Published in final edited form as:

J Cogn Neurosci. 2008 November ; 20(11): 1966–1979. doi:10.1162/jocn.2008.20133.

Combining modalities with different latencies for optimal motor control

Fredrik Bissmarck,
ATR International, Kyoto, Japan

Hiroyuki Nakahara,
RIKEN Brain Science Institute, Saitama, Japan

Kenji Doya, and
Okinawa Institute of Science and Technology, Okinawa, Japan

Okhide Hikosaka
National Eye Institute, NIH, Maryland

Abstract

Feedback signals may be of different modality, latency and accuracy. To learn and control motor tasks the feedback available may be redundant, and it would not be necessary to rely on every accessible feedback loop. Which feedback loops should then be utilized? In this article, we propose that the latency is a critical factor to determine which signals will be influential at different learning stages. We use a computational framework to study the role of feedback modules with different latencies in optimal motor control. Instead of explicit gating between modules, the reinforcement learning algorithm learns to rely on the more useful module. We tested our paradigm for two different implementations, which confirmed our hypothesis. In the first, we examined how feedback latency affects the competitiveness of two identical modules. In the second, we examined an example of visuomotor sequence learning, where a plastic, faster somatosensory module interacts with a preacquired, slower visual module. We found that the overall performance depended on the latency of the faster module alone, while the relative latency determines the independence of the faster from the slower. In the second implementation, the somatosensory module with shorter latency overtook the slower visual module, and realized better overall performance. The visual module played different roles in early and late learning. First, it worked as a guide for the exploration of the somatosensory module. Then, when learning had converged, it contributed to robustness against system noise and external perturbations. Overall, these results demonstrate that our framework successfully learns to utilize the most useful available feedback for optimal control.

1 Introduction

For motor control and learning, the brain relies on feedback signals of different modalities such as vision and somatosensation, and appears to use them selectively depending on the task demands and the extent of learning. For example, to play the piano, the novice must first rely on visual and somatosensory feedback for finger movement. With practice, she can gradually reduce the reliance of visual feedback, and as expert she does not need to look at the keyboard at all. How does the reliance of feedback change with learning? In this paper, we consider a hypothesis that feedback delay is a major factor in selection of feedback modality, and test if

appropriate feedback pathways can be selected through reinforcement learning to achieve the best real-time motor performance.

Recently, the optimal feedback control paradigm (Todorov & Jordan, 2002; Kording & Wolpert, 2006) has successfully predicted human motor behaviour (Todorov & Jordan, 2002; Kording & Wolpert, 2004; Liu & Todorov, 2007). Under this framework, execution is preceded by state estimation, by integration of available feedback modalities and models. Here the current state would have to be inferred from the delayed feedback in a recursive manner. In the context of well-learned, specialized motor skills, which are characterized by fast execution and minimum effort, this may be computationally expensive and time-consuming. In the present study, we propose an alternative, model-free architecture for learning and control of motor skills, where motor commands are computed in parallel by a modular circuit for each modality. Motor commands are mapped directly to the crude, delayed sensory feedback, and then integrated with the outputs of other modalities. This way, the quickest feedback is directly available to the controller for exploitation.

The actor-critic (Barto, 1995) is a reinforcement learning architecture proposed to be implemented by the basal gangliathalamocortical (BG-TC) system (Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Doya, 1999). Our framework is an actor-critic architecture with multiple actors (Nakahara, Doya, & Hikosaka, 2001), where each actor corresponds to a modality or submodality. We propose that the feedback latency constrains the utility of each actor. Further, we propose that the reinforcement learning algorithm plays a critical role in gating between inputs - only the better feedback signals, presumably those with shorter latency, would be reinforced. Thus, modular inputs are, once learned, gated implicitly by their latency. In our framework, the gating is realized by combination of population coded outputs, sharpened by a softmax function in favour of the module with highest confidence. This mechanism is different from explicit gating (Jacobs, Jordan, & Barto, 1991; Haruno, Wolpert, & Kawato, 2001), where explicit signals are computed to weight the influence of modules on the combined output.

The paper is outlined as follows: first, we present the general framework of our model (section 2). Then, we outline two implementations for validation of our model - "Experiment I" and "Experiment II" (section 3), with results of simulations presented in the following section (section 4). In Experiment I, we studied a very simple system to clearly understand the effect of feedback delays. Two modules, which are identical except for their feedback delays, were trained until convergence for a simple arm reaching task. We found that 1) performance was constrained by the latency of the faster module, and 2) that the contribution of a module depended on its latency relative to the other module. The faster module could effectively learn the task, without interference from the slower module. In Experiment II, we studied the interaction between vision and somatosensation in a sequential reaching task. Here, we assumed that a "somatosensory module", corresponding to a motor skill, is learned under the assistance of a "visual module", a pre-acquired, general but suboptimal controller guided by visual feedback. We found that for the somatosensory module to become independent of the visual module, it is critical that the latency of the somatosensory module has the shorter latency. During learning, albeit its longer latency, the visual module still functioned as a teacher for the somatosensory module. Once the motor skill was sufficiently acquired by the somatosensory module, it could execute the movement alone, independent from the visual module. At this mature stage of learning, the visual module still contributed by maintaining robustness against unexpected perturbations. Given these results, we propose that our framework, combining reinforcement learning modules by a softmax combination of population codes, realizes flexible learning and robust motor control by utilizing the best available feedback. We discuss its implications in section 5.

2 General framework

As a simple model of learning control using multiple delayed feedback channels, we consider a modular architecture as shown in Figure 1. The state $\mathbf{x}(t)$ of the physical environment evolves depending on the motor command $\mathbf{u}(t)$. The state is monitored through different sensory channels $\mathbf{y}^m(t)$ with different delays τ_m ($m = 1, \dots, M$). Each module outputs a population-coded motor command $a^m(t)$, and through their combination $\pi(t)$, the final motor command $u(t)$ is sent out to the physical environment. The goal of control is to maximize the cumulative reward $r(t)$, as in the standard reinforcement learning paradigm (Barto, 1995; Doya, 2000).

Below we outline the operation of the feedback control modules, combination of their outputs, and the learning algorithm. The architecture presented here is a modification of an earlier draft (Bissmarck et al., 2005).

2.1 Feedback control modules

Each module m has a characteristic feedback signal

$$\mathbf{y}^m(t) = \mathbf{f}^m(\mathbf{x}(t - \tau^m)) \quad (1)$$

where $\mathbf{f}^m()$ is an observation function and τ^m is a characteristic latency for the particular module. Each module gives as output a population code

$$\mathbf{a}^m(t) = \mathbf{g}(\mathbf{y}^m(t); \mathbf{w}^m) \quad (2)$$

where $\mathbf{g}(t)$ is a function approximator, with a set of trainable parameters $\mathbf{w}^m$. Each element a_j^m ($j = 1, 2, \dots, J$) corresponds to a preferred motor output $\bar{\mathbf{u}}_j$.

2.2 Combination of modular outputs

The motor command $\mathbf{u} \in R^D$ is represented by a combination of the population coded outputs of all modules with a softmax function:

$$\pi_j(t) = \frac{\exp(\beta \sum_{m=1}^M a_j^m(t) + n_j(t))}{\sum_{j=1}^J \exp(\beta \sum_{m=1}^M a_j^m(t) + n_j(t))} \quad (3)$$

where β is a constant that regulated the overlap of population codes. The noise term $n_j(t)$ makes the policy stochastic, i.e. it controls the exploration of the agent. The actual motor command $\mathbf{u}(t)$ is given by the weighted sum of the preferred motor commands $\bar{\mathbf{u}}^j$ corresponding to each population code:

$$\mathbf{u}(t) = \sum_{j=1}^J \pi_j(t) \bar{\mathbf{u}}^j. \quad (4)$$

The modular outputs a_j^m can be interpreted as the log-probability of selecting the output $\bar{\mathbf{u}}_j$,

$$a_j^m(t) = \log(P(\bar{\mathbf{u}}_j(t) | \mathbf{y}^m(t - \tau^m), \mathbf{w}^m)). \quad (5)$$

Summing over all modules, adding noise and exponentiating gives the full probability $P(\bar{\mathbf{u}}_j(t)) = \pi_j(t)$. This simple but straightforward interpretation gives a direct relationship between the activities of single neurons and distributional population codes (Pouget et al., 2003; Weiss & Fleet, 2002).

2.3 Actor-Critic learning

Our model implements a form of the continuous actor-critic (Doya, 2000). The goal of learning is to maximize the cumulative future rewards:

$$E \left[\int_0^\infty e^{-\frac{s}{\tau^{TD}}} r(t+s) ds \right], \quad (6)$$

where τ^{TD} determines how far into the future returns should be considered.

The role of the critic is to estimate the cumulative future reward from each state $\mathbf{x}(t)$ in the form of state value function:

$$V(\mathbf{x}(t)) = E \left[\int_0^\infty e^{-\frac{s}{\tau^{TD}}} r(t+s) ds \right] \quad (7)$$

for each state $\mathbf{x}(t)$. The critic learns to estimate the value function from available feedback:

$$V(\mathbf{y}^1(t), \mathbf{y}^2(t), \dots, \mathbf{y}^M(t); \mathbf{w}^c) \quad (8)$$

where $\mathbf{w}^c$ is a set of trainable parameters.

Learning of the critic and the feedback control modules is based on the temporal difference (TD) error

$$\delta^{TD}(t) = r(t) - \frac{1}{\tau^{TD}} V(t) + \dot{V}(t), \quad (9)$$

which signals the deviation of reward prediction. See appendix A for the update equations for the critic parameters $\mathbf{w}^c$ and the controller parameters $\mathbf{w}^m$.

3 Visuo-motor reaching tasks

We test the effects of different sensory feedback delays in two simulated experiments of arm reaching. In Experiment I, we used two somatosensory feedback control modules with different delays for a simple reaching task. The aim is to see how the minimal delay affect the control performance and how relative feedback delay affects the selection of the modules by learning. In Experiment II, we used both visual and somatosensory feedback modules for a sequential reaching task. The aim is to see whether and how transition from slow, task independent visual control to fast, task dependent somatosensory control happens under different feedback delays.

Figures 2 and 5 show the implementation of Experiments 1 and 2, respectively.

3.1 Reaching tasks

We use a 2DOF arm, where each link is 0.3 m long, 0.1 m in diameter, and 1 kg (see Figure 2). The state is defined by its shoulder and elbow joint angles θ_1 and θ_2 and angular velocities $\dot{\theta}_1$ and $\dot{\theta}_2$. The Cartesian hand position is $\xi^{hand}(\theta_1, \theta_2)$. The arm moves according to the motor command $\mathbf{u}(t) = (u_1(t), u_2(t))$. In Experiment I, we assume the system noise proportional to the motor command so that each joint torque is given by:

$$u_d^{actual}(t) = (1 + n_d(t))u_d(t) \quad (d=1, 2) \quad (10)$$

where $n_d(t)$ is white noise with unit variance and mean zero. In Experiment II, we assumed the system noise to be zero.

In Experiment I, the goal is to move the hand as quickly and accurately as possible to the target position T given the start position S. The reward signal is given by an exponential function of the distance of the hand to the target

$$r(t) = a \exp(-b \|\xi^{hand}(t) - \xi^{target}\|) + c \quad (11)$$

where $a = 6$, $b = 20$ and $c = -0.3$. Each trial lasts for 1.0 second.

In Experiment II, the task is to press three targets in consecutive order, which always appear one at the time at the same positions, marked 1, 2 and 3 in Figure 5. A target is pressed when the hand reaches a proximity of the target $\|\xi^{hand}(t) - \xi^{target}\| < \xi^{prox}$ at a low speed $\|\dot{\xi}^{hand}(t)\| < v^{prox}$ ($\xi^{prox} = 0.02$ m and $v^{prox} = 0.5$ m/s). After each successful target reaching, the agent is rewarded with an increasing amount (50, 100, and 150) and the next target appears immediately. Each trial ends after successful completion of the sequence, or after 5 seconds.

3.2 Feedback control modules

In Experiment I, we use two somatosensory feedback controllers, while in Experiment II, we use somatosensory and visual feedback controllers.

The somatosensory module—The somatosensory control module uses a population code representing joint angles θ and angular velocities $\dot{\theta}$ of the arm as the input:

$$y_k^m(t) = \frac{1}{Z} \exp \left(-\frac{1}{2} \left\{ \sum_d \left(\frac{\theta_d(t - \tau_m) - \bar{\theta}_{kd}}{\sigma_{kd}} \right)^2 + \sum_d \left(\frac{\dot{\theta}_d(t - \tau_m) - \bar{\omega}_{kd}}{\sigma'_{kd}} \right)^2 \right\} \right) \quad (12)$$

where $k = 1, 2, \dots, K$ is the index of the input units, $\bar{\theta}_{kd}$ and $\bar{\omega}_{kd}$ are their preferred joint angles and velocities, σ and σ' are their width parameters, and Z is a normalization term. In Experiment II, we introduced additional units representing the time since the target onset.

The output of module m is given by another population code

$$\mathbf{a}^m(t) = \mathbf{W}^m \mathbf{y}^m(t - \tau^m) \quad (13)$$

where $\mathbf{W}^m$ are trainable weight matrices. Initially all weights are zero. See appendix C for more details.

The visual module—The input for the visual feedback controller is the Cartesian positions of the hand and the target

$$\mathbf{y}^v(t) = \{\tilde{\xi}^{hand}(t), \xi^{target}(t - \tau^v)\}. \quad (14)$$

While the target position is subject to feedback delay τ^v , we assume that an estimate of the present hand position $\tilde{\xi}^{hand}$ is available, e.g., by simple linear prediction. The output is expressed as a population code $\mathbf{a}^v$.

We assume that the feedback control of the visual module (indexed by v) is pre-acquired and use a linear feedback controller with inverse dynamics compensation and output smoothing (see appendix D). The controller produces bell-shaped velocity profile similar to natural hand movement.

3.3 Actor-critic learning

To promote effective exploration, we use a low-pass filtered noise $\mathbf{n}$ in the action output (Equation 3)

$$\tau^n \dot{\mathbf{n}}(t) = -\mathbf{n}(t) + v\mathbf{N}(t) \quad (15)$$

where the time constant $\tau^n = 50$ ms and $N(t)$ is Gaussian noise with zero mean and unit variance. The amplitude v is fixed at 0.1 in Experiment I, and reduced as $1/(1 + 0.0001T)$ at trial T in Experiment II.

The critic takes the population coded somatosensory feedback as its input and use a linear weighting to produce the value estimate

$$V(t) = \mathbf{w}^c \mathbf{y}(t), \quad (16)$$

where $\mathbf{y}(t) = (\mathbf{y}^1(t), \mathbf{y}^2(t))$ in Experiment I and $\mathbf{y}(t) = \mathbf{y}^s(t)$ in Experiment II. (We verified that inclusion of visual input $\mathbf{y}^v$ in Experiment II did not affect the results in this example).

3.4 Performance measures

In order to compare the control performance under different settings of feedback delays, we use a number of performance measures, namely, the hand trajectory, the hand velocity profile, the cumulative reward, and the performance time.

The cumulative reward is given by

$$R = \int_0^T r(t) dt, \quad (17)$$

where T is the length of a trial ($T=1$ sec in Experiment I, $T=5$ sec in Experiment II).

To compare the relative contribution of different modules, we define the actor weight ratio, the output deviation, and the relative output proximity. The actor weight ratio (AWR) we define as the ratio of the absolute sum of actor weights of respective trained module:

$$AWR = \frac{\sum_k \sum_j |w_{jk}^1|}{\sum_k \sum_j |w_{jk}^2|}, \quad (18)$$

i.e. a value $AWR > 1$ indicates a relatively more influential actor of module 1. We also define the output deviation

$$d^m(\mathbf{x}(t)) = \mathbf{u}(\mathbf{x}(t)) - \mathbf{u}^m(\mathbf{x}(t)) \quad (m=v, s) \quad (19)$$

which shows how much module m 's output differ from the agent's at time t . Its time average over trajectories is given by

$$\langle d^m \rangle = \frac{1}{T} \int_0^T |d^m(t)| dt. \quad (20)$$

for a trial terminating at T . From the output deviation, we define the relative output proximity

$$p^v(t) = 1 - \frac{d^v(t)}{d^v(t) + d^s(t)} \quad (21)$$

$$p^s(t) = 1 - p^v(t) \quad (22)$$

which is a measure of the relative influence of visual and somatosensory modules over agent output, respectively. Both $p^v(t)$ and $p^s(t)$ are bounded between 0 and 1 and their relation is $p^v(t) + p^s(t) = 1$ by definition. Thus, the value of one module larger than that of the other module indicates that the former is a dominant module.

3.5 Simulation

In both experiments the same learning parameters were used: inverse temperature $\beta = 10$, time constants $\tau^{TD} = 200$ ms, $\tau^{ET} = 200$ ms, and learning rate $\alpha = 0.1$ s⁻¹.

All differential equations were approximated with the Euler forward method with a time step small enough not to affect the results (10 ms).

4 Simulation results

4.1 Experiment I: Simple reaching with somatosensory feedback modules with different delays

In Experiment I (Figure 2), we investigate how feedback latencies affect learning and control in the proposed framework. We use a simple implementation with two modules which are identical, except for their feedback latencies. The task is to learn a simple reaching movement

with a 2DOF arm. We train the networks with different pairs of feedback latencies and compare their performance and relative contribution of modules after learning has converged.

Reaching trajectories—Figure 3 shows examples of hand trajectories generated by the architecture with four different settings of feedback latencies. The 10 trajectories in the top row (Figure 3A) are generated with both modules after 100,000 training trials. Effective reaching movement is achieved by all four latency pairs in a robust manner. In the case of $(\tau_1, \tau_2) = (0, 50)$, the variability is higher than in the other examples. Since the reward function (see Section 3.1) does not explicitly penalize variability, this is still a performance optimally close to other well-performing agents (see below). However, in the cases of $(\tau_1, \tau_2) = (0, 0)$, and $(0, 50)$ ms the movement is much faster than $(50, 50)$, and $(50, 100)$ ms, as can be seen in the hand velocity plots in the bottom row (Figure 3D, solid lines (mean velocity of the samples in Figure 3A)). This shows that the shortest feedback delay is critical for the performance. This is not a trivial finding as the output of the module with the longer feedback delay can interfere with the feedback command generated by the module with shorter delay.

In order to see the relative contribution of the two modules, we compared the trajectories generated by either one of the modules (Figure 3B–C), with the other module's output set as $\mathbf{a}^m(t) = \mathbf{0}$. With the identical delays $(\tau_1, \tau_2) = (0, 0)$ and $(50, 50)$ ms, both module can realize comparable trajectories. On the other hand, with different delays $(\tau_1, \tau_2) = (0, 50)$ and $(50, 100)$ ms, while the module 1 with shorter delay can realize nice trajectories, the module 2 with longer delays generates very poor trajectories. This shows that the less desired outputs of the module with longer feedback delay is effectively shut down by reinforcement learning.

Effects of minimum and relative delays—To verify the critical role of the minimum feedback delay in control performance and the role of relative feedback delay for module selection, we measured the cumulative reward R and the actor weight ratio of trained agents, for 24 different pairs of feedback delays. Under the condition of $\tau_1 \leq \tau_2$, we plot those measures in the parameter space of $\tau_{min} = \tau_1$ and $\Delta\tau = \tau_2 - \tau_1$.

Figure 4A shows how the cumulative reward depends on τ_{min} and $\Delta\tau$. Each black dot corresponds to a trained agent with the specific latency pair. It is clearly seen that the longer τ_{min} results in reduced cumulative reward, while the relative delay $\Delta\tau$ has almost no effect on the performance. Figure 4B shows the actor weight ratio, which increases markedly with the increase of the relative delay $\Delta\tau$ (for $\tau_{min} = 0$, per definition $AWR \equiv 1$. Never was $AWR < 1$).

These results confirm that the performance of the modular learning control architecture is mostly determined by the module with shortest latency. This is achieved by the softmax combination of population coded outputs of modules (2.2) and tuning of modular outputs by actor-critic learning. It is noteworthy that potential problem of slower feedback module contaminating the good output of the faster module has been avoided by this scheme.

4.2 Experiment II: Sequential reaching with visual and somatosensory feedbacks

In Experiment II (Figure 5), we introduce a more realistic, complex implementation of a visuomotor sequence task. In motor skill acquisition, there is substantial evidence for a shift in cortical activity with experience, from prefrontal areas to motor areas (Petersen, Mier, Fiez, & Raichle, 1998; Jueptner, Frith, Brooks, Frackowiak, & Passingham, 1997; Doyon, Owen, Petrides, Sziklas, & Evans, 1996; Hikosaka, Nakamura, Sakai, & Nakahara, 2002a; Floyer-Lea & Matthews, 2005). Analogously, there should be a shift in modalities of feedback subserving these cortical areas; from extrinsic (visual) feedback needed for anticipation and proceduralization of task dynamics to intrinsic (somatosensory) feedback needed for optimization of motor control (Nakahara et al., 2001).

Here, we study the transfer between these two systems, a “visual module” and a “somatosensory module”, in a task of reaching a stereotyped sequence of three targets. The visual module relies on a general purpose controller which regulates a single reach to a given visual target. We assume that the module is preacquired and is not be optimized for any particular target sequence. The somatosensory module relies on somatosensory feedback and become optimized for repeated motor sequences. Architectures with different latency pairs τ^v and τ^s were trained for 100,000 trials, after which learning had converged in all cases. We investigate the relative contribution of the somatosensory module for different latency pairs and also compare the robustness against external perturbations of the composite system versus single module control.

Learning performance—Figure 6A–B compare the reaching trajectories before and after learning ($(\tau^v, \tau^s) = (100, 0)$ ms). Before learning, movements are variable and step-by-step - they are directed towards one target at the time. After learning, movements are stereotyped, and also coarticulated, as they are redirected towards targets 2 and 3 before preceding targets are concluded.

Figure 6C compares the performance time (the time it takes to complete one trial) before (black bars) and after (white bars) 100,000 trials of learning for 12 different latency pairs. Clearly, sequence-specific learning by the somatosensory module contributes to reduction of the performance time. Its potential to do so is primarily constrained by τ^s , which has a decreasing trend of performance time for 100, 50 and 0 ms with any latency τ^v of the visual module. In turn, τ^v is also a constraint for performance, as the performance times of learned modules are shorter with lower τ^v . The overall similar variance of performance time suggests that, within the range of investigated delays, robustness to noise is not affected by delay in our composite system.

Contribution of the somatosensory module—To elucidate the contribution of the somatosensory module, we compared the joint torque outputs of single modules (computed as in Experiment I) with the joint torque output of the agent. Figure 7A–B shows trajectories of generated joint torques over time (one trial), for the two extremes of relative latency in our study, $((\tau^v, \tau^s) = (100, 0)$ ms (A), $\Delta\tau = 100$ ms) and $((\tau^v, \tau^s) = (0, 100)$ ms, $\Delta\tau = -100$ ms) (B). In the first latency pair, the somatosensory module generates an output different from the visual module, but is evidently dominant as it is close to the agent output. In the second latency pair, the outputs of the two modules are close to each other, indicating that both modules equally contribute to the agent output. Figure 7C shows the quantitative picture, expressed as mean output deviation for 7 latency pairs. In cases of $\tau^v < \tau^s$, the visual module has the smaller output deviation ((50, 100) and (0, 100)), indicating larger contribution. In the case of mutual, long latency (100, 100), contribution is equal. Otherwise, the somatosensory module has lower output deviation. This result indicates that for the somatosensory module to learn an independent policy, it needs to have a shorter or equal latency τ^s relative to τ^v , i.e. $\tau^s \leq \tau^v$.

We then investigated how the learned behaviour is driven by the somatosensory module. We compared the normal behaviour of the learned agent with a condition with the visual module inactive. Figure 8A shows examples of hand trajectories for four latency pairs in the two conditions. With both modules, all agents are always successful. When the visual module is inactive, the ability to control the movement depends on the relative latency $\Delta\tau = \tau^v - \tau^s$. The success rate of the somatosensory module to complete a trial (given 100 trials) is shown in Figure 8B. We observe that the successful rate is high in the case of $\tau^s < \tau^v$, whereas none (or single trials in the case of (50, 50)) was successful in the case of $\tau^s \geq \tau^v$. These results further confirm our observation above (Figure 7) that the somatosensory module can become dominant as far as $\tau^s \leq \tau^v$.

Figure 8C compares the mean performance time of the two conditions. Two of the agents ((50, 0) and (100, 50)) can on average perform almost as well in the somatosensory only condition, but note the smaller variance of the normal condition. The visual module provides robustness also late in learning.

Robustness to perturbation—To further evaluate the robustness of the composite system, we perturbed a behaving agent ($(\tau^v, \tau^s) = (100, 0)$ ms) by applying a force on the end effector (hand). An impulse with constant force (400 N) was applied for 50 ms in a random direction (in the plane of the arm), for 50 ms, 0.3–0.6 s after trial start. Figure 9A–B shows an example trajectory, where the impulse (400 N up left, onset at 0.3 s) throws the agent off track to miss target 2 to the left. The green/blue colours of the trajectory indicate the relative proximity (see section 3.4) in Figure 9A of visual/somatosensory modules' output to the agent's, respectively. Note how the visual module predominates after the perturbation, to put back the trained movement on track (towards target 2), after which the somatosensory module regains influence anew. Figure 9C shows a comparison of the impact on performance time (mean of 1000 trials) for the random impulse, when operating with both or single modules active. The visual module functions as a safeguard against perturbations, since the somatosensory module alone (blue bar) cannot effectively recover, resulting in the significantly higher performance time (for which 56 % of the trials were timed out at 5.0 s). The somatosensory module contributes to speedup before and after perturbation recovery, which is why both modules (black bar) are performing faster than the visual only (green bar).

In summary, these results indicate that in this visuomotor sequence task, as learning progresses, the somatosensory module with the presumably shorter latency becomes dominant in motor control. After learning, the visual module provides stability when the effector ends up outside the well-trained regime. The memory transfer, or the degree of control by different modalities, critically depends on the difference in latencies between the visual and somatosensory modules.

5 Discussion

We have examined how feedback latency affects the relative importance of modules for the learning and control of real-time motor skills. With soft-max combination of population-coded output of multiple control modules, we demonstrated in simulations how the modules with shorter latency attain dominance in motor control. Although the result may sound straightforward, there are potential problems with conflicts of multiple modules, e.g., the longer latency output pulling back movement by the shorter latency module. It is noteworthy that appropriate module selection was achieved without any explicit gating and simply by reinforcement of the output of the module that best contributed to the performance. *removed: The last experiment showed that module weighting is highly flexible; the general-purpose visual module takes over the job when the trajectory-specific somatosensory module does not perform well.*

The use of population codes strongly contribute to the robustness and stability of our framework. On the level of perception, the robustness towards sensor noise of population coding (Georgopoulos et al. 1988, Pouget et al. 2000) give sensory estimates of high certainty, enhanced by the redundancy of multiple modalities. On the level of action selection, the population codes make the sharper distributions of modular inputs regarded as signal, while the impact of flatter distributions will be suppressed like noise. In experiment II, this explains the robustness of the composite system, demonstrated towards noise and force perturbations. The somatosensory module will predominate the output within its experienced, expert regime, but as soon as noise or mechanical force makes the arm go out of range, the visual module will resume control.

In dealing with delayed or noisy sensory signal, a recently popular paradigm is to use recursive Bayesian filters to estimate the hidden state (Todorov, 2004). Such a model may also explain more weight on the faster module, that is more informative about the current state. However, Bayesian inference requires the models of the physical dynamics and sensory delay and noise and also takes heavy on-line computation, except for linear Gaussian systems where Kalman filtering is possible. Instead, here we pursued much simpler approach of training feedback controllers specialized for given delays. Analysis of pros and cons of these approaches and their possible integration is the subject of our future study. *removed: It is also worth mentioning that speed up of the trajectory in experiment II is a natural outcome of delayed reward discounting.*

The effects of feedback delays are a relatively less investigated aspect of motor control and learning (Miall & Jackson, 2006). Some tracking experiments have been conducted, all showing worse performance for artificially imposed delays 100 ms and longer (Miall et al., 1985; Foulkes & Miall, 2000; Miall & Jackson, 2006; Ogawa et al., 2007). Kitazawa et al. also showed that learning speed and performance of prism adaptation of a reaching task worsened with delays of visual feedback of the end point error, for humans (Kitazawa et al., 1995) and monkeys (Kitazawa & Yin, 2002).

In experiment II, faster movements were learned even though reward was given only for key presses, regardless of time expenditure. The time discounting of reward used in our actor-critic learning algorithm can naturally explain why performance of numerous skill learning tasks (e.g. Anderson 1983) speed up, although speed is not an explicit performance criterion.

The mechanism of transfer from declarative to procedural memories is poorly understood (Doyon & Benali, 2005; Hikosaka et al., 2002b). In our framework, modules with shorter latency become dominant with learning. As demonstrated in experiment II, this allows specialized motor skills based on fast, intrinsic feedback loops to emerge under general purpose controllers based on slow, extrinsic feedback like vision or audition. If the difference in feedback latency is long enough, the faster modality will eventually become independent of the slower modality, which can then be used for other purposes.

There are two analogies between our framework and the BG-TC system: 1) its organization into modular circuits (Alexander & Crutcher, 1990), and 2) the actor-critic architecture (Houk et al., 1995). In previous experimental (Hikosaka et al., 1999, 2002a) and computational (Nakahara et al., 2001) work, we have proposed that prefrontal and motor BG-TC loops cooperate in motor sequence learning, encoding sequences in visual and motor coordinates, respectively.

The success of this rather simple modular learning control framework motivates future studies with agents comprising of more complex, heterogeneous features, such as different sensor noise levels, learning speeds, or inclusion of feedforward components. For example, given a slow, low-noise module and a fast, noisy module, the former would be used for precision tasks and the latter for speed tasks. To further test the generality of this prediction, delayed auditory feedback could be added as a third modality, and modality dependence could be tested under different pairs of feedback delays.

The brain receives possibly thousands of sensory signals, from which it has to make a sensible response. Biological reinforcement learning may not just be about selecting actions, but also about selecting sensory input. In this context, feedback latencies may be a critical factor for which input and output connections are formed.

Acknowledgments

FB wishes to thank Mitsuo Kawato, Erhan Oztup and Jun Morimoto for comments on an earlier draft. This research was funded by Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency.

References

- Alexander GE, Crutcher MD. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci* 1990;13(7):266–71. [PubMed: 1695401]
- Barto, A. Adaptive critics and the basal ganglia. In: Houk, J.; Davis, J.; Beiser, D., editors. *Models of information processing in the basal ganglia*. Cambridge, MA: MIT Press; 1995. p. 215-232.
- Bissmarck, F.; Nakahara, H.; Doya, K.; Hikosaka, O. Responding to modalities with different latencies. In: Saul, K.; Weiss, Y.; Bottou, L., editors. *Advances in neural information processing systems*. Vol. 17. Vancouver, Canada: Cambridge, MA: MIT Press; 2005. p. 169-176.
- Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw* 1999;12(7–8):961–974. [PubMed: 12662639]
- Doya K. Reinforcement learning in continuous time and space. *Neural Comput* 2000;12(1):219–45. [PubMed: 10636940]
- Doyon J, Benali H. Reorganization and plasticity in the adult brain during learning of motor skills. *Curr Opin Neurobiol* 2005;15(2):161–7. [PubMed: 15831397]
- Doyon J, Owen AM, Petrides M, Sziklas V, Evans AC. Functional anatomy of visuomotor skill learning in human subjects examined with positron emission tomography. *Eur J Neurosci* 1996;8(4):637–48. [PubMed: 9081615]
- Floyer-Lea A, Matthews PM. Distinguishable brain activation networks for short- and long-term motor skill learning. *J Neurophysiol* 2005;94(1):512–8. [PubMed: 15716371]
- Foulkes AJ, Miall RC. Adaptation to visual feedback delays in a human manual tracking task. *Exp Brain Res* 2000;131(1):101–10. [PubMed: 10759175]
- Haruno M, Wolpert DM, Kawato M. Mosaic model for sensorimotor learning and control. *Neural Comput* 2001;13(10):2201–20. [PubMed: 11570996]
- Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, et al. Parallel neural networks for learning sequential procedures. *Trends Neurosci* 1999;22(10):464–71. [PubMed: 10481194]
- Hikosaka O, Nakamura K, Sakai K, Nakahara H. Central mechanisms of motor skill learning. *Curr Opin Neurobiol* 2002a;12(2):217–22. [PubMed: 12015240]
- Hikosaka O, Nakamura K, Sakai K, Nakahara H. Central mechanisms of motor skill learning. *Curr Opin Neurobiol* 2002b;12(2):217–22. [PubMed: 12015240]
- Houk, J.; Adams, J.; Barto, A. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, J.; Davis, J.; Beiser, D., editors. *Models of information processing in the basal ganglia*. Cambridge, MA: MIT Press; 1995. p. 249-270.
- Jacobs R, Jordan M, Barto A. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science* 1991;15(15):219–250.
- Jueptner M, Frith CD, Brooks DJ, Frackowiak RS, Passingham RE. Anatomy of motor learning. ii. subcortical structures and learning by trial and error. *J Neurophysiol* 1997;77(3):1325–37. [PubMed: 9084600]
- Kitazawa S, Kohno T, Uka T. Effects of delayed visual information on the rate and amount of prism adaptation in the human. *J Neurosci* 1995;15(11):7644–52. [PubMed: 7472515]
- Kitazawa S, Yin PB. Prism adaptation with delayed visual error signals in the monkey. *Exp Brain Res* 2002;144(2):258–61. [PubMed: 12012163]
- Kording KP, Wolpert DM. Bayesian integration in sensori-motor learning. *Nature* 2004;427(6971):244–7. [PubMed: 14724638]
- Kording KP, Wolpert DM. Bayesian decision theory in sensorimotor control. *Trends Cogn Sci* 2006;10(7):319–26. [PubMed: 16807063]
- Liu D, Todorov E. Evidence for the flexible sensorimotor strategies predicted by optimal feedback control. *J Neurosci* 2007;27(35):9354–68. [PubMed: 17728449]

- Miall RC, Jackson JK. Adaptation to visual feedback delays in manual tracking: evidence against the smith predictor model of human visually guided action. *Exp Brain Res* 2006;172(1):77–84. [PubMed: 16424978]
- Miall RC, Weir DJ, Stein JF. Visuomotor tracking with delayed visual feedback. *Neuroscience* 1985;16(3):511–20. [PubMed: 4094689]
- Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci* 1996;16(5):1936–47. [PubMed: 8774460]
- Nakahara H, Doya K, Hikosaka O. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J Cogn Neurosci* 2001;13(5):626–47. [PubMed: 11506661]
- Ogawa K, Inui T, Sugio T. Neural correlates of state estimation in visually guided movements: an event-related fmri study. *Cortex* 2007;43(3):289–300. [PubMed: 17533753]
- Petersen E, van Mier H, Fiez A, Raichle E. The effect of practice on the functional anatomy of task performance. *Proceedings of the National Academy of Sciences* 1998;95:853–860.
- Pouget A, Dayan P, Zemel RS. Inference and computation with population codes. *Annu Rev Neurosci* 2003;26:381–410. [PubMed: 12704222]
- Todorov E. Optimality principles in sensorimotor control. *Nat Neurosci* 2004;7(9):907–15. [PubMed: 15332089]
- Todorov E, Jordan MI. Optimal feedback control as a theory of motor coordination. *Nat Neurosci* 2002;5(11):1226–35. [PubMed: 12404008]
- Weiss, Y.; Fleet, D. Velocity likelihoods in biological and machine vision. In: Rao, R.; Olshausen, B.; Lewicki, M., editors. *Statistical theories of the cortex*. Cambridge, MA: MIT Press; 2002. p. 77-96.

Appendix

A. Learning algorithm

Our model implements a form of the continuous actor-critic (Doya, 2000). The function of the critic is to estimate the cumulative sum of expected future reward $r(t)$, i.e. to learn the value function. For a given policy, the continuous value function is defined as

$$V(\mathbf{x}(t)) = E \left[\int_0^{\infty} e^{-\frac{s}{\tau^{TD}}} r(t+s) ds \right] \quad (23)$$

for each state $\mathbf{x}(t)$. The time constant τ^{TD} determines how far into the future returns should be considered. The critic implements a function approximator to estimate the value function from available feedback:

$$V = V(\mathbf{y}^1(t - \tau^1), \mathbf{y}^2(t - \tau^2), \dots, \mathbf{y}^M(t - \tau^M); \mathbf{w}^c) \quad (24)$$

where $\mathbf{w}^c$ is a set of trainable parameters. The temporal difference (TD) error δ^{TD} is the discrepancy between expected and actual return $r(t)$. In its continuous form (Doya, 2000):

$$\delta^{TD}(t) = r(t) - \frac{1}{\tau^{TD}} V(t) + \dot{V}(t). \quad (25)$$

The TD error is used to update the parameters in the critic (see below), and converges to zero when Equation 8 equals Equation 6. The TD error is also used to improve the policy $\hat{\pi}(t)$ of the actor, where circumflex denotes the noise-free, deterministic policy. The action deviation signal

$$E_j(t) = \frac{(\pi_j(t) - \hat{\pi}_j(t))^2}{2} \quad (26)$$

is the difference between the learned action and the action that was actually selected. The TD error reinforces or penalizes this deviation to update the policy $\hat{\pi}$. To control the time scale of states and actions to be updated we use eligibility traces:

$$\dot{e}_k^c(t) = -\frac{1}{\tau^{ET}} e_k^c + \frac{\partial V}{\partial w_k^c} \quad \dot{e}_{kj}^m(t) = -\frac{1}{\tau^{ET}} e_{kj}^m + \frac{\partial E_j(t)}{\partial w_{kj}^m} \quad (27)$$

for the critic and actor modules respectively. The parameters are indexed by k and τ^{ET} is a time constant. The trace for the m -th actor is given from

$$\frac{\partial E_j(t)}{\partial w_{kj}^m} = (\pi_j(t) - \hat{\pi}_j(t)) \frac{\partial \pi_j(t)}{\partial w_{kj}^m}. \quad (28)$$

The parameters are updated by gradient descent as

$$\dot{w}_k^c = \alpha \delta^{TD}(t) e_k^c(t) \quad \dot{w}_{kj}^m = \alpha \delta^{TD}(t) e_{kj}^m(t) \quad (29)$$

where α denotes the learning rate.

B. Population codes

In both experiments the population codes were equal. In the somatosensory modules, the preferred joint angles θ_{kd} and angular velocities $\dot{\omega}_{kd}$ were distributed uniformly in a $7 \times 7 \times 3 \times 3$ grid ($K_0 = 441$ nodes) for $k = 1, 2, \dots, K_0$ nodes, in the ranges $(-0.2:1.2, 1.2:1.6)$ rad and $(-1:1, -1:1)$ rad/s. The corresponding variances σ_{kd} and σ'_{kd} were half the distance to the closest node in each direction.

The preferred joint torques $\bar{\mathbf{u}}_j$ corresponding to action j were distributed symmetrically over the origin in a 5×5 grid, in the range $(-100:100, -100:100)$ Nm with the middle (0,0) unit removed. The corresponding variances σ''_{jd} were half the distance to the closest node in each direction.

The somatosensory module in experiment II also included “context units”. The context units consists of 3 tapped delay lines, each corresponding to a key in the sequence task. Each delay line had 8 units, i.e. 24 context units in all. For the k -th unit in the n -th delay line ($k > K_0, k \neq K_0 + 8(n-1) + 1$):

$$\dot{y}_k^m(t) = -\frac{1}{\tau^C} y_k^m(t) + y_{k-1}(t) \quad (30)$$

where $\tau^C = 30$ ms. Each delay line is initiated by the input at $(k = K_0 + 8(n-1) + 1)$:

$$y_k^m(t) = \delta(t - \tau_n^{keypress}) \quad (31)$$

where δ is the Dirac delta function, and $\tau_n^{keypress}$ is the instant the n -th key was pressed.

C. The visual controller in Experiment II

The feedback signal $\mathbf{y}^v$ to the visual module consists of the hand kinematics ξ^{hand} , $\dot{\xi}^{hand}$ and the target position ξ^{target} . Since the computed torque control law itself does require at least a good estimation of the current motor kinematics, the delayed feedback signals will not produce satisfactory control: the delays will cause oscillations, and become unstable at some 50–100 ms. To overcome this problem, we assumed that the agent has a good model of its own internal dynamics, and can cancel out the delay of ξ^{hand} with a prediction $\tilde{\xi}^{hand}(t) = \xi^{hand}(t)$. The target position ξ^{target} is assumed not to be predictable. Thus, with the onset of a new target, it takes τ^v ms before the visual module reacts towards that target. The control is further perturbed by a decoding error, by modification of the somatosensory module and by the stochasticity of action selection.

The joint torques are first computed by

$$\dot{\mathbf{u}}^{visual}(t) = -\frac{1}{\tau^{CT}} \mathbf{u}^{visual}(t) + \lambda \mathbf{u}^{visual'}(\tilde{\xi}^{hand}, \dot{\tilde{\xi}}^{hand}, \mathbf{e}) \quad (32)$$

where τ^{CT} and λ are constants, $\mathbf{e} = \xi^{target}(t - \tau^v) - \tilde{\xi}^{hand}(t)$ and the input to the filter is the inverse dynamics equation

$$\mathbf{u}^{visual'}(t) = \mathbf{J}^T (\mathbf{M}(\tilde{\xi}^{hand}) \ddot{\tilde{\xi}}^{hand} + \mathbf{K}_1 \dot{\tilde{\xi}}^{hand} - \mathbf{K}_2 \mathbf{e} + \mathbf{C} \dot{\tilde{\xi}}^{hand}) \quad (33)$$

in Cartesian coordinates, where $\mathbf{J}$ is the Jacobian ($\partial \theta / \partial \xi^{hand}$), $\mathbf{M}$ the moment of inertia matrix and $\mathbf{C}$ the Coriolis matrix. Using a filter by Equation 32, more bell-shaped velocity profiles of the hand, similar to biological motion are generated, in contrast to using Equation 33 directly.

The module output is an expansion of the joint torque $\mathbf{u}^{visual}$ on a population vector

$$a_j^v(t) = \frac{1}{Z} \exp\left(-\frac{1}{2} \left\{ \sum_d \left(\frac{u_d^{visual}(t) - \bar{u}_{jd}}{\sigma_{jd}''} \right)^2 \right\} \right) \quad (34)$$

where Z is the normalization term, $\bar{u}_{jd}$ is a preferable joint torque for Cartesian dimension d for vector element j , σ_{jd}'' the corresponding variance.

The parameters of equations of 32 and 33 were $\tau^{CT} = 50$ ms, $\lambda = 100$, $\mathbf{K}_1 = [10 \ 0; 0 \ 10]$, $\mathbf{K}_2 = [50 \ 0; 0 \ 50]$.

Figure 1.

The modular network architecture for learning control with multiple feedback pathways (see text).

Figure 2.

The implementation of $(\tau_1, \tau_2) = \text{Experiment I}$. The agent controls a 2DOF arm by applying joint torques to shoulder and elbow joints, with angles θ_1 and θ_2 respectively (ξ_1 and ξ_2 define the Cartesian coordinates). The task is to reach from the start hand position S to the target position T, as quickly and accurately as possible (according to the reward signal r). The feedback signals y^1 and y^2 are identical, a population code of joint angles and velocities. See text for further details.

Figure 3.

Trajectory samples generated by four different settings of latencies $(\tau_1, \tau_2) = (0, 0)$, $(0, 50)$, $(50, 50)$, and $(50, 100)$ milliseconds (ms) after 100,000 trials of training. S - start, T - target. (A) 10 trajectory samples generated by both modules under system noise. (B) System noise free trajectory generated by module 1 only. (C) System noise free trajectory generated by module 2 only. (D) Velocities of both modules (solid line, mean of 10 samples), module 1 only (dashed line) and module 2 (dotted line).

Figure 4.

Latency-dependent performance measures, displayed as surface plots constrained by 24 latency pairs (black dots) over latency space (τ_{min} , $\Delta\tau$). (A) Cumulative reward R , indicating behavioural performance of agent. (B) Actor weight ratio (AWR), indicating relative contribution of modules. A value below the x-y plane ($AWR = 1$) indicates relatively larger contribution of slower module 2, above plane larger contribution of faster module 1.

Figure 5.

The implementation of Experiment II. Here, with the arm as in Experiment I, the goal is to press targets 1, 2 and 3, presented in consequent order, starting from S. Reward is given only at the time when a key is pressed. The agent consists of two modules called “visual module” and “somatosensory module”. The visual module is a fixed controller, receiving feedback about the current target position ξ^{target} and hand position ξ^{hand} to control a reaching movement. The somatosensory module is similar to the modules in Experiment I. See text for further details.

Figure 6.

Performance before and after learning. (A) 5 sample trajectories before learning. (B) 5 sample trajectories after learning. (C) Performance times of 12 latency pairs before learning (black bars) and after learning (white bars), compared with equal levels of exploratory noise ($v = 0.01$). Note that the initial performance is the same for agents with equal τ^v , since the somatosensory module is inactive before learning.

Figure 7.

A comparison of contribution to joint torque outputs between the visual and somatosensory modules. (A–B) Example trajectories of shoulder (top) and elbow (bottom) torques over time for the latency pairs $(\tau^V, \tau^S) = (100, 0)$ (A) and $(0, 100)$ ms (B). The green, blue and black lines correspond to the outputs of the visual module, somatosensory module and agent, respectively. (C) A comparison of mean output deviation (100 trials, noise amplitude $\nu = 0.02$) of visual and somatosensory modules, for 7 latency pairs.

Figure 8.

(A) Learned behaviour of the agent in normal execution (“both modules”, top row, solid lines) and in execution with the visual module inactive (“somatosensory module only”, middle row, dashed lines) for four different latency pairs (noise-free). The bottom row shows corresponding, absolute hand velocities (first 1.5 seconds) over time. (B) Success rates of the “somatosensory module only” condition for 7 different latency pairs ($v = 0.02$). For the agents of $\tau^s = 100$, there were no successful trials in this condition. (C) Comparison of performance times between normal (white bars) and somatosensory module only (gray bars) conditions, for successful trials.

Figure 9.

External force perturbation imposed on the end effector (hand), for a trained agent (τ^v, τ^s) = (100, 0) ms. (A–B) Example trajectory when an impulse (400 N, 50 ms) perturbs the composite system. (A) Spatial movement trajectory. The two arrows indicate the direction of the force perturbation, drawn from the position of start and stop of the impulse. The green/blue colour indicate relative proximity to visual/somatosensory modules' output, respectively, i.e. green indicates $p^v(t) > p^s(t)$ and blue $p^v(t) < p^s(t)$. (B) Temporal trajectories (corresponding to (A)) of impulse (top), and output deviations of visual (green) and somatosensory (blue) modules. Note that output deviation is inverse proportional to proximity in (A). (C) Mean PT for perturbed trials with an impulse (400 N, 50 ms) of random direction and random onset (0.3–0.6 s from trial start), for the agent (τ^v, τ^s) = (100, 0) ms, comparing control with both versus single module.