adjusted based on the rating (Davis, 1989). In the domain of evolutionary optimization of learning systems, such a procedure was successfully applied by Igel and Kreutz (1999, 2001) for the adaptation of the probabilities to apply different mutation operators. In our context, this procedure is not suitable, because the number of different operators needed to represent the various learning times would be too high. Further, the learning operators strongly differ in their computational costs. To allow for an ongoing probability adaptation, none of these operators should be allowed to become extinct. Hence, even when some expensive operators are not suitable in the current phase of optimization, they have to be applied every now and then and may dominate the average computational complexity. Simulations support this assumption.

Instead of the operator approach, we adapt the learning time more directly. To explore different learning strategies, the value of $\tau_i^{(g)}$ is drawn independently for each individual from a Poisson distribution with the expectation $m^{(g)}$.[1] The parameter $m^{(g)}$ is subject to adaptation, which permits a smooth adjustment of the learning periods. The expectation of the learning time in adaptation cycle $g+1$ is given by

$$m^{(g+1)} = \max\left[(1-\gamma)m^{(g)} + \gamma\,\tilde{\tau}^{(g)}, m_{\min}\right] . \qquad (1)$$

The variable $\gamma \in [0,1]$ determines the influence of $\tilde{\tau}^{(g)}$, the value which would have been the most suitable choice of $m^{(g)}$ in the last adaptation cycle. In the end, (1) is a weighted average over the whole history, but the influence of past generations is exponentially suppressed depending on γ. This weighted average is similar to the evolution path in the CMA evolution strategy. The lower bound $m_{\min}$ in (1) enables a minimum diversity of the learning times to allow for continuing adaptation.

Now one is only left with the estimation of $\tilde{\tau}^{(g)}$. The efficiency of learning for τ iterations is measured by the benefit $B^{(g)}(\tau)$ (Tuson and Ross, 1998), normalized to the costs of learning $c(\tau)$ as proposed by Igel and Kreutz (1999):

$$B^{(g)}(\tau) = \frac{1}{N_\tau^{(g)}} \sum_\iota \max\left[\frac{\phi(\text{parent}(\iota)) - \phi(\iota)}{c(\tau)}, 0\right] . \qquad (2)$$

The sum runs over all $N_\tau^{(g)}$ offspring ι in adaptation cycle g that have been trained for τ iterations; $\phi(.)$ assigns each individual a fitness value. As an alternative to (2), one might relate $B^{(g)}(\tau)$ only to the fitness gain achieved by learning, i.e., the difference of the offspring's fitness before and after learning replaces the numerator in (2). However, (2) has empirically proven to be more efficient, as it allows for evaluating the number of iterations in the context of mutations. For instance, it is able to take into account the time necessary to counterbalance mutational disturbances.

The computational costs $c(\tau)$ depend on the implementation of the feed-forward neural network. For simplicity, we utilize an approximation. It takes roughly twice as much time to calculate the gradient of the network error with respect to all weights than to compute the network's error itself (Rummelhart et al., 1986): First, the input is "propagated forward" through the network and thereafter it is "propagated backward" through it. Additionally, one "forward-propagation" has to be performed after the last iteration of learning to calculate the individual's fitness $\phi(\iota)$. As we use "propagations" as the unit of the costs, we set

$$c(\tau) = 2\tau + 1 \quad . \qquad (3)$$

The learning time $\tilde{\tau}^{(g)}$ should be the time for which the improvements have been maximal in the near past and therefore might also be in the near future. This seems to be fulfilled for $\tilde{\tau}^{(g)} = \arg\left[\max_\tau\left\{B^{(g)}(\tau)\right\}\right]$. However, this might not be optimal, as in the next generations not only learning times equal to $\tilde{\tau}^{(g)}$ are applied, but also learning times randomly drawn from a Poisson distribution. In the limiting case of an isolated maximum of $B^{(g)}(\tilde{\tau}^{(g)})$, learning times $\tau_i^{(g+1)} = \tilde{\tau}^{(g)}$ would yield a maximum improvement, but slightly differing learning times would mainly lead to an exploration of bad strategies. Therefore, we do not consider the maximum of $B^{(g)}(\tau)$, but the maximum of

$$b^{(g)}(\tau) = \sum_{\tau'=0}^{\infty} B^{(g)}(\tau') \cdot \frac{\tau^{\tau'}}{\tau'!}\,\mathrm{e}^{-\tau} \quad , \qquad (4)$$

the convolution of $B^{(g)}(\tau)$ with the Poisson distribution with mean τ. The value of $b^{(g)}(\tau)$ is an estimation of the expected improvement in the case of $m^{(g)} = \tau$, as the distribution of learning times is taken into account. As a side effect, the convolution yields a smoothing of $B^{(g)}(\tau)$, which makes the evaluation of the benefit more robust. Finally, the estimation of the optimal learning time is given by

$$\tilde{\tau}^{(g)} = \arg\left[\max_\tau\left[b^{(g)}(\tau)\right]\right] \quad . \qquad (5)$$

[1] The Poisson distribution of $\tau_i^{(g)} \in \mathbb{N}_0$ is given by $p_{\tau_i^{(g)}} = \frac{\left(m^{(g)}\right)^{\tau_i^{(g)}}}{\tau_i^{(g)}!}\,\mathrm{e}^{-m^{(g)}}$. As the expectation as well as the variance are equal to $m^{(g)}$, the width of the distribution increases with its expectation.