derivatives w.r.t. some of the input variables. The basic idea is to replace the sigmoid of a sum by a product of either softplus or sigmoid functions over each of the dimensions (using the softplus over the convex dimensions and the sigmoid over the others):

$$ {}_{c,n}\hat{\mathcal{N}}_{++} = \{f(x) = e^{b_0} + \sum_{i=1}^{H} e^{w_i} (\prod_{j=1}^{c} \zeta(b_{ij} + e^{v_{ij}} x_j))(\prod_{j=c+1}^{n} h(b_{ij} + e^{v_{ij}} x_j))\} \tag{7} $$

One can readily check that the output is necessarily positive, the first derivatives w.r.t. x_j are positive, and the second derivatives w.r.t. x_j for $j \leq c$ are positive. However, this class of functions has other properties. Let $(j_1, \cdots, j_m)$ be a set of indices with $1 \leq j_i \leq c$ (convex dimensions), and let $(j'_1, \cdots, j'_p)$ be a set of indices $c+1 \leq j'_i \leq n$ (the other dimensions), then

$$ \frac{\partial^{m+p} f}{\partial x_{j_1} \cdots \partial x_{j_m} \partial x_{j'_1} \cdots x_{j'_p}} \geq 0, \quad \frac{\partial^{2m+p} f}{\partial x_{j_1}^2 \cdots \partial x_{j_m}^2 \partial x_{j'_1} \cdots x_{j'_p}} \geq 0 \tag{8} $$

The set of functions that respect these derivative conditions will be refered to as ${}_{c,n}\mathcal{F}_{++}$. Note that m or p can be 0, so as special cases we find that f is positive, and that it is monotonically increasing w.r.t. all its inputs, and convex w.r.t. the first c inputs. Also note that this set of equations, when applied to our particular case where $n = 2, c = 1$, corresponds to equations 1 and 2.

2.3 Universality of ${}_{c,n}\hat{\mathcal{N}}_{++}$

We now state the universality theorem and present the associated proof:

Theorem Within the set ${}_{c,n}\mathcal{F}_{++}$ of continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$ whose set of derivatives as specified by equation 8 are non-negative, the class ${}_{c,n}\hat{\mathcal{N}}_{++}$ is a universal approximator.

Proof

We develop a constructive proof for which we define the threshold functions $\theta(x) = I_{x \geq 0}$ and the positive part $x_+ = max(0, x)$. These two functions are part of the closure of the set ${}_{c,n}\hat{\mathcal{N}}_{++}$ since

$$ \theta(x) = \lim_{a \to \infty} h(ax) \tag{9} $$
$$ x_+ = \lim_{a \to \infty} \zeta(ax) \tag{10} $$

Let D be the compact domain over which we wish to obtain an approximation error below ϵ in every point. Suppose the existence of an oracle allowing us to evaluate the function in a certain number of points. Let T be the smallest hyperrectangle encompassing D. Let us partition T in hypercubes with sides of length L so that the variation of the function between two neighboring points is bounded above by ϵ. For example, given s, an upper bound on the gradient of the function in any direction, setting $L \leq \epsilon/s$ would do the trick. The number of hypercubes is N_1 over the x_1 axis, N_2 over the x_2 axis, ..., N_n over the x_n axis. The number of points on the treillis formed within T is $H =$

$(N_1+1)\cdot(N_2+1)\cdot\ldots\cdot(N_n+1)$. We define treillis points $\vec{a}=(a_1,a_2,\cdots,a_n)$ and $\vec{b}=(b_1,b_2,\cdots,b_n)$ as the innermost (closest to origin) and outermost corners of T, respectively. Figure 1 illustrates these values. The points of the grid are defined as such: $\vec{p}_1=a, \vec{p}_2=(a_1,a_2,\ldots,a_n+L),\ldots,\vec{p}_{N_n+1}=(a_1,a_2,\ldots,b_n),\ldots,\vec{p}_H=b$.

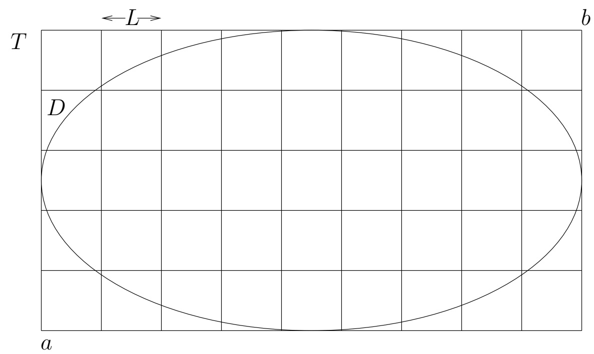

Figure 1: *Two dimensional illustration of the proof of universality: The ellipse D corresponds to the domain of observation over which we wish to obtain a universal approximator. The rectangle T encompasses D and is partitioned in squares of length L. Points a and b are the innermost (closest to origin) and outermost corners of T, respectively.*

Starting with an approximating function $\hat{f}_0=f(\vec{a})$, we scan the grid in an orderly manner, according to the definition of the set of points $\{\vec{p}_h\}$. At each point along the grid, we add a term to the current approximating function so that it becomes exact at that point:

$$\begin{aligned}\hat{f}_h &= g_h+\hat{f}_{h-1}\\ &= \sum_{k=1}^{h} g_k \qquad (11)\end{aligned}$$

The increment term g_h must be such that $\hat{f}_h(\vec{p}_h)=f(\vec{p}_h)$. We therefore compute the term δ_h as the difference between the value of the function evaluated at point $\vec{p}_h$ and the value of the currently accumulated approximating function $\hat{f}_{h-1}$ at the same point:

$$\delta_h = f(\vec{p}_h)-\hat{f}_{h-1}(\vec{p}_h) \qquad (12)$$

However, g_h must not affect the value of the approximating function at grid points that have already been visited. According to our sequencing of the grid points, this corresponds to having $g_h(\vec{p}_k)=0$ for $0<k<h$. Enforcing this constraint ensures that $\hat{f}_h(\vec{p}_k)=\hat{f}_k(\vec{p}_k)=f(\vec{p}_k)$, $0<k<h$. We define

$$\beta(\vec{y},\vec{z}) = \prod_{j=1}^{c}(y(j)-z(j)+L)_+\cdot\prod_{j=c+1}^{n}\theta(y(j)-z(j)) \qquad (13)$$

where $y(j)$ is the j^{th} coordinate of $\vec{y}$ and similarly for $\vec{z}$. We have assumed, without loss of generality, that the convex dimensions are the first c ones. One can readily verify that $\beta(\vec{p}_k, \vec{p}_h) = 0$ for $0 < k < h$. We can now define the incremental term:

$$g_h(\vec{p}) \quad = \quad \delta_h \beta(\vec{p}, \vec{p}_h) \tag{14}$$

so that after all treillis points have been visited, our final approximation is

$$\hat{f}_H(\vec{p}) \quad = \quad \sum_{k=1}^{H} g_k(\vec{p}) \tag{15}$$

with $f(\vec{p}) = \hat{f}_H(\vec{p})$ for all treillis points.

So far, we have devised a way to approximate the target function as a sum of terms from the set ${}_{c,n}\hat{\mathcal{N}}_{++}$. We know our approximation to be exact in every point of a grid tight enough so that the function varies at most by ϵ between neighbor points. The target function respects a set of constraints on its derivatives as expressed in equation 8. The terms of our final approximate function $\hat{f}_H$ also individually respect the constraints. One essential question remains: does $\hat{f}_H$, the sum of all these terms, also respect these constraints ? In order to ensure this, we need to show that $\delta_h \geq 0 \; \forall h$.

Let $p_k(j)$ be the value of the j^{th} coordinate of point $\vec{p}_k$. Also, let $p_k(j) = a(j) + i_k(j)L$. In other words, $\vec{p}_k = \vec{a} + L \cdot (i_k(1), i_k(2), \ldots, i_k(n))$ and this defines the bijective relationship between k and the $\{i_k(j)\}_{j=1}^{n}$ values. Note that for any pair of points $(\vec{y}, \vec{z})$, we have $y(j) - z(j) = L(i_y(j) - i_z(j))$. Let us first express the target function in terms of the δ values.

$$\begin{aligned}
f(\vec{p}_h) &= \hat{f}_H(\vec{p}_h) \\
&= \sum_{k=1}^{H} g_k(\vec{p}_h) \\
&= \sum_{\{k : p_k(j) \leq p_h(j) \, \forall j\}} g_k(\vec{p}_h) \\
&= \sum_{\{k : p_k(j) \leq p_h(j) \, \forall j\}} \delta_k \prod_{j=1}^{c} (p_h(j) - p_k(j) + L)_+ \cdot \prod_{j=c+1}^{n} \theta(p_h(j) - p_k(j)) \\
&= \sum_{\{k : p_k(j) \leq p_h(j) \, \forall j\}} \delta_k \prod_{j=1}^{c} (i_h(j) - i_k(j) + 1)L \cdot \prod_{j=c+1}^{n} 1 \\
&= \sum_{i_k(1)=1}^{i_h(1)} \sum_{i_k(2)=1}^{i_h(2)} \cdots \sum_{i_k(n)=1}^{i_h(n)} \delta_k L^c \prod_{j=1}^{c} (i_h(j) - i_k(j) + 1)
\end{aligned} \tag{16}$$

Then, we define the finite difference of the function along the l^{th} axis as such:

$$\Delta_l f(\vec{p}_h) \quad = \quad f(\vec{p}_h) - f(\vec{p}_{k : i_k(l) = i_h(l) - 1}) \tag{17}$$

so that k is a neighbor of h on the hypergrid of points within T. All coordinates of k and h are the same except along the l^{th} axis where $i_k(l) = i_h(l) - 1$. Using equations 17 and 16 we get:

$$\begin{aligned}\Delta_l f(\vec{p}_h) &= \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(l)=1}^{i_h(l)} \cdots \sum_{i_k(n)=1}^{i_h(n)} \delta_k L^c \prod_{j=1}^{c} (i_h(j) - i_k(j) + 1) \\ &- \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(l)=1}^{i_h(l)-1} \cdots \sum_{i_k(n)=1}^{i_h(n)} \delta_k L^c \prod_{j=1}^{c} (i_h(j) - i_k(j) + 1) \end{aligned} \tag{18}$$

We then reorder the summations so that the l^{th} axis comes last. We also factor out the terms that are independent of l. Let us first consider the case where $l \leq c$, i.e., the l^{th} dimension bears the convexity property:

$$\Delta_l f(\vec{p}_h) = \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(n)=1}^{i_h(n)} L^c \prod_{j=1, j\neq l}^{c} (i_h(j) - i_k(j) + 1) \cdot \alpha(l)$$

where,

$$\begin{aligned}\alpha(l) &= \sum_{i_k(l)=1}^{i_h(l)} \delta_k (i_h(l) - i_k(l) + 1) - \sum_{i_k(l)=1}^{i_h(l)-1} \delta_k (i_h(l) - 1 - i_k(l) + 1) \\ &= \sum_{i_k(l)=1}^{i_h(l)} \delta_k \end{aligned} \tag{19}$$

so that $\Delta_l f(\vec{p}_h)$ still bears a dependency on its l^{th} dimension. We therefore differentiate once more along the l^{th} axis:

$$\begin{aligned}\Delta_l^2 f(\vec{p}_h) &= \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(n)=1}^{i_h(n)} L^c \prod_{j=1, j\neq l}^{c} (i_h(j) - i_k(j) + 1) \cdot \left(\sum_{i_k(l)=1}^{i_h(l)} \delta_k - \sum_{i_k(l)=1}^{i_h(l)-1} \delta_k \right) \\ &= \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(n)=1}^{i_h(n)} \delta_{k:i_k(l)=i_h(l)} L^c \prod_{j=1, j\neq l}^{c} (i_h(j) - i_k(j) + 1) \end{aligned} \tag{20}$$

so that we remove any dependency on the l^{th} dimension. Now in case $l > c$, our task is simpler since:

$$\begin{aligned}\alpha(l) &= \sum_{i_k(l)=1}^{i_h(l)} \delta_k - \sum_{i_k(l)=1}^{i_h(l)-1} \delta_k \\ &= \delta_{k:i_k(l)=i_h(l)} \end{aligned}$$

and we obtain:

$$\Delta_l f(\vec{p}_h) = \sum_{i_k(1)=1}^{i_h(1)} \cdots \sum_{i_k(n)=1}^{i_h(n)} \delta_{k:i_k(l)=i_h(l)} L^c \prod_{j=1, j\neq l}^{c} (i_h(j) - i_k(j) + 1) \tag{21}$$

Note the similarity of equations 20 and 21. Both remove dependency along the dimension of differentiation. They address the cases where $l \leq c$ and $l > c$, respectively. This procedure can be applied recursively over all dimensions so that in the end,

$$\begin{aligned}\Delta_1^2 \dots \Delta_c^2 \Delta_{c+1} \dots \Delta_n f(\vec{p}_h) &= L^c \delta_{k:i_k(1)=i_h(1)\dots i_k(n)=i_h(n)} \\ &= L^c \delta_h \end{aligned} \tag{22}$$

and we have finally isolated the value of δ_h as a function of L and a finite difference of order $n+c$ of the function. The value of δ_h is non negative iff this finite difference value also is.

Now, according to the mean value theorem,

$$\Delta f = \frac{f(b)-f(a)}{b-a} = \frac{1}{b-a} \int_a^b f' \, dx \tag{23}$$

so that if $f' \geq 0$ over the range $[a, b]$, then consequentely, $\Delta f \geq 0$. Applying this to our case, we set

$$\frac{\partial^{n+c} f(p_h)}{\partial x_1^2 \partial x_2^2 \dots \partial x_c^2 \partial x_{c+1} \dots \partial x_n} \geq 0 \tag{24}$$

over the hyperrectangle T so that

$$\Delta_1^2 \dots \Delta_c^2 \Delta_{c+1} \dots \Delta_n f(\vec{p}_h) \geq 0 \tag{25}$$

over T as well and consequently, $\delta_h \geq 0 \, \forall h$ and the proof is complete.

Corollary Within the set of positive continuous functions from $\mathbb{R}$ to $\mathbb{R}$ whose first and second derivatives are non-negative, the class ${}_{1,1}\hat{\mathcal{N}}_{++}$ is a universal approximator.

2.4 Illustration of proof for ${}_{1,2}\hat{\mathcal{N}}_{++}$

In order give the reader a better intuition as to how we were able to isolate the δ_h factor in equation 22, we apply the finite difference method to ${}_{1,2}\hat{\mathcal{N}}_{++}$, the set of functions that include call price functions, i.e., positive convex w.r.t. the first variable and monotone increasing w.r.t. both variables. Figure 2 illustrates the two dimensional setting of our example with the points of the grid labelled in the order in which they are scanned according the constructive procedure. We will show how to isolate δ_6.

For the set ${}_{1,2}\hat{\mathcal{N}}_{++}$, we have,

$$f(\vec{p}_h) = \sum_{k=1}^{H} \delta_k \cdot (p_h(1) - p_k(1) + L)_+ \cdot \theta(p_h(2) - p_k(2)) \tag{26}$$

Applying this to the six grid points of Figure 2, we obtain $f(\vec{p}_1) = L\delta_1$, $f(\vec{p}_2) = L(\delta_1 + \delta_2)$, $f(\vec{p}_3) = L(2\delta_1 + \delta_3)$, $f(\vec{p}_4) = L(2\delta_1 + 2\delta_2 + \delta_3 + \delta_4)$, $f(\vec{p}_5) = L(3\delta_1 + 2\delta_3 + \delta_5)$, $f(\vec{p}_6) = L(3\delta_1 + 3\delta_2 + 2\delta_3 + 2\delta_4 + \delta_5 + \delta_6)$.

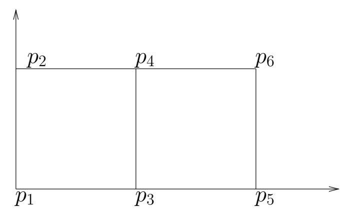

Figure 2: *Illustration in two dimensions of the constructive proof. The grid is scanned along the abscissa axis, then along the ordinates axis. The points are labelled accordingly from 1 to 6. The function is known to be convex w.r.t. to the first variable (abscissa) and monotone increasing w.r.t. both variables.*

Differentiating w.r.t. the second variable, then the first, we have:

$$\begin{aligned}
\Delta_2 f(\vec{p}_6) &= f(\vec{p}_6) - f(\vec{p}_5) \\
\Delta_1 \Delta_2 f(\vec{p}_6) &= (f(\vec{p}_6) - f(\vec{p}_4)) - (f(\vec{p}_5) - f(\vec{p}_3)) \\
\Delta_1 \Delta_2 f(\vec{p}_6) &= (f(\vec{p}_6) - f(\vec{p}_4)) - (f(\vec{p}_4) - f(\vec{p}_2)) \\
&- (f(\vec{p}_5) - f(\vec{p}_3)) + (f(\vec{p}_3) - f(\vec{p}_1)) \\
&= \delta_6
\end{aligned}$$

This procedure can be repeated for any point on the grid. The conclusion associated with this result is that the third finite difference of the function must be positive in order for δ_6 to be positive as well. As stated above, enforcing the corresponding derivative is a slightly stronger condition which is respected by all element functions of ${}_{1,2}\hat{\mathcal{N}}_{++}$.

For points close to the boundary, it is simpler to isolate the δ_h value as fewer finite difference values need to be computed. The constraint for the associated δ_h values is therefore set on derivatives of lower orders which are still respected by all elements of ${}_{1,2}\hat{\mathcal{N}}_{++}$.

3. Estimating Call Option Prices

An option is a contract between two parties that entitles the buyer to a claim at a future date T that depends on the future price, S_T of an underlying asset whose price at time t is S_t. In this paper we consider the very common European call options, in which the value of the claim *at maturity* (time T) is $\max(0, S_T - K)$, i.e. if the price is above the *strike price* K, then the seller of the option owes $S_T - K$ dollars to the buyer, otherwise, the option expires worthless. In the no-arbitrage framework, the call function is believed to be a function of the actual market price of the security (S_t), the strike price (K), the remaining time to maturity ($\tau = T - t$), the risk free interest rate (r), and the volatility of the return (σ). The challenge is to evaluate the value of the option prior to the expiration date before entering a transaction. The risk free interest rate (r) needs to be somehow extracted from the term structure of interest rates and the volatility (σ) needs to be forecasted, this latest task being a field of research in itself. We have (Dugas et al. (2000)) previously tried to feed in neural networks with estimates of the volatility using historical averages but so far, the gains remained insignificant. We therefore drop these two features and rely on the ones that can be observed: S_t, K, τ. One more important result is that under mild conditions, the call option function is homogeneous of degree one with respect to the strike price and so our final approximation depends on two variables: the moneyness ($M = S_t/K$) and the time to maturity (τ).

$$C_t/K = f(M, \tau) \tag{27}$$

Simple arbitrage theory imposes the properties of equation 1 on the call option function[1] (Garcia and Gençay (1998)). Stronger parametric assumptions yield the Black-Scholes formula (Black and Scholes (1973)):

$$f(M, \tau, r, \sigma) = M\mathcal{N}(d_1) - e^{-r\tau}\mathcal{N}(d_2) \tag{28}$$

where $\mathcal{N}(\cdot)$ is the cumulative gaussian function evaluated in points

$$d_1, d_2 = \frac{\ln M + (r \pm \sigma^2/2)\tau}{\sigma\sqrt{\tau}} \tag{29}$$

i.e., $d_1 = d_2 + \sigma\sqrt{\tau}$. Let's confront this formula to our set of constraints.

$$\frac{\partial f}{\partial M} = \mathcal{N}(d_1) \tag{30}$$

$$\frac{\partial^2 f}{\partial M^2} = \frac{\mathcal{N}'(d_1)}{\sqrt{\tau}M\sigma} \tag{31}$$

$$\frac{\partial f}{\partial \tau} = e^{-r\tau}\left(\frac{\mathcal{N}'(d_2)\sigma}{2\sqrt{\tau}} + r\mathcal{N}(d_2)\right) \tag{32}$$

$$\frac{\partial^2 f}{\partial M \partial \tau} = \frac{\mathcal{N}'(d_1)}{2\sigma\tau^{3/2}}\left((r + \sigma^2/2)\tau - \ln M\right) \tag{33}$$

$$\frac{\partial^3 f}{\partial M^2 \partial \tau} = \frac{\mathcal{N}'(d_1)}{2M\sigma^3\tau^{5/2}}\left(\ln^2 M - \sigma^2\tau - (r + \sigma^2/2)^2\tau^2\right) \tag{34}$$

where $\mathcal{N}'(\cdot)$ is the gaussian density function. Equations 30, 31 and 32 confirm that the Black-Scholes formula is in accordance with our prior knowledge of the call option function: all three derivatives are positive. Equations 33 and 34 are the cross derivatives which will be positive for any function chosen from ${}_{1,2}\hat{\mathcal{N}}_{++}$. When applied to the Black-Scholes formula, it is less clear whether these values are positive, too. In particular, one can easily see that both cross derivatives can not be simultaneously positive. Thus, the Black-Scholes formula is not within the set ${}_{1,2}\hat{\mathcal{N}}_{++}$. Then again, it is known that the Black-Scholes formula does not adequately represent the market pricing of options, but it is considered as a useful guide for evaluating call option prices. So, it is not clear whether these constraints on the cross derivatives should or not be present in the true price function.

4. Experimental Setup

As a reference model, we use a simple multi-layered perceptron with one hidden layer (eq. 3). We also compare our results with a recently proposed model (Garcia and Gençay (1998))

1. The convexity of the call option w.r.t. the moneyness is a consequence of the butterfly spread strategy.

that closely resembles the Black-Scholes formula for option pricing (i.e. another way to incorporate possibly useful prior knowledge):

$$
\begin{aligned}
y^{BS} &= \alpha + M \cdot \sum_{i=1}^{n_h} \beta_{1,i} \cdot h(\gamma_{i,0} + \gamma_{i,1} \cdot M + \gamma_{i,2} \cdot \tau) \\
&+ e^{-r\tau} \cdot \sum_{i=1}^{n_h} \beta_{2,i} \cdot h(\gamma_{i,3} + \gamma_{i,4} \cdot M + \gamma_{i,5} \cdot \tau).
\end{aligned} \tag{35}
$$

We evaluate two new architectures incorporating some or all of the constraints defined in equation 8.

We used european call option data from 1988 to 1993. A total of 43518 transaction prices on european call options on the S&P500 index were used. In section 5, we report results on 1988 data. In each case, we used the first two quarters of 1988 as a training set (3434 examples), the third quarter as a validation set (1642 examples) for model selection and the fourth quarter as a test set (each with around 1500 examples) for final generalization error estimation. In tables 1 and 2, we present results for networks with unconstrained weights on the left-hand side, and weights constrained to positive and monotone functions through exponentiation of parameters on the right-hand side. For each model, the number of hidden units varies from one to nine. The mean squared error results reported were obtained as follows: first, we randomly sampled the parameter space 1000 times. We picked the best (lowest training error) model and trained it up to 1000 more epochs. Repeating this procedure 10 times, we selected and averaged the performance of the best of these 10 models (those with training error no more than 10% worse than the best out of 10). In figure 3, we present tests of the same models on each quarter up to and including 1993 (20 additional test sets) in order to assess the persistence (conversely, the degradation through time) of the trained models.

5. Forecasting Results

As can be seen in tables 1 and 2, the positivity constraints through exponentiation of the weights allow the networks to avoid overfitting. The training errors are generally slightly lower for the networks with unconstrained weights, the validation errors are similar but final test errors are disastrous for unconstrained networks, compared to the constrained ones. This "liftoff" pattern when looking at training, validation and testing errors has triggered our attention towards the analysis of the evolution of the test error through time. The unconstrained networks obtain better training, validation and testing (test 1) results but fail in the extra testing set (test 2). Constrained architectures seem more robust to changes in underlying econometric conditions. The constrained Black-Scholes similar model performs slightly better than other models on the second test set but then fails on latter quarters (figure 3). All in all, at the expense of slightly higher initial errors our proposed architecture allows us to forecast with increased stability much farther in the future. This is a very welcome property as new derivative products have a tendency to lock in values for much longer durations (up to 10 years) than traditional ones.

Simple Multi-Layered Perceptrons

Mean Squared Error Results on Call Option Pricing ($\times 10^{-4}$)

Units	Unconstrained weights				Constrained weights			
	Train	Valid	Test1	Test2	Train	Valid	Test1	Test2
1	2.38	1.92	2.73	6.06	2.67	2.32	3.02	3.60
2	1.68	1.76	1.51	5.70	2.63	2.14	3.08	3.81
3	1.40	1.39	1.27	27.31	2.63	2.15	3.07	3.79
4	1.42	1.44	1.25	27.32	2.65	2.24	3.05	3.70
5	1.40	1.38	**1.27**	**30.56**	2.67	2.29	3.03	3.64
6	1.41	1.43	1.24	33.12	2.63	2.14	**3.08**	**3.81**
7	1.41	1.41	1.26	33.49	2.65	2.23	3.05	3.71
8	1.41	1.43	1.24	39.72	2.63	2.14	3.07	3.80
9	1.40	1.41	1.24	38.07	2.66	2.27	3.04	3.67

Black-Scholes Similar Networks

Mean Squared Error Results on Call Option Pricing ($\times 10^{-4}$)

Units	Unconstrained weights				Constrained weights			
	Train	Valid	Test1	Test2	Train	Valid	Test1	Test2
1	1.54	1.58	1.40	4.70	2.49	2.17	2.78	3.61
2	1.42	1.42	1.27	24.53	1.90	1.71	2.05	3.19
3	1.40	1.41	1.24	30.83	1.88	1.73	2.00	3.72
4	1.40	1.39	**1.27**	**31.43**	1.85	1.70	1.96	3.15
5	1.40	1.40	1.25	30.82	1.87	1.70	2.01	3.51
6	1.41	1.42	1.25	35.77	1.89	1.70	2.04	3.19
7	1.40	1.40	1.25	35.97	1.87	1.72	1.98	3.12
8	1.40	1.40	1.25	34.68	1.86	1.69	**1.98**	**3.25**
9	1.42	1.43	1.26	32.65	1.92	1.73	2.08	3.17

Table 1: *Left: the parameters are free to take on negative values. Right: parameters are constrained through exponentiation so that the resulting function is both positive and monotone increasing everywhere w.r.t. to both inputs. Top: regular feed-forward artificial neural networks. Bottom: neural networks with an architecture resembling the Black-Scholes formula as defined in equation 35. The number of units varies from 1 to 9 for each network architecture. The first two quarters of 1988 were used for training, the third of 1988 for validation and the fourth of 1988 for testing. The first quarter of 1989 was used as a second test set to assess the persistence of the models through time (figure 3). In bold: test results for models with best validation results.*

6. Conclusions

Motivated by prior knowledge on the derivatives of the function that gives the price of European options, we have introduced new classes of functions similar to multi-layer neural

Products of SoftPlus and Sigmoid Functions

Mean Squared Error Results on Call Option Pricing ($\times 10^{-4}$)

Units	Unconstrained weights				Constrained weights			
	Train	Valid	Test1	Test2	Train	Valid	Test1	Test2
1	2.27	2.15	2.35	3.27	2.28	2.14	2.37	3.51
2	1.61	1.58	1.58	14.24	2.28	2.13	2.37	3.48
3	1.51	1.53	1.38	18.16	2.28	2.13	2.36	3.48
4	1.46	1.51	1.29	20.14	1.84	1.54	**1.97**	**4.19**
5	1.57	1.57	1.46	10.03	1.83	1.56	1.95	4.18
6	1.51	1.53	1.35	22.47	1.85	1.57	1.97	4.09
7	1.62	1.67	1.46	7.78	1.86	1.55	2.00	4.10
8	1.55	1.54	1.44	11.58	1.84	1.55	1.96	4.25
9	1.46	1.47	**1.31**	**26.13**	1.87	1.60	1.97	4.12

Sums of SoftPlus and Sigmoid functions

Mean Squared Error Results on Call Option Pricing ($\times 10^{-4}$)

Units	Unconstrained weights				Constrained weights			
	Train	Valid	Test1	Test2	Train	Valid	Test1	Test2
1	1.83	1.59	1.93	4.10	2.30	2.19	2.36	3.43
2	1.42	1.45	**1.26**	**25.00**	2.29	2.19	2.34	3.39
3	1.45	1.46	1.32	35.00	1.84	1.58	1.95	4.11
4	1.56	1.69	1.33	21.80	1.85	1.56	1.99	4.09
5	1.60	1.69	1.42	10.11	1.85	1.52	**2.00**	**4.21**
6	1.57	1.66	1.39	14.99	1.86	1.54	2.00	4.12
7	1.61	1.67	1.48	8.00	1.86	1.60	1.98	3.94
8	1.64	1.72	1.48	7.89	1.85	1.54	1.98	4.25
9	1.65	1.70	1.52	6.16	1.84	1.54	1.97	4.25

Table 2: *Similar results as in table 1 but for two new architectures. Top: products of softplus along the convex axis with sigmoid along the monotone axis. Bottom: the softplus and sigmoid functions are summed instead of being multiplied. Top right: the fully constrained proposed architecture.*

networks that have those properties. We have shown universal properties for these classes, and we have shown that using this *a priori* knowledge can help in improving generalization performance. In particular, we have found that the models that incorporate this *a priori* knowledge generalize in a more stable way over time.

References

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

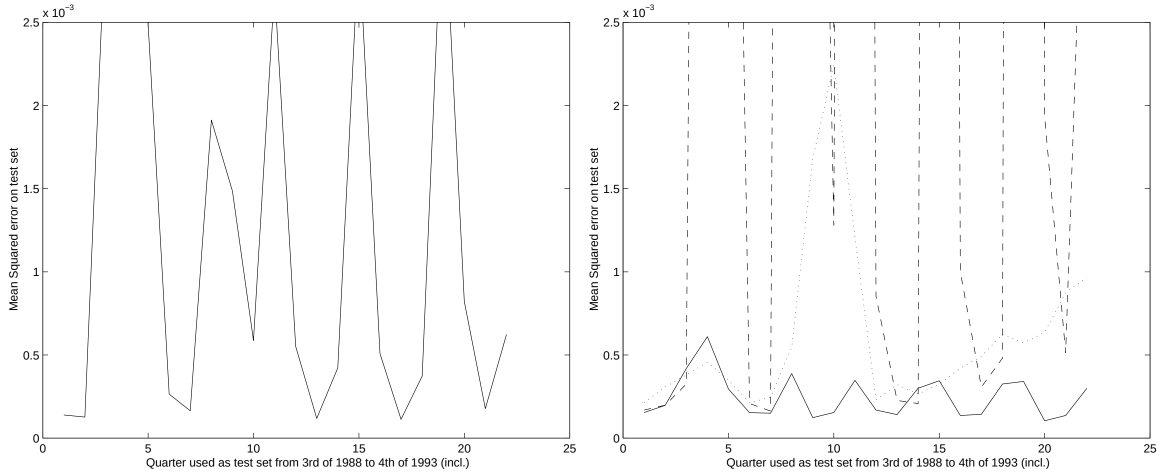

Figure 3: *Out-of-sample results from the third quarter of 1988 to the fourth of 1993 (incl.) for models with best validation results. Left: unconstrained models: results for the Black-Scholes similar network. Other unconstrained models exhibit similar swinging result patterns and levels of errors. Right: constrained models: the fully constrained proposed architecture (solid). The model with sums over dimensions obtains similar results. The regular neural network (dotted) is significantly worse. The constrained Black-Scholes model obtains very poor results (dashed).*

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

G. Cybenko. Continuous valued neural networks with two hidden layers are sufficient. Technical report, Department of Computer Science, Tufts University, Medford, MA, 1988.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

C. Dugas, O. Bardou, and Y. Bengio. Analyses empiriques sur des transactions d'options. Technical Report 1176, Départment d'informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec, Canada, 2000.

R. Garcia and R. Gençay. Pricing and Hedging Derivative Securities with Neural Networks and a Homogeneity Hint. Technical Report 98s-35, CIRANO, Montréal, Québec, Canada, 1998.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

J. Moody. Prediction risk and architecture selection for neural networks. In *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. Springer, 1994.

Liste des publications au CIRANO*

Série Scientifique / *Scientific Series* (ISSN 1198-8177)

* Consultez la liste complète des publications du CIRANO et les publications elles-mêmes sur notre site Internet :

http://www.cirano.qc.ca

2002s-28 The Causal Effect of High School Employment on Educational Attainment in Canada / Daniel Parent

2002s-27 Employer-Supported Training in Canada and Its Impact on Mobility and Wages / Daniel Parent

2002s-26 Restructuring and Economic Performance: The Experience of the Tunisian Economy / Sofiane Ghali and Pierre Mohnen

2002s-25 What Type of Enterprise Forges Close Links With Universities and Government Labs? Evidence From CIS 2 / Pierre Mohnen et Cathy Hoareau

2002s-24 Environmental Performance of Canadian Pulp and Paper Plants : Why Some Do Well and Others Do Not ? / Julie Doonan, Paul Lanoie et Benoit Laplante

2002s-23 A Rule-driven Approach for Defining the Behavior of Negotiating Software Agents / Morad Benyoucef, Hakim Alj, Kim Levy et Rudolf K. Keller

2002s-22 Occupational Gender Segregation and Women's Wages in Canada: An Historical Perspective / Nicole M. Fortin et Michael Huberman

2002s-21 Information Content of Volatility Forecasts at Medium-term Horizons / John W. Galbraith et Turgut Kisinbay

2002s-20 Earnings Dispersion, Risk Aversion and Education / Christian Belzil et Jörgen Hansen

2002s-19 Unobserved Ability and the Return to Schooling / Christian Belzil et Jörgen Hansen

2002s-18 Auditing Policies and Information Systems in Principal-Agent Analysis / Marie-Cécile Fagart et Bernard Sinclair-Desgagné

2002s-17 The Choice of Instruments for Environmental Policy: Liability or Regulation? / Marcel Boyer, Donatella Porrini

2002s-16 Asymmetric Information and Product Differentiation / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-15 Entry Preventing Locations Under Incomplete Information / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-14 On the Relationship Between Financial Status and Investment in Technological Flexibility / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-13 Modeling the Choice Between Regulation and Liability in Terms of Social Welfare / Marcel Boyer et Donatella Porrini

2002s-12 Observation, Flexibilité et Structures Technologiques des Industries / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-11 Idiosyncratic Consumption Risk and the Cross-Section of Asset Returns / Kris Jacobs et Kevin Q. Wang

2002s-10 The Demand for the Arts / Louis Lévy-Garboua et Claude Montmarquette

2002s-09 Relative Wealth, Status Seeking, and Catching Up / Ngo Van Long, Koji Shimomura

2002s-08 The Rate of Risk Aversion May Be Lower Than You Think / Kris Jacobs

2002s-07 A Structural Analysis of the Correlated Random Coefficient Wage Regression Model / Christian Belzil et Jörgen Hansen

2002s-06 Information Asymmetry, Insurance, and the Decision to Hospitalize / Åke Blomqvist et Pierre Thomas Léger

2002s-05 Coping with Stressful Decisions: Individual Differences, Appraisals and Choice / Ann-Renée Blais

2002s-04 A New Proof Of The Maximum Principle / Ngo Van Long et Koji Shimomura

2002s-03 Macro Surprises And Short-Term Behaviour In Bond Futures / Eugene Durenard et David Veredas

2002s-02 Financial Asset Returns, Market Timing, and Volatility Dynamics / Peter F. Christoffersen et Francis X. Diebold

2002s-01 An Empirical Analysis of Water Supply Contracts / Serge Garcia et Alban Thomas

2001s-71 A Theoretical Comparison Between Integrated and Realized Volatilities Modeling / Nour Meddahi

2001s-70 An Eigenfunction Approach for Volatility Modeling / Nour Meddahi

2001s-69 Dynamic Prevention in Short Term Insurance Contracts / M. Martin Boyer et Karine Gobert

2001s-68 Serial Cost Sharing in Multidimensional Contexts / Cyril Téjédo et Michel Truchon

2001s-67 Learning from Strike / Fabienne Tournadre et Marie-Claire Villeval

2001s-66 Incentives in Common Agency / Bernard Sinclair-Desgagné

2001s-65 Detecting Mutiple Breaks in Financial Market Volatility Dynamics / Elena Andreou et Eric Ghysels

2001s-64 Real Options, Preemption, and the Dynamics of Industry Investments / Marcel Boyer, Pierre Lasserre, Thomas Mariotti et Michel Moreaux

2001s-63 Dropout, School Performance and Working while in School: An Econometric Model with Heterogeneous Groups / Marcel Dagenais, Claude Montmarquette et Nathalie Viennot-Briot

2001s-62 Derivatives Do Affect Mutual Funds Returns : How and When? / Charles Cao, Eric Ghysels et Frank Hatheway

2001s-61 Conditional Quantiles of Volatility in Equity Index and Foreign Exchange Data / John W. Galbraith, Serguei Zernov and Victoria Zinde-Walsh

2001s-60 The Public-Private Sector Risk-Sharing in the French Insurance "Cat. Nat. System" / Nathalie de Marcellis-Warin et Erwann Michel-Kerjan

2001s-59 Compensation and Auditing with Correlated Information / M. Martin Boyer et Patrick González

2001s-58 Resistance is Futile: An Essay in Crime and Commitment / M. Martin Boyer

2001s-57 The Unreliability of Output Gap Estimates in Real Time / Athanasios Orphanides et Simon van Norden

2001s-56 Exact Nonparametric Two-Sample Homogeneity Tests for Possibly Discrete Distributions / Jean-Marie Dufour et Abdeljelil Farhat

2001s-55 Les coûts de la réglementation : une revue de la littérature / Robert Gagné, Paul Lanoie, Pierre-Carl Micheud et Michel Patry

2001s-54 Testing for structural Change in the Presence of Auxiliary Models / Eric Ghysels et Alain Guay

2001s-53 Environmental Regulation and Productivity: New Findings on the Porter Hypothesis / Paul Lanoie, Michel Patry et Richard Lajeunesse