Reinforcement Learning for Continuous Stochastic Control Problems

Rémi Munos
CEMAGREF, LISC, Parc de Tourvoie,
BP 121, 92185 Antony Cedex, FRANCE.
Remi.Munos@cemagref.fr

Paul Bourgine
Ecole Polytechnique, CREA,
91128 Palaiseau Cedex, FRANCE.
Bourgine@poly.polytechnique.fr

Abstract

This paper is concerned with the problem of Reinforcement Learning (RL) for continuous state space and time stochastic control problems. We state the Hamilton-Jacobi-Bellman equation satisfied by the value function and use a Finite-Difference method for designing a convergent approximation scheme. Then we propose a RL algorithm based on this scheme and prove its convergence to the optimal solution.

1 Introduction to RL in the continuous, stochastic case

The objective of RL is to find -thanks to a reinforcement signal- an optimal strategy for solving a dynamical control problem. Here we sudy the continuous time, continuous state-space stochastic case, which covers a wide variety of control problems including target, viability, optimization problems (see [FS93], [KP95]) for which a formalism is the following. The evolution of the *current state* $x(t) \in \bar{O}$ (the *state-space*, with O open subset of $\mathbb{R}^d$), depends on the *control* $u(t) \in U$ (compact subset) by a stochastic differential equation, called the *state dynamics:*

$$dx = f(x(t), u(t))dt + \sigma(x(t), u(t))dw \tag{1}$$

where f is the local drift and $\sigma.dw$ (with w a brownian motion of dimension r and σ a $d \times r$-matrix) the stochastic part (which appears for several reasons such as lake of precision, noisy influence, random fluctuations) of the diffusion process.

For initial state x and control $u(t)$, (1) leads to an infinity of possible trajectories $x(t)$. For some trajectory $x(t)$ (see figure 1), let τ be its *exit time* from $\bar{O}$ (with the convention that if $x(t)$ always stays in $\bar{O}$, then $\tau = \infty$). Then, we define the *functional* J of initial state x and control $u(.)$ as the expectation for all trajectories of the discounted cumulative reinforcement :

$$J(x; u(.)) = E_{x,u(.)} \left\{ \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau)) \right\}$$

where $r(x,u)$ is the *running reinforcement* and $R(x)$ the *boundary reinforcement*. γ is the *discount factor* ($0 \leq \gamma < 1$). In the following, we assume that f, σ are of class $\mathcal{C}^2$, r and R are Lipschitzian (with constants L_r and L_R) and the boundary ∂O is $\mathcal{C}^2$.

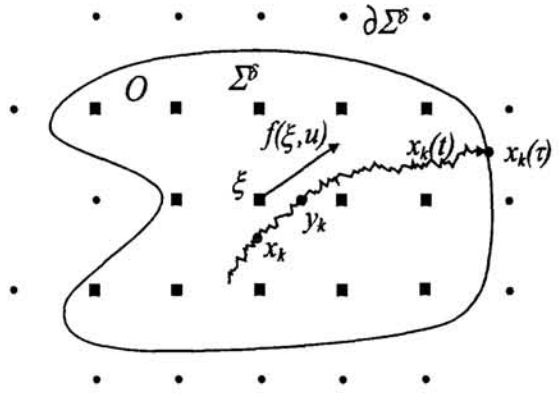

Figure 1: The state space, the discretized Σ^δ(the square dots) and its frontier $\partial\Sigma^\delta$ (the round ones). A trajectory $x_k(t)$ goes through the neighbourhood of state ξ.

RL uses the method of Dynamic Programming (DP) which generates an optimal (feed-back) control $u^*(x)$ by estimating the *value function* (VF), defined as the maximal value of the functional J as a function of initial state x :

$$V(x) = \sup_{u(.)} J(x; u(.)). \tag{2}$$

In the RL approach, the state dynamics is unknown from the system ; the only available information for learning the optimal control is the reinforcement obtained at the current state. Here we propose a model-based algorithm, i.e. that learns on-line a model of the dynamics and approximates the value function by successive iterations.

Section 2 states the Hamilton-Jacobi-Bellman equation and use a Finite-Difference (FD) method derived from Kushner [Kus90] for generating a convergent approximation scheme. In *section 3*, we propose a RL algorithm based on this scheme and prove its convergence to the VF in *appendix A*.

2 A Finite Difference scheme

Here, we state a second-order nonlinear differential equation (obtained from the DP principle, see [FS93]) satisfied by the value function, called the *Hamilton-Jacobi-Bellman* equation.

Let the $d \times d$ matrix $a = \sigma.\sigma'$ (with $'$ the transpose of the matrix). We consider the *uniformly parabolic* case, i.e. we assume that there exists $c > 0$ such that $\forall x \in \overline{O}, \forall u \in U, \forall y \in \mathbb{R}^d, \sum_{i,j=1}^d a_{ij}(x,u) y_i y_j \geq c||y||^2$. Then V is $\mathcal{C}^2$ (see [Kry80]). Let V_x be the gradient of V and $V_{x_i x_j}$ its second-order partial derivatives.

Theorem 1 (Hamilton-Jacobi-Bellman) *The following HJB equation holds :*

$$V(x) \ln \gamma + \sup_{u \in U} \left[r(x,u) + V_x(x).f(x,u) + \tfrac{1}{2} \sum_{i,j=1}^n a_{ij} V_{x_i x_j}(x) \right] = 0 \text{ for } x \in O$$

Besides, V satisfies the following boundary condition : $V(x) = R(x)$ for $x \in \partial O$.

Remark 1 *The challenge of learning the VF is motivated by the fact that from V, we can deduce the following optimal feed-back control policy:*

$$u^*(x) \in \arg\sup_{u \in U} \left[r(x,u) + V_x(x).f(x,u) + \tfrac{1}{2} \sum_{i,j=1}^{n} a_{ij} V_{x_i x_j}(x) \right]$$

In the following, we assume that O is bounded. Let $e_1, ..., e_d$ be a basis for $\mathbb{R}^d$. Let the positive and negative parts of a function ϕ be : $\phi^+ = \max(\phi, 0)$ and $\phi^- = \max(-\phi, 0)$. For any *discretization step* δ, let us consider the lattices : $\delta\mathbb{Z}^d = \left\{ \delta . \sum_{i=1}^{d} j_i e_i \right\}$ where $j_1, ..., j_d$ are any integers, and $\Sigma^\delta = \delta\mathbb{Z}^d \cap O$. Let $\partial\Sigma^\delta$, the *frontier* of Σ^δ denote the set of points $\{\xi \in \delta\mathbb{Z}^d \setminus O$ such that at least one adjacent point $\xi \pm \delta e_i \in \Sigma^\delta\}$ (see figure 1).

Let $U^\delta \subset U$ be a finite control set that approximates U in the sense: $\delta \leq \delta' \Rightarrow U^{\delta'} \subset U^\delta$ and $\overline{\cup_\delta U^\delta} = U$. Besides, we assume that: $\forall i = 1..d$,

$$a_{ii}(x,u) - \sum_{j \neq i} |a_{ij}(x,u)| \geq 0. \tag{3}$$

By replacing the gradient $V_x(\xi)$ by the forward and backward first-order finite-difference quotients: $\Delta_{x_i}^{\pm} V(\xi) = \frac{1}{\delta} [V(\xi \pm \delta e_i) - V(\xi)]$ and $V_{x_i x_j}(\xi)$ by the second-order finite-difference quotients:

$$\begin{aligned}
\Delta_{x_i x_i} V(\xi) &= \tfrac{1}{\delta^2} [V(\xi + \delta e_i) + V(\xi - \delta e_i) - 2V(\xi)] \\
\Delta_{x_i x_j}^{\pm} V(\xi) &= \tfrac{1}{2\delta^2} [V(\xi + \delta e_i \pm \delta e_j) + V(\xi - \delta e_i \mp \delta e_j) \\
&\quad - V(\xi + \delta e_i) - V(\xi - \delta e_i) - V(\xi + \delta e_j) - V(\xi - \delta e_j) + 2V(\xi)]
\end{aligned}$$

in the HJB equation, we obtain the following : for $\xi \in \Sigma^\delta$,

$$V^\delta(\xi) \ln\gamma + \sup_{u \in U^\delta} \left\{ r(\xi,u) + \sum_{i=1}^{d} \left[f_i^+(\xi,u).\Delta_{x_i}^+ V^\delta(\xi) - f_i^-(\xi,u).\Delta_{x_i}^- V^\delta(\xi) \right.\right.$$
$$\left.\left. + \tfrac{a_{ii}(\xi,u)}{2} \Delta_{x_i x_i} V(\xi) + \sum_{j \neq i} \left(\tfrac{a_{ij}^+(\xi,u)}{2} \Delta_{x_i x_j}^+ V(\xi) - \tfrac{a_{ij}^-(\xi,u)}{2} \Delta_{x_i x_j}^- V(\xi) \right) \right] \right\} = 0$$

Knowing that $(\Delta t \ln\gamma)$ is an approximation of $(\gamma^{\Delta t} - 1)$ as Δt tends to 0, we deduce:

$$V^\delta(\xi) = \sup_{u \in U^\delta} \left[\gamma^{\tau(\xi,u)} \sum_{\zeta \in \Sigma^\delta} p(\xi,u,\zeta) V^\delta(\zeta) + \tau(\xi,u) r(\xi,u) \right] \tag{4}$$

$$\text{with } \tau(\xi,u) = \frac{\delta^2}{\sum_{i=1}^{d} \left[\delta |f_i(\xi,u)| + a_{ii}(\xi,u) - \frac{1}{2} \sum_{j \neq i} |a_{ij}(\xi,u)| \right]} \tag{5}$$

which appears as a DP equation for some finite Markovian Decision Process (see [Ber87]) whose state space is Σ^δ and probabilities of transition :

$$\begin{aligned}
p(\xi,u,\xi \pm \delta e_i) &= \tfrac{\tau(\xi,u)}{2\delta^2} \left[2\delta |f_i^\pm(\xi,u)| + a_{ii}(\xi,u) - \sum_{j \neq i} |a_{ij}(\xi,u)| \right], \\
p(\xi,u,\xi + \delta e_i \pm \delta e_j) &= \tfrac{\tau(\xi,u)}{2\delta^2} a_{ij}^\pm(\xi,u) \text{ for } i \neq j, \\
p(\xi,u,\xi - \delta e_i \pm \delta e_j) &= \tfrac{\tau(\xi,u)}{2\delta^2} a_{ij}^\mp(\xi,u) \text{ for } i \neq j, \\
p(\xi,u,\zeta) &= 0 \text{ otherwise.}
\end{aligned} \tag{6}$$

Thanks to a contraction property due to the discount factor γ, there exists a unique solution (the fixed-point) V^δ to equation (4) for $\xi \in \Sigma^\delta$ with the boundary condition $V^\delta(\xi) = R(\xi)$ for $\xi \in \partial\Sigma^\delta$. The following theorem (see [Kus90] or [FS93]) insures that V^δ is a convergent approximation scheme.

Theorem 2 (Convergence of the FD scheme) *V^δ converges to V as $\delta \downarrow 0$:*

$$\lim_{\substack{\delta \downarrow 0 \\ \xi \to x}} V^\delta(\xi) = V(x) \text{ uniformly on } \overline{O}$$

Remark 2 *Condition (3) insures that the $p(\xi, u, \zeta)$ are positive. If this condition does not hold, several possibilities to overcome this are described in [Kus90].*

3 The reinforcement learning algorithm

Here we assume that f is bounded from below. As the state dynamics (f and a) is unknown from the system, we approximate it by building a *model* $\widetilde{f}$ and $\widetilde{a}$ from samples of trajectories $x_k(t)$: we consider series of successive states $x_k = x_k(t_k)$ and $y_k = x_k(t_k + \tau_k)$ such that:

- $\forall t \in [t_k, t_k + \tau_k]$, $x(t) \in N(\xi)$ neighbourhood of ξ whose diameter is inferior to $k_N.\delta$ for some positive constant k_N,

- the control u is constant for $t \in [t_k, t_k + \tau_k]$,

- τ_k satisfies for some positive k_1 and k_2,

$$k_1 \delta \leq \tau_k \leq k_2 \delta. \tag{7}$$

Then incrementally update the model :

$$\begin{aligned}
\widetilde{f_n}(\xi, u) &= \tfrac{1}{n} \sum_{k=1}^{n} \frac{y_k - x_k}{\tau_k} \\
\widetilde{a_n}(\xi, u) &= \tfrac{1}{n} \sum_{k=1}^{n} \frac{\left(y_k - x_k - \tau_k.\widetilde{f_n}(\xi, u)\right)\left(y_k - x_k - \tau_k.\widetilde{f_n}(\xi, u)\right)'}{\tau_k} \\
\widetilde{r}(\xi, u) &= \tfrac{1}{n} \sum_{k=1}^{n} r(x_k, u)
\end{aligned} \tag{8}$$

and compute the approximated time $\widetilde{\tau}(x, u)$ and the approximated probabilities of transition $\widetilde{p}(\xi, u, \zeta)$ by replacing f and a by $\widetilde{f}$ and $\widetilde{a}$ in (5) and (6).

We obtain the following updating rule of the V^δ-value of state ξ:

$$V^\delta_{n+1}(\xi) = \sup_{u \in U^\delta} \left[\gamma^{\widetilde{\tau}(x,u)} \sum_\zeta \widetilde{p}(\xi, u, \zeta) V^\delta_n(\zeta) + \widetilde{\tau}(x, u)\widetilde{r}(\xi, u)\right] \tag{9}$$

which can be used as an off-line (synchronous, Gauss-Seidel, asynchronous) or on-time (for example by updating $V^\delta_n(\xi)$ as soon as a trajectory exits from the neighbourood of ξ) DP algorithm (see [BBS95]).

Besides, when a trajectory hits the boundary ∂O at some exit point $x_k(\tau)$ then update the closest state $\xi \in \partial\Sigma^\delta$ with:

$$V^\delta_n(\xi) = R(x_k(\tau)) \tag{10}$$

Theorem 3 (Convergence of the algorithm) *Suppose that the model as well as the V^δ-value of every state $\xi \in \Sigma^\delta$ and control $u \in U^\delta$ are regularly updated (respectively with (8) and (9)) and that every state $\xi \in \partial\Sigma^\delta$ are updated with (10) at least once. Then $\forall \varepsilon > 0$, $\exists \Delta$ such that $\forall \delta \leq \Delta$, $\exists N, \forall n \geq N$,*

$$\sup_{\xi \in \Sigma^\delta} |V^\delta_n(\xi) - V(\xi)| \leq \varepsilon \text{ with probability 1}$$

4 Conclusion

This paper presents a model-based RL algorithm for continuous stochastic control problems. A model of the dynamics is approximated by the mean and the covariance of successive states. Then, a RL updating rule based on a convergent FD scheme is deduced and in the hypothesis of an adequate exploration, the convergence to the optimal solution is proved as the discretization step δ tends to 0 and the number of iteration tends to infinity. This result is to be compared to the model-free RL algorithm for the deterministic case in [Mun97]. An interesting possible future work should be to consider model-free algorithms in the stochastic case for which a Q-learning rule (see [Wat89]) could be relevant.

A Appendix: proof of the convergence

Let M_f, M_a, M_{f_x} and M_{σ_x} be the upper bounds of f, a, f_x and σ_x and m_f the lower bound of f. Let $E^\delta = \sup_{\xi\in\Sigma^\delta} |V^\delta(\xi) - V(\xi)|$ and $E_n^\delta = \sup_{\xi\in\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right|$.

A.1 Estimation error of the model $\widetilde{f_n}$ and $\widetilde{a_n}$ and the probabilities $\widetilde{p}_n$

Suppose that the trajectory $x_k(t)$ occured for some occurence $w_k(t)$ of the brownian motion: $x_k(t) = x_k + \int_{t_k}^t f(x_k(t),u)dt + \int_{t_k}^t \sigma(x_k(t),u)dw_k$. Then we consider a trajectory $z_k(t)$ starting from ξ at t_k and following the same brownian motion: $z_k(t) = \xi + \int_{t_k}^t f(z_k(t),u)dt + \int_{t_k}^t \sigma(z_k(t),u)dw_k$.

Let $z_k = z_k(t_k+\tau_k)$. Then $(y_k - x_k) - (z_k - \xi) = \int_{t_k} [f(x_k(t),u) - f(z_k(t),u)]\, dt + \int_{t_k}^{t_k+\tau_k} [\sigma(x_k(t),u) - \sigma(z_k(t),u)]\, dw_k$. Thus, from the $\mathcal{C}^1$ property of f and σ,

$$\|(y_k - x_k) - (z_k - \xi)\| \leq (M_{f_x} + M_{\sigma_x}).k_N.\tau_k.\delta. \tag{11}$$

The diffusion processes has the following property (see for example the Itô-Taylor majoration in [KP95]) : $E_x[z_k] = \xi + \tau_k.f(\xi,u) + O(\tau_k^2)$ which, from (7), is equivalent to : $E_x\left[\frac{z_k-\xi}{\tau_k}\right] = f(\xi,u) + O(\delta)$. Thus from the law of large numbers and (11):

$$\begin{aligned}\limsup_{n\to\infty} \left\|\widetilde{f_n}(\xi,u) - f(\xi,u)\right\| &= \limsup_{n\to\infty} \left\|\tfrac{1}{n}\textstyle\sum_{k=1}^n \left[\frac{y_k-x_k}{\tau_k} - \frac{z_k-\xi}{\tau_k}\right]\right\| + O(\delta) \\ &= (M_{f_x} + M_{\sigma_x}).k_N.\delta + O(\delta) = O(\delta) \text{ w.p. } 1 \end{aligned} \tag{12}$$

Besides, diffusion processes have the following property (again see [KP95]): $E_x\left[(z_k-\xi)(z_k-\xi)'\right] = a(\xi,u)\tau_k + f(\xi,u).f(\xi,u)'.\tau_k^2 + O(\tau_k^3)$ which, from (7), is equivalent to: $E_x\left[\frac{(z_k-\xi-\tau_k f(\xi,u))(z_k-\xi-\tau_k f(\xi,u))'}{\tau_k}\right] = a(\xi,u) + O(\delta^2)$. Let $r_k = z_k - \xi - \tau_k f(\xi,u)$ and $\widetilde{r_k} = y_k - x_k - \tau_k \widetilde{f_n}(\xi,u)$ which satisfy (from (11) and (12)) :

$$\|r_k - \widetilde{r_k}\| = (M_{f_x} + M_{\sigma_x}).\tau_k.k_N.\delta + \tau_k.O(\delta) \tag{13}$$

From the definition of $\widetilde{a_n}(\xi,u)$, we have: $\widetilde{a_n}(\xi,u) - a(\xi,u) = \frac{1}{n}\sum_{k=1}^n \frac{\widetilde{r_k}.\widetilde{r_k}'}{\tau_k} - E_x\left[\frac{r_k.r_k'}{\tau_k}\right] + O(\delta^2)$ and from the law of large numbers, (12) and (13), we have:

$$\begin{aligned}\limsup_{n\to\infty} \quad & \|\widetilde{a_n}(\xi,u) - a(\xi,u)\| = \limsup_{n\to\infty} \left\|\tfrac{1}{n}\textstyle\sum_{k=1}^n \frac{\widetilde{r_k}.\widetilde{r_k}'}{\tau_k} - \frac{r_k.r_k'}{\tau_k}\right\| + O(\delta^2) \\ = \quad & \|\widetilde{r_k} - r_k\| \limsup_{n\to\infty} \tfrac{1}{n} \sum_{k=1}^n \left(\left\|\frac{\widetilde{r_k}}{\tau_k}\right\| + \left\|\frac{r_k}{\tau_k}\right\|\right) + O(\delta^2) = O(\delta^2)\end{aligned}$$

with probability 1. Thus there exists k_f and k_a s.t. $\exists \Delta_1, \forall \delta \leq \Delta_1, \exists N_1, n \geq N_1$,

$$\begin{array}{ll} \left\| \widetilde{f_n}(\xi,u) - f(\xi,u) \right\| & \leq k_f.\delta \text{ w.p. } 1 \\ \left\| \widetilde{a_n}(\xi,u) - a(\xi,u) \right\| & \leq k_a.\delta^2 \text{ w.p. } 1 \end{array} \tag{14}$$

Besides, from (5) and (14), we have:

$$|\tau(\xi,u) - \widetilde{\tau}_n(\xi,u)| \leq \frac{d.(k_f.\delta^2 + d.k_a\delta^2)}{(d.m_f.\delta)^2}\delta^2 \leq k_\tau.\delta^2 \tag{15}$$

and from a property of exponential function,

$$\left| \gamma^{\tau(\xi,u)} - \gamma^{\widetilde{\tau}_n(\xi,u)} \right| = k_\tau.\ln\tfrac{1}{\gamma}.\delta^2. \tag{16}$$

We can deduce from (14) that:

$$\limsup_{n\to\infty} |p(\xi,u,\zeta) - \widetilde{p_n}(\xi,u,\zeta)| \leq \frac{(2.\delta.M_f + d.M_a)(2.k_f + d.k_a)\delta^2}{\delta m_f - (2.k_f + d.k_a)\delta^2} \leq k_p\delta \text{ w.p. } 1 \tag{17}$$

with $k_p = 4(d.M_a)(2.k_f + d.k_a)$ for $\delta \leq \Delta_2 = \min\left\{ \frac{m_f}{2.k_f + d.k_a}, \frac{d.M_a}{2.\delta.M_f} \right\}$.

A.2 Estimation of $|V^\delta_{n+1}(\xi) - V^\delta(\xi)|$

After having updated $V^\delta_n(\xi)$ with rule (9), let Λ denote the difference $\left|V^\delta_{n+1}(\xi) - V^\delta(\xi)\right|$. From (4), (9) and (8),

$$\begin{aligned} \Lambda \leq\ & \gamma^{\tau(\xi,u)} \sum_\zeta \left[p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)\right] V^\delta(\zeta) + \left(\gamma^{\tau(\xi,u)} - \gamma^{\widetilde{\tau}(\xi,u)}\right) \sum_\zeta \widetilde{p}(\xi,u,\zeta) V^\delta(\zeta) \\ & + \gamma^{\widetilde{\tau}(\xi,u)}.\sum_\zeta \widetilde{p}(\xi,u,\zeta)\left[V^\delta(\zeta) - V^\delta_n(\zeta)\right] + \sum_\zeta \widetilde{p}(\xi,u,\zeta).\widetilde{\tau}(\xi,u)\left[r(\xi,u) - \widetilde{r}(\xi,u)\right] \\ & + \textstyle\sum_\zeta \widetilde{p}(\xi,u,\zeta)\left[\widetilde{\tau}(\xi,u) - \tau(\xi,u)\right] r(\xi,u) \text{ for all } u \in U^\delta \end{aligned}$$

As V is differentiable we have : $V(\zeta) = V(\xi) + V_x.(\zeta - \xi) + o(\|\zeta - \xi\|)$. Let us define a linear function $\widetilde{V}$ such that: $\widetilde{V}(x) = V(\xi) + V_x.(x - \xi)$. Then we have: $[p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)]\, V^\delta(\zeta) = [p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)].\left[V^\delta(\zeta) - V(\zeta)\right] + [p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)]\, V(\zeta)$, thus: $\sum_\zeta [p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)]\, V^\delta(\zeta) = k_p.E^\delta.\delta + \sum_\zeta [p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)]\left[\widetilde{V}(\zeta) + o(\delta)\right] = \left[\widetilde{V}(\eta) - \widetilde{V}(\widetilde{\eta})\right] + k_p.E^\delta.\delta + o(\delta) = \left[\widetilde{V}(\eta) - \widetilde{V}(\widetilde{\eta})\right] + o(\delta)$ with: $\eta = \sum_\zeta p(\xi,u,\zeta)\,(\zeta - \xi)$ and $\widetilde{\eta} = \sum_\zeta \widetilde{p}(\xi,u,\zeta)\,(\zeta - \xi)$. Besides, from the convergence of the scheme (theorem 2), we have $E^\delta.\delta = o(\delta)$. From the linearity of $\widetilde{V}$, $\left|\widetilde{V}(\zeta) - \widetilde{V}(\widetilde{\zeta})\right| \leq \left\|\zeta - \widetilde{\zeta}\right\|.M_{V_x} \leq 2k_p\delta^2$. Thus $\left|\sum_\zeta [p(\xi,u,\zeta) - \widetilde{p}(\xi,u,\zeta)]\, V^\delta(\zeta)\right| = o(\delta)$ and from (15), (16) and the Lipschitz property of r,

$$\Lambda = \left|\gamma^{\widetilde{\tau}(\xi,u)}.\textstyle\sum_\zeta \widetilde{p}(\xi,u,\zeta)\left[V^\delta(\zeta) - V^\delta_n(\zeta)\right]\right| + o(\delta).$$

As $\gamma^{\widetilde{\tau}(\xi,u)} \leq 1 - \frac{\widetilde{\tau}(\xi,u)}{2}\ln\frac{1}{\gamma} \leq 1 - \frac{\tau(\xi,u) - k_\tau\delta^2}{2}\ln\frac{1}{\gamma} \leq 1 - \left(\frac{\delta}{2d(M_f + d.M_a)} - \frac{k_\tau}{2}\delta^2\right)\ln\frac{1}{\gamma}$, we have:

$$\Lambda = (1 - k.\delta)E^\delta_n + o(\delta) \tag{18}$$

with $k = \frac{1}{2d(M_f + d.M_a)}$.

A.3 A sufficient condition for $\sup_{\xi\in\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right| \leq \varepsilon_2$

Let us suppose that for all $\xi \in \Sigma^\delta$, the following conditions hold for some $\alpha > 0$

$$E_n^\delta > \varepsilon_2 \Rightarrow \left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq E_n^\delta - \alpha \quad (19)$$

$$E_n^\delta \leq \varepsilon_2 \Rightarrow \left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq \varepsilon_2 \quad (20)$$

From the hypothesis that all states $\xi \in \Sigma^\delta$ are regularly updated, there exists an integer m such that at stage $n+m$ all the $\xi \in \Sigma^\delta$ have been updated at least once since stage n. Besides, since all $\xi \in \partial G^\delta$ are updated at least once with rule (10), $\forall \xi \in \partial G^\delta, |V_n^\delta(\xi) - V^\delta(\xi)| = |R(x_k(\tau)) - R(\xi)| \leq 2.L_R.\delta \leq \varepsilon_2$ for any $\delta \leq \Delta_3 = \frac{\varepsilon_2}{2.L_R}$. Thus, from (19) and (20) we have:

$$E_n^\delta > \varepsilon_2 \Rightarrow E_{n+m}^\delta \leq E_n^\delta - \alpha$$

$$E_n^\delta \leq \varepsilon_2 \Rightarrow E_{n+m}^\delta \leq \varepsilon_2$$

Thus there exists N such that : $\forall n \geq N$, $E_n^\delta \leq \varepsilon_2$.

A.4 Convergence of the algorithm

Let us prove theorem 3. For any $\varepsilon > 0$, let us consider $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$. Assume $E_n^\delta > \varepsilon_2$, then from (18), $\Lambda = E_n^\delta - k.\delta.\varepsilon_2 + o(\delta) \leq E_n^\delta - k.\delta.\frac{\varepsilon_2}{2}$ for $\delta \leq \Delta_3$. Thus (19) holds for $\alpha = k.\delta.\frac{\varepsilon_2}{2}$. Suppose now that $E_n^\delta \leq \varepsilon_2$. From (18), $\Lambda \leq (1 - k.\delta)\varepsilon_2 + o(\delta) \leq \varepsilon_2$ for $\delta \leq \Delta_3$ and condition (20) is true.

Thus for $\delta \leq \min\{\Delta_1, \Delta_2, \Delta_3\}$, the sufficient conditions (19) and (20) are satisfied. So there exists N, for all $n \geq N$, $E_n^\delta \leq \varepsilon_2$. Besides, from the convergence of the scheme (theorem 2), there exists Δ_0 st. $\forall \delta \leq \Delta_0, \sup_{\xi\in\Sigma^\delta} |V^\delta(\xi) - V(\xi)| \leq \varepsilon_1$.

Thus for $\delta \leq \min\{\Delta_0, \Delta_1, \Delta_2, \Delta_3\}$, $\exists N$, $\forall n \geq N$,

$$\sup_{\xi\in\Sigma^\delta} |V_n^\delta(\xi) - V(\xi)| \leq \sup_{\xi\in\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| + \sup_{\xi\in\Sigma^\delta} |V^\delta(\xi) - V(\xi)| \leq \varepsilon_1 + \varepsilon_2 = \varepsilon.$$

References

[BBS95] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, (72):81–138, 1995.

[Ber87] Dimitri P. Bertsekas. *Dynamic Programming : Deterministic and Stochastic Models.* Prentice Hall, 1987.

[FS93] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions.* Applications of Mathematics. Springer-Verlag, 1993.

[KP95] Peter E. Kloeden and Eckhard Platen. *Numerical Solutions of Stochastic Differential Equations.* Springer-Verlag, 1995.

[Kry80] N.V. Krylov. *Controlled Diffusion Processes.* Springer-Verlag, New York, 1980.

[Kus90] Harold J. Kushner. Numerical methods for stochastic control problems in continuous time. *SIAM J. Control and Optimization*, 28:999–1048, 1990.

[Mun97] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case based on a finite difference method. *International Joint Conference on Artificial Intelligence*, 1997.

[Wat89] Christopher J.C.H. Watkins. *Learning from delayed reward.* PhD thesis, Cambridge University, 1989.

Use of a Multi-Layer Perceptron to Predict Malignancy in Ovarian Tumors

Herman Verrelst,
Yves Moreau and Joos Vandewalle
Dept. of Electrical Engineering
Katholieke Universiteit Leuven
Kard. Mercierlaan 94
B-3000 Leuven, Belgium

Dirk Timmerman
Dept. of Obst. and Gynaec.
University Hospitals Leuven
Herestraat 49
B-3000 Leuven, Belgium

Abstract

We discuss the development of a Multi-Layer Perceptron neural network classifier for use in preoperative differentiation between benign and malignant ovarian tumors. As the Mean Squared classification Error is not sufficient to make correct and objective assessments about the performance of the neural classifier, the concepts of sensitivity and specificity are introduced and combined in Receiver Operating Characteristic curves. Based on objective observations such as sonomorphologic criteria, color Doppler imaging and results from serum tumor markers, the neural network is able to make reliable predictions with a discriminating performance comparable to that of experienced gynecologists.

1 Introduction

A reliable test for preoperative discrimination between benign and malignant ovarian tumors would be of considerable help to clinicians. It would assist them to select patients for whom minimally invasive surgery or conservative management suffices versus those for whom urgent referral to a gynecologic oncologist is needed.

We discuss the development of a neural network classifier/diagnostic tool. The neural network was trained by supervised learning, based on data from 191 thoroughly examined patients presenting with ovarian tumors of which 140 were benign and 51 malignant. As inputs to the network we chose indicators that in recent studies have proven their high predictive value [1, 2, 3]. Moreover, we gave preference to those indicators that can be obtained in an objective way by any gynecologist. Some of these indicators have already been used in attempts to make one single protocol or decision algorithm [3, 4].

In order to make reliable assessments on the practical performance of the classifier, it is necessary to work with other concepts than Mean Squared classification Error (MSE), which is traditionally used as a measure of goodness in the training of a neural network. We will introduce notions as specificity and sensitivity and combine them into Receiver Operating Characteristic (ROC) curves. The use of ROC-curves is motivated by the fact that they are independent of the relative proportion of the various output classes in the sample population. This enables an objective validation of the performance of the classifier. We will also show how, in the training of the neural network, MSE optimization with gradient methods can be refined and/or replaced with the help of ROC-curves and simulated annealing techniques.

The paper is organized as follows. In Section 2 we give a brief description of the selected input features. In Section 3 we state some drawbacks to the MSE criterion and introduce the concepts of sensitivity, specificity and ROC-curves. Section 4 then deals with the technicalities of training the neural network. In Section 5 we show the results and compare them to human performance.

2 Data acquisition and feature selection

The data were derived from a study group of 191 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 to August 1996. Table 1 lists the different indicators which were considered, together with their mean value and standard deviations or together with the relative presence in cases of benign and malignant tumors.

Table 1	**Indicator**	**Benign**	**Malignant**
Demographic	Age	49.3 ± 16.0	58.3 ± 14.3
	Postmenopausal	40%	70.6%
Serum marker	CA 125 (log)	2.8 ± 1.1	5.2 ± 1.9
CDI	Blood flow present	72.9%	100%
Morphologic	Abdominal fluid	12.1%	52.9%
	Bilateral mass	11.4%	35.3%
	Unilocular cyst	42.1%	5.9%
	Multiloc/solid cyst	16.4%	49.0%
	Smooth wall	58.6%	2.0%
	Irregular wall	32.1%	76.5%
	Papillations	7.9%	74.5%

Table 1: Demographic, serum marker, color Doppler imaging and morphologic indicators. For the continuous valued features the mean and standard deviation for each class are reported. For binary valued indicators, the last two columns give the presence of the feature in both classes e.g. only 2% of malignant tumors had smooth walls.

First, all patients were scanned with ultrasonography to obtain detailed gray-scale images of the tumors. Every tumor was extensively examined for its morphologic characteristics. Table 1 lists the selected morphologic features: presence of abdominal fluid collection, papillary structures (> 3mm), smooth internal walls, wall irregularities, whether the cysts were unilocular, multilocular-solid and/or present on both pelvic sides. All outcomes are binary valued: every observation relates to the presence (1) or absence (0) of these characteristics.

Secondly, all tumors were entirely surveyed by color Doppler imaging to detect presence or absence of blood flow within the septa, cyst walls, solid tumor areas or ovarian tissue. The outcome is also binary valued (1/0).

Thirdly, in 173 out of the total of 191 patients, serum CA 125 levels were measured, using CA 125 II immunoradiometric assays (Centocor, Malvern, PA). The CA 125 antigen is a glycoprotein that is expressed by most epithelial ovarian cancers. The numerical value gives the concentration in U/ml. Because almost all values were situated in a small interval between 0 and 100, and because a small portion took values up to 30,000, this variable was rescaled by taking its logarithm.

Since age and menopausal status of the patient are considered to be highly relevant, these are also included. The menopausal score is -1 for premenopausal, $+1$ for postmenopausal. A third class of patients were assigned a 0 value. These patients had had an hysterectomy, so no menopausal status could be appointed to them.

It is beyond the scope of this paper to give a complete account of the meaning of the different features that are used or the way in which the data were acquired. We will limit ourselves to this short description and refer the reader to [2, 3] and gynecological textbooks for a more detailed explanation.

3 Receiver Operating Characteristics

3.1 Drawbacks to Mean Squared classification Error

Let us assume that we use a one-hidden-layer feed-forward NN with m inputs x_k^i, n_h hidden neurons with the tanh(.) as activation function, and one output $\hat{y}_k$,

$$y_k(\theta) = \sum_{j=1}^{n_h} w_j \tanh(\sum_{i=1}^{m} v_{ij} x_k^i + \beta_j), \tag{1}$$

parameterized by the vector θ consisting of the network's weights w_j and v_{ij} and bias terms β_j. The cost function is often chosen to be the squared difference between the desired d_k and the actual response y_k, averaged over all N samples [12],

$$J(\theta) = \frac{1}{N} \sum_{k=1}^{N} (d_k - y_k(\theta))^2. \tag{2}$$

This type of cost function is continuous and differentiable, so it can be used in gradient based optimization techniques such as steepest descent (back-propagation), quasi-Newton or Levenberg-Marquardt methods [8, 9, 11, 12]. However there are some drawbacks to the use of this type of cost function.

First of all, the MSE is heavily dependent on the relative proportion of the different output classes in the training set. In our dichotomic case this can easily be demonstrated by writing the cost function, with superscripts b and m respectively meaning benign and malignant, as

$$J(\theta) = \underbrace{\frac{N_b}{N_b + N_m}}_{\lambda} \frac{1}{N_b} \sum_{k=1}^{N_b} (d_k^b - y_k)^2 + \underbrace{\frac{N_m}{N_b + N_m}}_{(1-\lambda)} \frac{1}{N_m} \sum_{k=1}^{N_m} (d_k^m - y_k)^2 \tag{3}$$

If the relative proportion in the sample population is not representative for reality, the λ parameter should be adjusted accordingly. In practice this real proportion is often not known accurately or one simply ignores the meaning of λ and uses it as a design parameter in order to bias the accuracy towards one of the output classes.

A second drawback of the MSE cost function is that it is not very informative towards practical usage of the classification tool. A clinician is not interested in the averaged deviation from desired numbers, but thinks in terms of percentages found, missed or misclassified. In the next section we will introduce the concepts of sensitivity and specificity to express these more practical measures.

3.2 Sensitivity, specificity and ROC-curves

If we take the desired response to be 0 for benign and 1 for malignant cases, the way to make clearcut (dichotomic) decisions is to compare the numerical outcome of the neural network to a certain threshold value T between 0 and 1. When the outcome is above the threshold T, the prediction is said to be *positive*. Otherwise the prediction is said to be *negative*. With this convention, we say that the prediction was

True Positive (TP)	if the prediction was positive when the sample was malignant.
True Negative (TN)	if the prediction was negative when the sample was benign.
False Positive (FP)	if the prediction was positive when the sample was benign.
False Negative (FN)	if the prediction was negative when the sample was malignant.

To every of the just defined terms TP, TN, FP and FN, a certain subregion of the total sample space can be associated, as depicted in Figure 1. In the same sense, we can associate to them a certain number counting the samples in each subregion. We can then define sensitivity as $\frac{TP}{TP+FN}$, the proportion of malignant cases that

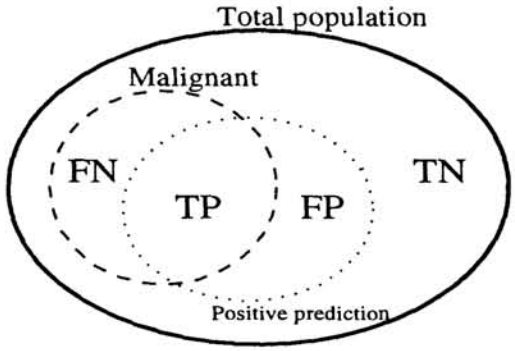

Figure 1: The concepts of true and false positive and negative illustrated. The dashed area indicates the malignant cases in the total sample population. The positive prediction of an imperfect classification (dotted area) does not fully coincide with this sub area.

are predicted to be malignant and specificity as $\frac{TN}{FP+TN}$, the proportion of benign cases that are predicted to be benign. The false positive rate is 1−specificity.

When varying the threshold T, the values of TP, TN, FP, FN and therefore also sensitivity and specificity, will change. A low threshold will detect almost all malignant cases at the cost of many false positives. A high threshold will give less false positives, but will also detect less malignant cases. Receiver Operating Characteristic (ROC) curves are a way to visualize this relationship. The plot gives the sensitivity versus false positive rate for varying thresholds T (e.g. Figure 2).

The ROC-curve is useful and widely used device for assessing and comparing the value of tests [5, 7]. The proportion of the whole area of the graph which lies below the ROC-curve is a one-value measure of the accuracy of a test [6]. The higher this proportion, the better the test. Figure 2 shows the ROC-curves for two simple classifiers that use only one single indicator. (Which means that we classify a tumor being malignant when the value of the indicator rises above a certain value.) It is seen that the CA 125 level has high predictive power as its ROC-curve spans 87.5% of the total area (left Figure 2). For the age parameter, the ROC-curve spans only 65.6% (right Figure 2). As indicated by the horizontal line in the plot, a CA 125 level classification will only misclassify 15% of all benign cases to reach a 80% sensitivity, whereas using only age, one would then misclassify up to 50% of them.

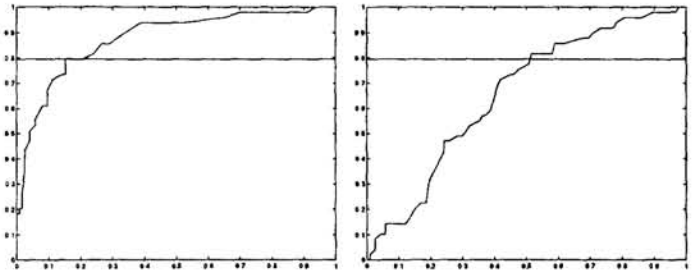

Figure 2: The Receiver Operating Characteristic (ROC) curve is the plot of the sensitivity versus the false positive rate of a classifier for varying thresholds used. Only single indicators (left: CA 125, right: age) are used for these ROC-curves. The horizontal line marks the 80% specificity level.

Since for every set of parameters of the neural network the area under the ROC-curve can be calculated numerically, this one-value measure can also be used for supervised training, as will be shown in the next Section.

4 Simulation results

4.1 Inputs and architecture

The continuous inputs were standardized by subtracting their mean and dividing by their standard deviation (both calculated over the entire population). Binary valued inputs were left unchanged. The desired outputs were labeled 0 for benign examples, 1 for malignant cases. The data set was split up: 2/3 of both benign and malignant samples were randomly selected to form the training set. The remaining examples formed the test set. The ratio of benign to all examples is $\lambda \approx \frac{2}{3}$.

Since the training set is not large, there is a risk of overtraining when too many parameters are used. We will limit the number of hidden neurons to $n_h = 3$ or 5. As the CA 125 level measurement is more expensive and time consuming, we will investigate two different classifiers: one which does use the CA 125 level and one which does not. The one-hidden-layer MLP architectures that are used, are 11-3-1 and 10-5-1. A tanh(.) is taken for the activation function in the hidden layer.

4.2 Training

A first way of training was MSE optimization using the cost function (3). By taking $\lambda = \frac{1}{3}$ in this expression, the role of malignant examples is more heavily weighted. The parameter vector θ was randomly initialized (zero mean Gaussian distribution, standard deviation $\sigma = 0.01$). Training was done using a quasi-Newton method with BFGS-update of the Hessian (*fminu* in Matlab) [8, 9]. To prevent overtraining, the training was stopped before the MSE on the test set started to rise. Only few iterations (≈ 100) were needed.

A second way of training was through the use of the area spanned by the ROC-curve of the classifier and simulated annealing techniques [10]. The area-measure A^{ROC} was numerically calculated for every set of trial parameters: first the sensitivity and false positive rate were calculated for 1000 increasing values of the threshold T between 0 and 1, which gave the ROC-curve; secondly the area A^{ROC} under the curve was numerically calculated with the trapezoidal integration rule.

We used Boltzmann Simulated Annealing to maximize the ROC-area. At time k a trial parameter set of the neural network θ_{k+1} is randomly generated in the neighborhood of the present set θ_k (Gaussian distribution, $\sigma = 0.001$). The trial set θ_{k+1} is always accepted if the area $A_{k+1}^{ROC} \geq A_k^{ROC}$. If $A_{k+1}^{ROC} < A_k^{ROC}$, θ_{k+1} is accepted if

$$e^{(\frac{A_{k+1}^{ROC} - A_k^{ROC}}{A_k^{ROC}})/T_e} > \alpha$$

with α a uniformly distributed random variable $\in [0, 1]$ and T_e the temperature. As cooling schedule we took

$$T_e = 1/(100 + 10k),$$

so that the annealing was low-temperature and fast-cooling. The optimization was stopped before the ROC-area calculated for the test set started to decrease. Only a few hundred annealing epochs were allowed.

4.3 Results

Table 2 states the results for the different approaches. One can see that adding the CA 125 serum level clearly improves the classifier's performance. Without it, the ROC-curve spans about 96.5% of the total square area of the plot, whereas with the CA 125 indicator it spans almost 98%. Also, the two training methods are seen to give comparable results. Figure 3 shows the ROC-curve calculated for the total population for the 11-3-1 MLP case, trained with simulated annealing

Table 2	Training set	Test set	Total population
10-5-1 MLP, MSE	96.7%	96.4%	96.5%
10-5-1 MLP, SA	96.6%	96.2%	96.4%
11-3-1 MLP, MSE	97.9%	97.4%	97.7%
11-3-1 MLP, SA	97.9%	97.5%	97.8%

Table 2: For the two architectures (10-5-1 and 11-3-1) of the MLP and for the gradient (MSE) and the simulated annealing (SA) optimization techniques, this table gives the resulting areas under the ROC-curves.

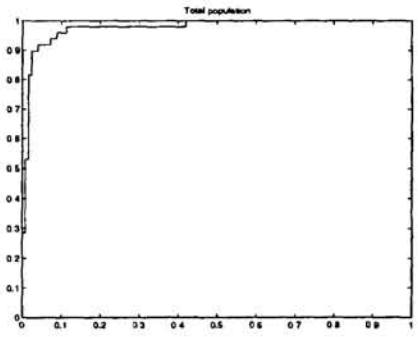

Figure 3: ROC-curves of 11-3-1 MLP (with CA 125 level indicator), trained with simulated annealing. The curve, calculated for the total population, spans 97.8% of the total region.

All patients were examined by two gynecologists, who gave their subjective impressions and also classified the ovarian tumors into (probably) benign and malignant. Histopathological examinations of the tumors afterwards showed these gynecologists

to have a sensitivity up to 98% and a false positive rate of 13% and 12% respectively. As can be seen in Figure 3, the 11-3-1 MLP has a similar performance. For a sensitivity of 98%, its false positive rate is between 10% and 15%.

5 Conclusion

In this paper we have discussed the development of a Multi-Layer Perceptron neural network classifier for use in preoperative differentiation between benign and malignant ovarian tumors. To assess the performance and for training the classifiers, the concepts of sensitivity and specificity were introduced and combined in Receiver Operating Characteristic curves. Based on objective observations available to every gynecologist, the neural network is able to make reliable predictions with a discriminating performance comparable to that of experienced gynecologists.

Acknowledgments

This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the Katholieke Universiteit Leuven, in the following frameworks: the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4-02 and IUAP P4-24), a Concerted Action Project MIPS (Modelbased Information Processing Systems) of the Flemish Community and the FWO (Fund for Scientific Research - Flanders) project G.0262.97 : Learning and Optimization: an Interdisciplinary Approach. The scientific responsibility rests with its authors.

References

[1] Bast R. C., Jr., Klug T.L. St. John E., et al, "A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer," *N. Engl. J. Med.*, Vol. 309, pp. 883-888, 1983

[2] Timmerman D., Bourne T., Tailor A., Van Assche F.A., Vergote I., "Preoperative differentiation between benign and malignant adnexal masses," *submitted*

[3] Tailor A., Jurkovic D., Bourne T.H., Collins W.P., Campbell S., "Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis," *Ultrasound Obstet. Gynaecol.* in press, 1997

[4] Jacobs I., Oram D., Fairbanks J., et al., "A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer," *Br. J. Obstet. Gynaecol.*, Vol. 97, pp. 922-929, 1990

[5] Hanley J.A., McNeil B., "A method of comparing the areas under the receiver operating characteristics curves derived from the same cases," *Radiology*, Vol. 148, pp. 839-843, 1983

[6] Swets J.A., "Measuring the accuracy of diagnostic systems," *Science*, Vol. 240, pp. 1285-1293, 1988

[7] Galen R.S., Gambino S., *Beyond normality: the predictive value and efficiency of medical diagnosis*, John Wiley, New York, 1975.

[8] Gill P., Murray W., Wright M., *Practical Optimization*, Acad. Press, New York, 1981

[9] Fletcher R., *Practical methods of optimization*, 2nd ed., John Wiley, New York, 1987.

[10] Kirkpatrick S., Gelatt C.D., Vecchi M., "Optimization by simulated annealing," *Science*, Vol. 220, pp. 621-680, 1983.

[11] Rumelhart D.E., Hinton G.E., Williams R.J., "Learning representations by back-propagating errors," *Nature*, Vol. 323, pp. 533-536, 1986.

[12] Bishop C., *Artificial Neural Networks for Pattern Recognition*, OUP, Oxford, 1996