OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Interference and Shaping in Sensorimotor Adaptations with Rewards

Ran Darshan[1,2,3], Arthur Leblois[2], David Hansel[2,3,4]*

1 Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel, **2** Laboratoire de Neurophysique et Physiologie UMR8119-CNRS Université René Descartes, Paris, France, **3** Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem, Israel, **4** The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel

Abstract

When a perturbation is applied in a sensorimotor transformation task, subjects can adapt and maintain performance by either relying on sensory feedback, or, in the absence of such feedback, on information provided by rewards. For example, in a classical rotation task where movement endpoints must be rotated to reach a fixed target, human subjects can successfully adapt their reaching movements solely on the basis of binary rewards, although this proves much more difficult than with visual feedback. Here, we investigate such a reward-driven sensorimotor adaptation process in a minimal computational model of the task. The key assumption of the model is that synaptic plasticity is gated by the reward. We study how the learning dynamics depend on the target size, the movement variability, the rotation angle and the number of targets. We show that when the movement is perturbed for multiple targets, the adaptation process for the different targets can interfere destructively or constructively depending on the similarities between the sensory stimuli (the targets) and the overlap in their neuronal representations. Destructive interferences can result in a drastic slowdown of the adaptation. As a result of interference, the time to adapt varies non-linearly with the number of targets. Our analysis shows that these interferences are weaker if the reward varies smoothly with the subject's performance instead of being binary. We demonstrate how shaping the reward or shaping the task can accelerate the adaptation dramatically by reducing the destructive interferences. We argue that experimentally investigating the dynamics of reward-driven sensorimotor adaptation for more than one sensory stimulus can shed light on the underlying learning rules.

Citation: Darshan R, Leblois A, Hansel D (2014) Interference and Shaping in Sensorimotor Adaptations with Rewards. PLoS Comput Biol 10(1): e1003377. doi:10.1371/journal.pcbi.1003377

Editor: Gelsy Torres-Oviedo, University of Pittsburgh, United States of America

Received February 10, 2013; **Accepted** October 14, 2013; **Published** January 9, 2014

Funding: This work was carried out within the framework of the France-Israel Laboratory of Neuroscience (LEA-FILNe) and supported by a grant of the France-Israel High Council for Scientific and Technological cooperation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.hansel@univ-paris5.fr

Introduction

Transformations that map sensory inputs to motor commands are referred to as sensorimotor mappings [1]. While sensorimotor mappings are already formed at early stages of development [2], they are subject to modifications, since the brain, the body and/or the environment are constantly changing. Plasticity in sensorimotor mappings has been extensively studied in situations where subjects receive sensory feedback during the task, allowing them to correct their motor actions and to adapt to the induced perturbation. These include visuomotor rotation [3], reaching movements under forcefields [4], adaptation in a smooth pursuit eye movements [5], prism adaptation [6], and pitch perturbation in songbirds [7] and in humans [8].

Although these studies involve different sensory modalities and different effectors, they are similar in the sense that they all have sensory goals (targets) and a motor gesture is made to reach the target. They consist of three phases namely a standard phase, in which subjects perform the task under regular conditions followed by an adaptation phase, where subjects perform the same task under the perturbed condition and a washout phase during which the perturbation is removed, and the subject readapts toward baseline. Remarkably, in all these three phases, movements display substantial trial to trial variability. Recent theoretical as well as experimental studies suggested that this variability plays a crucial role in sensorimotor learning and adaptation processes [9–11].

Another issue concerns the ability of subjects to generalize the adaptation from one context condition to a different context. This has been investigated by testing how subjects perform upon presentation of sensory stimuli that were not present during the adaptation phase [12,13]. Generalization is usually good for sensory stimuli that are similar to the one used during adaptation and degrades as the sensory stimuli become different [3,14]. Remarkably, subjects can even perform worse than in baseline (negative generalization) for sensory stimuli which are very different from those which was presented to the subject during adaptation. This has been observed, for instance, in motor reaching tasks, when the tested stimulus is presented in a direction which is opposite to the adapted direction [4,14].

The above mentioned studies implicitly assumed that the neural mechanisms for adaptation are driven by a sensory feedback, which supplies a continuous error signal to the subject. Yet, recent studies show that adaptation is possible even without any sensory feedback, when only a binary reward that informs on a success or a failure of a trial is provided to the subject [15–17]. Moreover, recent experimental works suggest that reward based mechanisms

Author Summary

The brain has a robust ability to adapt to external perturbations imposed on acquired sensorimotor transformations. Here, we used a mathematical model to investigate the reward-based component in sensorimotor adaptations. We show that the shape of the delivered reward signal, which in experiments is usually binary to indicate success or failure, affects the adaptation dynamics. We demonstrate how the ability to adapt to perturbations by relying solely on binary rewards depends on motor variability, size of perturbation and the threshold for delivering the reward. When adapting motor responses to multiple sensory stimuli simultaneously, on-line interferences between the motor performance in response to the different stimuli occur as a result of the overlap in the neural representation of the sensory stimuli, as well as the physical distance between them. Adaptation may be extremely slow when perturbations are induced to a few stimuli that are physically different from each other because of destructive interferences. When intermediate stimuli are introduced, the physical distance between neighbor stimuli is reduced, and constructive interferences can emerge, resulting in faster adaptation. Remarkably, adaptation to a widespread sensorimotor perturbation is accelerated by increasing the number of sensory stimuli during training, i.e. learning is faster if one learns more.

also affect the adaptation dynamics in sensorimotor tasks even when a sensory feedback is available [18,19].

However, and not surprisingly, adaptation relying solely on rewards at the end of a trial is more difficult than when a sensory feedback on the performance is provided continuously during the task, as adapting with sensory feedback conveys more information regarding errors. For instance, when visual feedback is available in visuomotor rotation tasks, subjects adapt to large perturbation (e.g. 30 degrees) in a few dozen trials [3,20], while in the absence of such feedback, but with binary (success or a failure) reward feedback, subjects find it notoriously difficult to adapt. Recent studies, nevertheless, have shown that it is possible to adapt to large perturbations relying solely on rewards if the size of the perturbation is slowly increased between rewarded blocks of trials [17,21]. The fact that progressively increasing the amount of perturbation makes it possible to adapt, even when the perturbation is large, is reminiscent of the classical shaping strategy [22]. In shaping, the difficulty of the task is increased gradually in order to accelerate learning, or to even make it possible. Although shaping is routinely used in laboratories when training animals to perform complex sensorimotor and cognitive tasks [23–25], it is only in recent years that it started to be explored in a theoretical framework [26–28].

What neural mechanisms could be involved in this reward based learning? Recent experimental evidence [29–31] indicates that rewards modulate local synaptic plasticity via global neuromodulatory signals, e.g. dopamine. When combined with the popular idea that synapses are modified according to Hebbian rules, this leads to the hypothesis that reward signals interact with local neuronal activity to modulate synaptic efficacies [32,33]. This theoretical paper aims to provide qualitative as well as quantitative insights into the conditions in which sensorimotor adaptation relying solely on rewards can take place. More specifically, we assume that a local learning rule based on the coactivation of pre and postsynaptic neurons is gated by a binary reward signal is the neural basis for modifications of synaptic efficacies [32,34,35].

We focus here on adaptation to a rotation during reaching movements where subjects are asked to move a cursor on a screen to bring it within a circular target while the cursor trajectory is rotated (perturbed) by some angle with respect to the hand trajectory. These perturbation tasks are classically used in behavioral studies of sensorimotor adaptation [3]. We consider a simplified network model of this task where adaptation relies solely on binary rewards [17]. The simplicity of the model allows us to analytically study several aspects of the adaptation dynamics. Combining these results with numerical simulations enables us to investigate the ways in which the learning dynamics depend on the model parameters. The key question is how the dynamics of adaptation are affected when the task involves multiple targets. Four main findings are reported: interferences can occur when adapting to multiple stimuli, interferences can slow down the adaptation dynamics dramatically, this depends on the (binary, stochastic) reward, and the slow down can be overcome by using shaping strategies.

Results

We consider the classical rotation experiment [3] in which a subject has to move a cursor on a screen to bring it within a circular target with a radius of $\sqrt{\epsilon}$; see Figure 1A. At the beginning of the experiment there is no discrepancy between the movement of the hand and the movement of the cursor. We assume that the subject is able to generate the appropriate hand movement to perform the task correctly. A perturbation is then introduced, so that the cursor trajectory is rotated by an angle γ with respect to the hand trajectory. The subject has to adapt his movements to this new condition.

In the present work, we focus on the case where the subject receives no visual feedback about the trajectory of the cursor. The only information on performance is a reward provided by the experimentalist at the end of a trial, according to the location of the cursor with respect to the desired target.

Our simplified model for a network which generates the reaching movement is depicted in Figure 1B. Its input layer consists of sensory neurons tuned to the location of the target. It has the geometry of a ring: the preferred direction (between 0^o and 360^o) of a neuron corresponds to its location on the ring (see Eq(2)). Hence, when a target appears, the population activity profile in the input layer peaks around a location which is also the target direction. For simplicity we assume that the tuning curves of all the neurons have the same shape. Therefore, the shapes of the population activity profile and the tuning curves are identical. In particular, the tuning width, ρ, is also the width of the activity profile.

The output layer consists of two linear units. Their activity encodes the $(r_1,r_2)=\boldsymbol{r}$ coordinates of the endpoint of the hand movement in the two dimensional environment. The connectivity matrix implementing the sensorimotor mapping between the input and the output layer is denoted by $\boldsymbol{W}\in\mathbb{R}^{2\times N}$. In addition to their feedfoward inputs from the first layer, the output units also receive a Gaussian noise, $\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma^2\boldsymbol{I})$ (see Eq(4)), where σ is the SD of the noise (also referred to hereafter as the *noise level*). The vector representing the endpoint of the cursor is obtained by rotating the output vector of the second layer, $\boldsymbol{r}$, by an angle γ (2×2 rotation matrix- $\boldsymbol{D}_\gamma$).

The reward, R, delivered at the end of the movement, depends on the distance between the cursor and the target. Unless specified otherwise it is binary: $R=1$ for a successful trial, *i.e.* if the squared distance is smaller than the target size, and $R=0$, otherwise. The

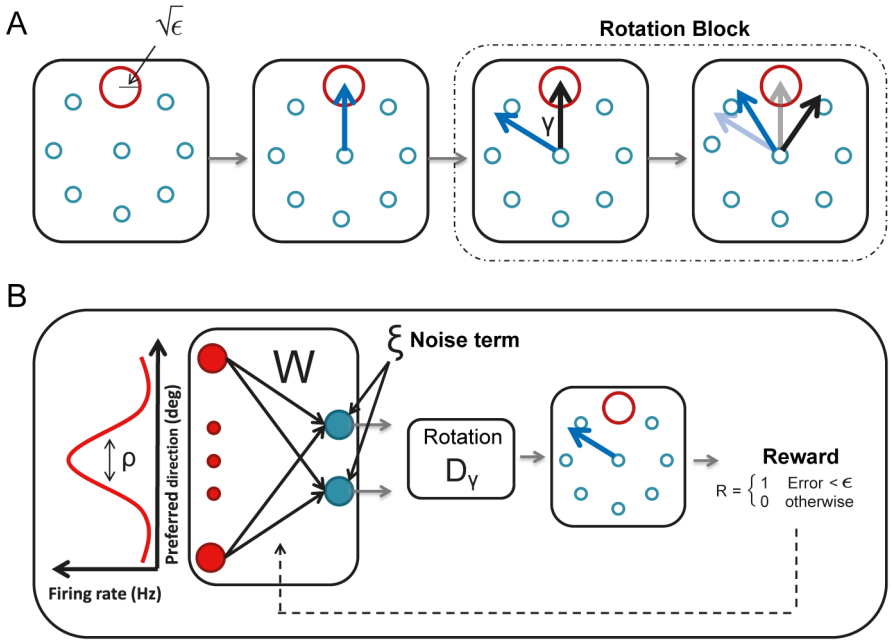

Figure 1. Schematic description of the sensorimotor adaptation task and the model. A. The rotation task. From left to right: 1) A circular target (red circle) of radius $\sqrt{\epsilon}$ appears on the screen at direction θ (here $\theta=0^{\circ}$) to instruct the subject where to move the cursor. 2) The subject moves the cursor, which is invisible to him, toward the target (blue arrow). The only information available to the subject on his performance is the reward, delivered only if the cursor falls within the target. 3) A perturbation is introduced: the cursor is rotated by an angle γ with respect to the direction of the subject's hand movement (black arrow). 4) A learning phase follows where the subject progressively adapts to the perturbation, reducing the distance between the cursor endpoint and the target. **B.** Schematic description of the model. When the target appears, the activity profile of the input layer (red neurons) peaks around the target direction. The parameter ρ controls the width of the activity profile. The connectivity matrix between the input and the output (blue neurons) layers is denoted by $\boldsymbol{W}$. A Gaussian noise with zero mean and a standard deviation of σ is added to the output layer of the network. The two-dimensional output vector rotated by the matrix $\boldsymbol{D}_{\gamma}$ represents the cursor endpoint. A reward is delivered if the distance between the cursor endpoint and the center of the target is smaller than $\sqrt{\epsilon}$. The connectivity matrix $\boldsymbol{W}$ is then changed according to a reward-modulated plasticity rule (see Eq(8)).
doi:10.1371/journal.pcbi.1003377.g001

target size is controlled by the parameter ϵ and therefore ϵ is referred to as the target size in the text.

Following trial t, the network adapts to the rotation by modifying the connectivity matrix, $\boldsymbol{W}$, according to the reward-gated synaptic plasticity rule [32,36–38]:

$$\boldsymbol{W}(t)=\boldsymbol{W}(t-1)+\eta R(t)\boldsymbol{\xi}(t)\boldsymbol{F}^{T}(\theta(t))$$

where η is the learning rate, $\boldsymbol{\xi}$ is the noise in the output layer and $\boldsymbol{F}(\theta)$ is the activity of the input layer in response to the presentation of a target in direction θ. We will assume that the initial value of the connectivity matrix is such that without noise, the network performs the task perfectly for all target directions when $\gamma=0^{\circ}$ (See Eq(9)). More details about the model are given in Materials and Methods.

The simplicity of the model allows for analytical calculations in the limit of small targets and a better understanding of the learning dynamics. However, the results reported here are grounded on the assumption of a reward-modulated learning rule and are qualitatively independent of the simplifying assumptions used to construct the model. For instance, as shown in Figure S2, the results still hold qualitatively in a more complicated network architecture with a different decoding scheme.

The learning dynamics for one target

We first consider the case where the network has to adapt to a rotation of the cursor when only one target is presented. Figure 2A (left) plots the evolution of the error (see Eq.(5)) with the number of trials, hereafter referred to as the learning curve, while the network adapts to an imposed rotation with an angle $\gamma=30^{o}$. On the right panel we plotted for the same parameters the learning curve of the directional error, which takes into account only the direction of the movement.

The error is large at the beginning of the process and decreases with the number of trials. Importantly, the dynamics strongly depend on the noise. For a low noise level (Figure 2A, $\sigma=0.1$), the error remains large for many trials and learning is slow. When the noise level is higher (Figure 2B, $\sigma=0.2$) the error declines faster. However, this comes at the cost of increasing the error after learning: the median of this error, called hereafter the *final error* (see Materials and Methods), is larger when the noise level is larger. Similarly, the probability that the network will perform the task successfully, improves more rapidly with the number of trials for $\sigma=0.2$ than for $\sigma=0.1$, but at very long time it is larger in the latter (0.824 ± 0.001) than in the former (0.443 ± 0.004) case.

The learning curves plotted in Figure 2A–B were obtained for particular realizations of the noise, $\boldsymbol{\xi}(t)$. To provide a statistical characterization of these dynamics, we estimated the distributions of the logarithm of the learning duration (τ_L) over many realizations of the noise (see Materials and Methods). As shown in Figure 2D, this distribution shifts toward longer learning duration as the noise level decreases.

Figures 2A and 2C plot the learning curves for $\epsilon=0.05$ and $\epsilon=0.1$ for the same noise level. The learning is substantially faster for $\epsilon=0.1$ but the final error is larger in this case. This is because when the target size is large, a reward might also be delivered for

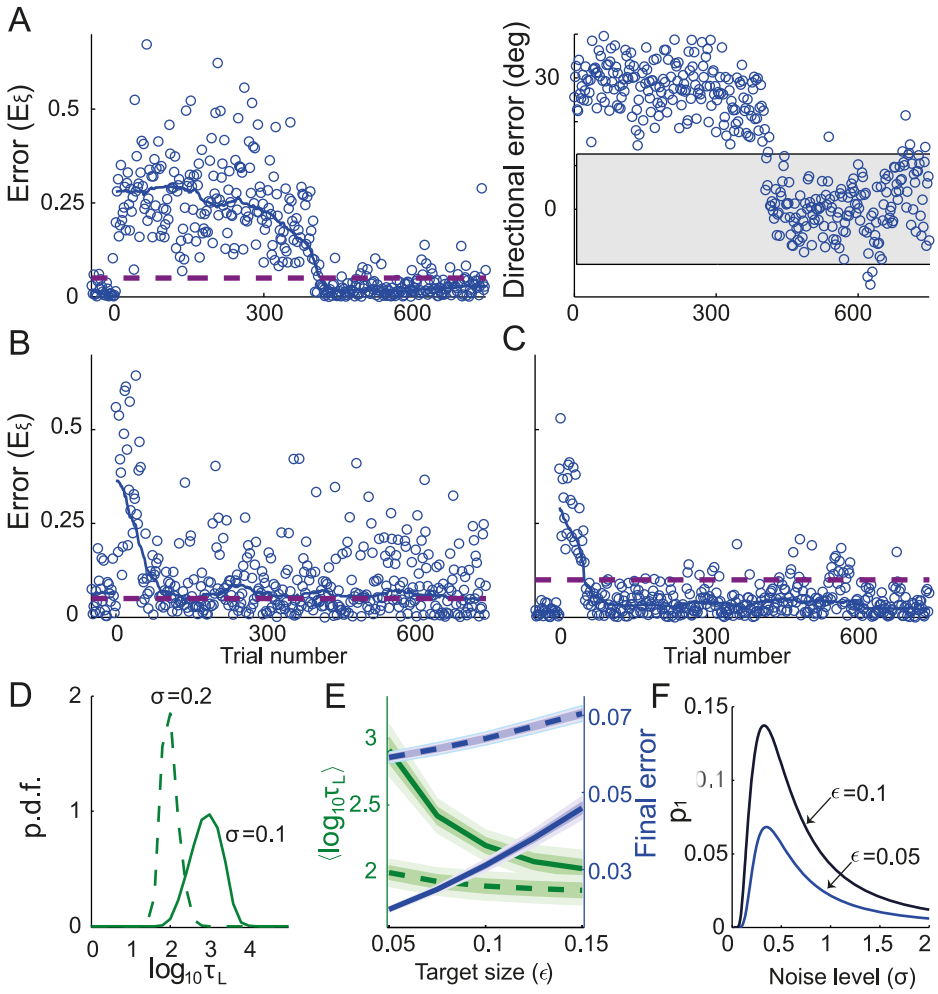

Figure 2. Learning dynamics when the network adapts to the rotation for one target. A. An examples of a learning curve for $\epsilon = 0.05$, $\sigma = 0.1$. Left: the error is calculated as the squared distance between the cursor endpoint and the target (see Eq. (5)) and plotted as a function of the trial number. The rotation perturbation is applied on trials following t = 0. For display purposes, only one in four trials is displayed. The solid line represents the error, smoothed with a 100 trials sliding median window. Final error of 0.02 ± 0.001 (mean$\pm$ SE, computed as explained in Materials and Methods). Dashed purple line: Target size. Right: as in left, but only the directional part of the error is plotted against the trial number. The shaded area corresponds to the target size. **B.** Same as in the left panel of **A.** but with $\epsilon = 0.05, \sigma = 0.2$ and corresponding final error of 0.06 ± 0.001. **C.** Same as in the left panel of **A.** but with $\epsilon = 0.1$, $\sigma = 0.1$ and a corresponding final error of 0.03 ± 0.001. **D.** Probability density function (p.d.f.) of the logarithm of the learning duration. The learning duration (τ_L) is defined as the number of trials it takes to learn the task (see: Materials and Methods). Target size is $\epsilon = 0.05$. **E.** Trade-off between learning duration and final error. Average of $\log_{10} \tau_L$ distribution (green) and the final error (blue) are plotted against the target size. The shaded area around the averages corresponds to half SD of the distributions. Solid lines: $\sigma = 0.1$. Dashed lines: $\sigma = 0.2$. **F.** The probability of getting the first reward, p_1 (see Eq. (10)), *vs.* the noise level, σ for two values of the target size. In all the panels: $\gamma = 30^\circ$.
doi:10.1371/journal.pcbi.1003377.g002

less precise movement, *i.e.*, for large errors. Figure 2E plots the log learning duration and the final error averaged over 1,000 realizations *vs.* the target size: when increasing the target size, the learning duration rapidly decreases, whereas the final error increases.

When the noise level or the target size are increased, the dynamics are typically faster because the probability of generating rewarded trials at the beginning of the learning is larger. As this probability increases, the time for the network to generate a rewarded trial decreases, leading to more updates in the connectivity matrix $\boldsymbol{W}$; hence the probability of the following trials to be rewarded increases further. This argument can be made more quantitative if one considers how the time to get the first reward depends on σ and ϵ. It has a geometrical distribution with a parameter p_1 (see Eq.(10)), which is the probability to get the first reward. Lower values of p_1 increase the expectation time to the first reward, and thereby the learning duration. When the noise level is low and the initial error is larger than the target size, the network explores a small region of the two dimensional space and the probability of getting a reward is small. In contrast, for very large noise the target is missed most of the time. The probability p_1 therefore varies non-monotonically with the noise level (Figure 2F). The dependency on target size is simpler: p_1 increases monotonically with target size, as it is more likely to reach a larger target.

Performance depends on the learning rate parameter. Obviously, the number of trials required to adapt also depends on η, which scales the increment in synaptic strength following a rewarded trial. If the rate is too small, the adaptation will be extremely long, even for large noise or big target size. On the other hand, if this rate is too large learning is likely to be impossible.

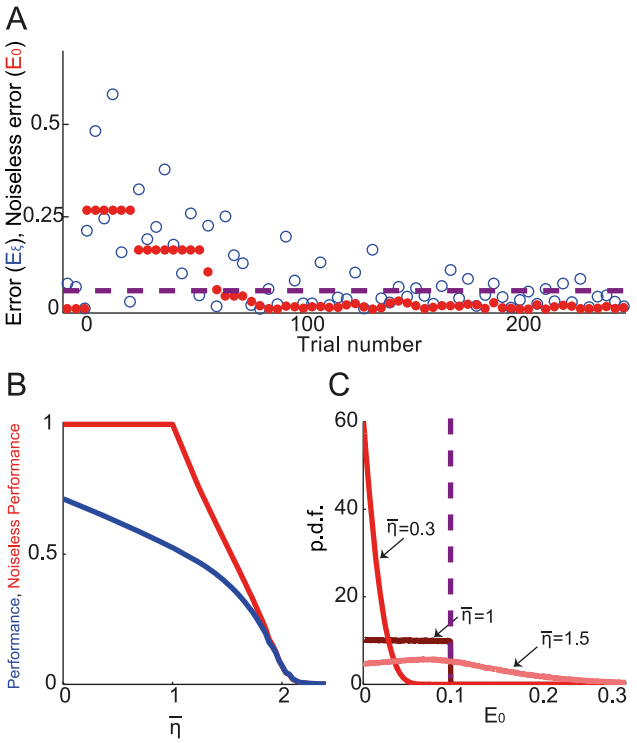

Figure 3. Performance and noiseless performance after learning depends on the learning rate. A. An example of the variations of the error (blue) and the noiseless error (red) with the number of trials for $\epsilon=0.05$ (purple dashed line), $\sigma=0.15$ and a normalized learning rate ($\bar{\eta}=\alpha\eta$, see Eq. (12)) of 0.3. For display purposes, only one in four trials is displayed. **B.** The performance (blue), *i.e.*, the probability that $E_\xi<\epsilon$ and the noiseless performance (red), *i.e.*, the probability that $E_0<\epsilon$ are plotted against the normalized learning rate. These quantities were estimated from simulations of 10^7 trials, while excluding the transient learning phase. Note that for $\bar{\eta}<1$ the noiseless performance is perfect. The standard error of the mean is too small to notice. **C.** Distribution of the noiseless error, E_0, at the end of the learning phase. For $\bar{\eta}=0.3$, the support of the distribution is bounded by ϵ. For $\bar{\eta}=1$, the distribution is uniform for $E_0<\epsilon$ and zero otherwise. For $\bar{\eta}=1.5$ the support of the distribution is bounded but extends beyond ϵ. In **B** and **C**: $\epsilon=0.1$; $\sigma=0.2$.
doi:10.1371/journal.pcbi.1003377.g003

To analyze how η affects the learning of the task it is convenient to decompose the error at trial t, $E_\xi(t)$, (Eq.(5)) into:

$$E_\xi = E_0 + 2\boldsymbol{\xi}^T\boldsymbol{E}_0 + \|\boldsymbol{\xi}\|^2$$

where $E_0=\|\boldsymbol{E}_0\|^2$ (Eq.(6)) on trial t does not depend on the noise, $\boldsymbol{\xi}(t)$ (for more details, see Materials and Methods). We therefore refer to E_0 as the *noiseless error*. Changes in the noiseless error are due to updates in the connectivity matrix, $\boldsymbol{W}$, and only occur after rewarded trials. In particular, the noiseless error rarely changes at the beginning of learning, when the probability of getting a reward is low (Figure 3A). The two other terms depend on the noise at trial t.

We also define the *noiseless performance* after learning as the probability that the noiseless error will be smaller than the target size at large time. In Figure 3A, the noiseless performance corresponds to the number of trials (red circles) that fall below target size, divided by the number of trials (see also Materials and Methods), when the number of trials is large.

Figure 3B plots the performance (blue) and the noiseless performance (red) for $\epsilon=0.1$ and $\sigma=0.2$ *vs.* the *normalized* learning rate, $\bar{\eta}=\eta\alpha$ (where α is a constant; see Materials and Methods). The noiseless performance is perfect for $\bar{\eta}<1$. It quickly deteriorates when $\bar{\eta}$ increases beyond 1, until it becomes extremely small around $\bar{\eta}=2$. Performance decreases monotonically with $\bar{\eta}$ until it reaches 0 around $\bar{\eta}=2$. Similar qualitatively results were obtained for other values of ϵ and σ (results not shown).

To better understand how the noiseless performance changes with $\bar{\eta}$, we solved the learning dynamics in the limit of small target size ($\epsilon\to 0$) analytically. In this limit, the time between rewarded trials diverges. Using the fact that when a trial t is rewarded, the noise, $\boldsymbol{\xi}(t)$, is uniquely determined in this limit, we computed the trajectory of the noiseless error analytically as a function of the number of rewarded trials; see Materials and Methods. In particular, the noiseless error goes to zero for a large number of trials if $\bar{\eta}$ is smaller than 2 and diverges for $\bar{\eta}$ larger than 2.

When $\epsilon\neq 0$ the noiseless error continues to fluctuate with time (as in Figure 3A) in the range $(0, E_{max})$, where E_{max} depends on ϵ and $\bar{\eta}$. This maximal value can be calculated analytically as shown in Materials and Methods:

$$E_{max}=\begin{cases}\epsilon & \bar{\eta}\le 1\\ \dfrac{\epsilon}{(2/\bar{\eta}-1)^2} & 1<\bar{\eta}<2\\ \infty & \bar{\eta}\ge 2\end{cases}$$

The dependency of noiseless performance with $\bar{\eta}$ (Figure 3B) stems from this result. When $\bar{\eta}<1$ the noiseless error is always smaller than the target size (see example in Figure 3C). Therefore the noiseless performance is always 1. For $\bar{\eta}=1$ the distribution of the noiseless error, can be calculated analytically (the proof is beyond the scope of this paper). It is uniform in the range $(0,\epsilon)$ (blue line in Figure 3C). For $1<\bar{\eta}<2$ the noiseless error can be larger than the target size (see example in Figure 3C) and noiseless performance is no longer perfect. In fact as $\bar{\eta}$ increases, the distribution becomes wider (its SD increases) and noiseless performance decreases monotonously. Finally, when $\bar{\eta}>2$ the above equation predicts that the support of the noiseless error distribution is unbounded, and simulations show that it becomes wider; hence the probability of getting a reward is substantially smaller than for $\bar{\eta}<2$.

While noiseless performance is always perfect for $\bar{\eta}<1$, performance can be improved by taking smaller values of $\bar{\eta}$ (Figure 3B). This is because the distribution of the noiseless error is sharper when $\bar{\eta}$ is smaller. However, decreasing $\bar{\eta}$ has the obvious consequence of increasing the learning duration. Figure 2 shows that for $\bar{\eta}=0.3$ it takes only a few dozen trials to adapt perfectly if the target size is $\epsilon=0.1$. Nevertheless, for smaller ϵ the number of trials increases dramatically. For instance, for $\epsilon=0.02$ this number becomes extremely large (much larger than 10^8) even if $\bar{\eta}=0.3$.

Accelerating the adaptation by shaping the task or the reward. Shaping is a well-known strategy in the context of operant conditioning, which allows a subject to learn difficult tasks in a reasonable amount of time [22]. In shaping strategies, the difficulty of the task is progressively increased. For a given degree of difficulty, the subject has to learn to perform the task, his performance is monitored, and when it is considered sufficiently satisfactory by the experimentalist, the difficulty of the task is increased. A shaping strategy has recently been successfully applied to allow subjects to learn the sensorimotor rotation task relying solely on a reward signal in the absence of visual feedback [17]. In this section, we apply shaping strategies in our model to examine to what extent learning can be facilitated or accelerated.

In the specific case of our sensorimotor adaptation task, the difficulty of the task depends on the target size, the rotation angle and the noise level. For fixed noise level and rotation angle, learning can be shaped by initiating the adaptation process with a large target size and then reducing the size progressively until it becomes as small as desired. This can be implemented as follows. The learning process begins with an initial value of the target size $\epsilon=\epsilon_0$, which is large enough for adaptation to be easy and fast. The target size is kept constant, while monitoring the running average of the reward. When the latter approaches a steady state, the target size is decreased by $\Delta\epsilon$ (and the running average of the reward is reinitialized to zero). We repeat this step until the target size reaches the desired value ϵ_d. An example of such a shaping strategy is depicted in Figure 4A. Here we plot the learning curve for $\epsilon_d=0.02$, when the adaptation is performed in the presence of very small noise ($\sigma=0.05$), starting with $\epsilon_0=0.2$. Within fewer than 200 trials the network has adapted and reached a performance of 0.893 ± 0.001. In fact, if the adaptation had been performed with fixed value of $\epsilon=\epsilon_d=0.02$, the probability of getting the first reward in fewer than 10^8 trials would essentially be zero ($Pr(\tau_L<10^8)=10^{-8}$), making the network unable to adapt without a tremendous number of trials.

Another example of acceleration by shaping is depicted in Figure 4B. Here, as in Figure 4A, the network has to adapt to a rotation of 30°. We used similar parameters as in [17] ($\epsilon=0.027$, corresponding to a target with a 3° radius and $\sigma=0.05$). Adaptation is performed using a constant target size, but at the beginning of the learning the angle of the rotation is small and is progressively increased with steps of $\Delta\gamma=4.2^\circ$ every block of 25 trials. The figure shows that the network adapts in fewer than 200 trials. However, in some of the realizations the network was unable to follow the gradual rotation (see inset). To avoid such cases, one can take smaller rotation steps for longer block of trials, as in [17]. Another possibility is to monitor the running average reward and to change the rotation angle when the latter approaches a steady state, similarly to what we did with the adaptive target size above.

Binary rewards, as typically used in operant conditioning, provide the subject with a limited amount of information about his performance. For instance, in our model, a binary reward does not convey any information regarding the exact distance between the cursor and the center of the target in case of a miss nor in the case of a success. One way to accelerate adaptation is to shape the reward, *i.e.*, to perform the learning using a reward that depends smoothly on the error. One possibility is to use a deterministic reward given by

$$R=\frac{1}{1+e^{(E_\xi-\epsilon)/T}} \qquad (1)$$

where T is a *smoothing parameter*. Figure 5A plots the learning curves for $T=0.01$ (top panel) and $T=0.05$ (bottom panel), for fixed values of target size and noise level ($\epsilon=0.05$, $\sigma=0.1$). The network improves substantially faster in the latter case than in the former. However, after the error has stabilized, it is comparable in both cases. Figure 5B plots the average logarithm of the learning duration as a function of T. It shows that the learning duration increases rapidly for $T\to0$, the limit where the reward becomes binary. Note that the learning duration varies non-monotonically with T (it is minimum at $T=0.05$). This is because the learning duration also increases for large T since a reward which is overly smoothed is less informative.

Remarkably, performance remains very close to 0.8 up to $T=0.05$. Therefore, using a smooth reward with $T=0.05$ reduces the learning duration substantially without affecting the performance of the network. For T above 0.05 performance drops rapidly and the learning duration becomes larger. Hence, in this case $T\simeq0.05$ is optimal. We found similar behavior for other values of noise level and target size (not shown).

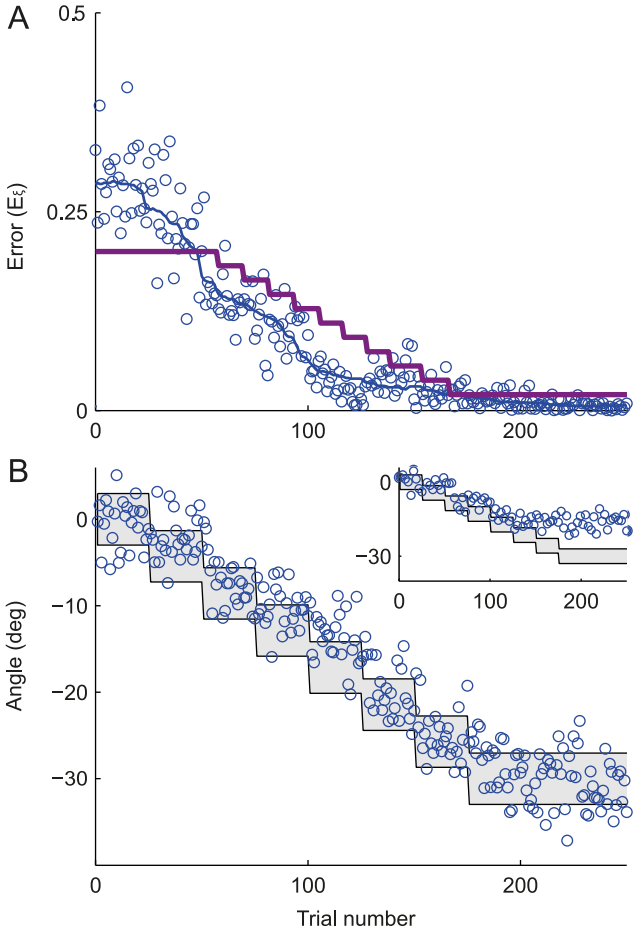

Figure 4. Shaping the task allows the network to adapt to a large rotation angle (here $\gamma=30^\circ$) even if the target size and the noise level are extremely small. A. Shaping by decreasing the target size, as explained in the text. Parameters: $\epsilon_0=0.2$; $\epsilon_d=0.02$; $\Delta\epsilon=0.018$; $\sigma=0.05$. Blue: The error is sampled every 3 trials (dots) and smoothed with a 50 trials median sliding window (line) *vs.* the number of trials. Purple: The size of the target. **B.** Reach angle (in degrees) as a function of the trial number when the rotation angle is progressively increased (see Results). The target size is fixed: $\epsilon=\epsilon_d=0.0027$. At $t=0$, $\gamma=0^\circ$. The rotation angle is increased by 4.2° every 25 trials up to $\gamma=-30^\circ$. The shaded area corresponds to the target size ($\pm3^\circ$ around the target center). Inset: the network is unable to follow the gradual rotation for a different realization of the noise with the same parameters. In both panels: $\sigma=0.05$.
doi:10.1371/journal.pcbi.1003377.g004

Another way to provide more information to the subject on his performance, still relying on a binary reward as traditionally used in operant conditioning, is to deliver it stochastically with a probability decreasing smoothly with the error. This can also accelerate adaptation as shown in Figure 5B (dashed lines). Here the reward is a Bernoulli random variable with a parameter $1/(1+e^{\frac{E_\xi-\epsilon}{T}})$.

Altogether, our modeling study predicts that reward shaping strategies, *e.g.*, providing a reward that is a smooth function of the error, as well as other shaping strategies, should be efficient in

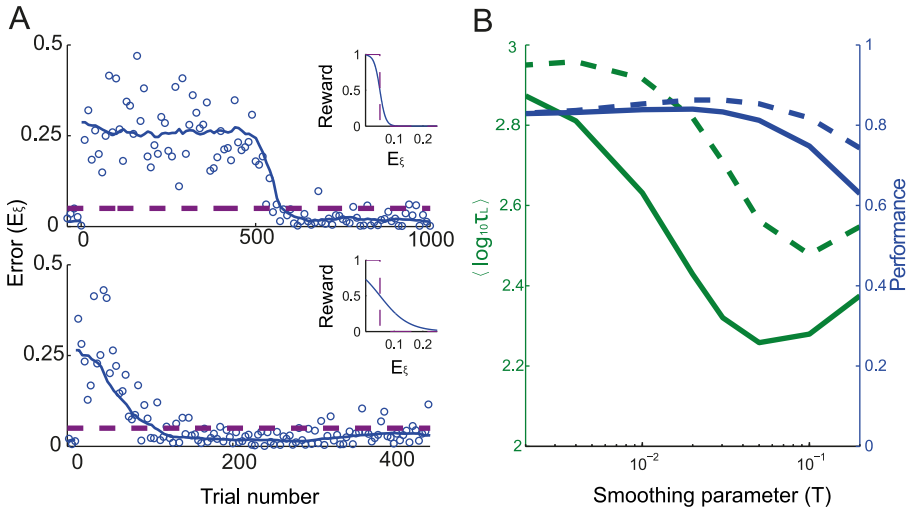

Figure 5. Shaping the reward function accelerates adaptation without impairing performance. A. The reward is given by $R=\frac{1}{1+e^{(E_\xi-\epsilon)/T}}$. Top: learning curve for a reward function that changes abruptly around target size ($T=10^{-2}$). Bottom, main panel: learning curve for a gradual reward function ($T=5\times10^{-2}$). Note the change in the abscissa scale. Inset: The reward function *vs.* the error. The target size is dashed purple line. **B.** The learning duration and the performance *vs.* the smoothing parameter, *T*. Solid lines: Deterministic smooth reward function as in **A**. Dashed lines: Stochastic binary reward delivered with a probability that depends on E_ξ (see Results). In **A** and **B**: $\epsilon=0.05$; $\sigma=0.1$.
doi:10.1371/journal.pcbi.1003377.g005

enabling or accelerating such reward-driven sensorimotor adaptation.

Generalization error. How does the network generalize the rotation for movement toward targets that were not presented during the adaptation process? To investigate this question we computed the generalization error, *G.E.* (see Materials and Methods) as a function of the angular distance, $\Delta\theta$, between the target to which the network had adapted and a test target to which it did not adapt. For small target size *G.E.* can be calculated analytically (Eq. (20)). Figure 6A plots the results for different widths of the tuning curves, ρ. For narrow tuning curves (dashed-dotted line), *G.E.* is almost one (*i.e.*, perfect generalization) only when the learned and the test targets are very close. When they are far apart, *G.E.* is almost zero. This is because the ability to generalize depends on the overlap, $\hat{\alpha}(\Delta\theta)$ (see Eq. (21)), between the activity profiles in the input layer of the network upon presentation of the learned and test targets. When the tuning curves are narrow, $\hat{\alpha}(\Delta\theta)$ is substantially different from zero only for very close targets and when they are far it is essentially zero. The range in the angular distance in which the generalization error is positive becomes broader when ρ increases (solid line). However, for wide tuning curves, *G.E.* is negative when the targets are far apart. This means that the network performance on far targets deteriorates compared to what it was before adaptation. Note that for intermediate values of ρ the generalization error can vary non-monotonically with $\Delta\theta$ (dashed lines).

The generalization error described here reveals possible interactions between the learning processes for two distinct targets, since adapting for a rotation in one target modifies performance toward others. In what follows, we evaluate the impact of such interactions when adapting the reaching movements to two targets simultaneously and dissect the mechanisms underlying *on-line* positive and negative interactions.

The learning dynamics for two targets

What is the learning dynamics when the subject has to perform the task for two targets ? How does learning the task for one of the

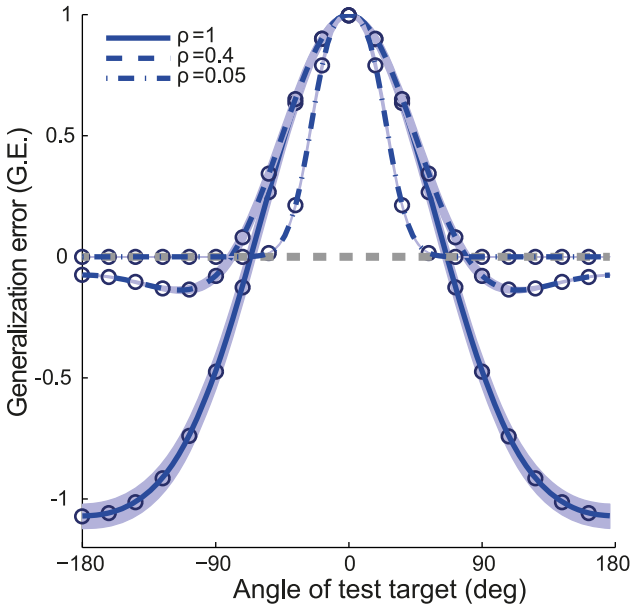

Figure 6. The generalization error (*G.E.*) for a new target (defined as the test target), presented after the network has adapted to one target. *G.E.* is plotted as a function of the angle of the test target after adaptation to a target in direction $\theta=0^\circ$. Perfect generalization is when $G.E.=1$. Lines: Analytical result for $\epsilon\to0$ (see Eq.(19)). Circles: Simulation results for $\epsilon=0.01$. For clarity, the results are displayed for test targets sampled every 15 degrees. The generalization error was averaged over 200 realizations of the noise. Shaded area represents one *SD* around the averages. Gray line: zero *G.E.*. The mapping between ρ and the half-bandwidth, $\theta_{\frac{1}{2}}$, is given in Eq. (3). For instance, $\rho=0.1$ corresponds to $\theta_{\frac{1}{2}}\simeq20^\circ$ and $\rho=1$ to $\theta_{\frac{1}{2}}\simeq65^\circ$. Parameters: $\sigma=0.14$; $\gamma=30^\circ$.
doi:10.1371/journal.pcbi.1003377.g006

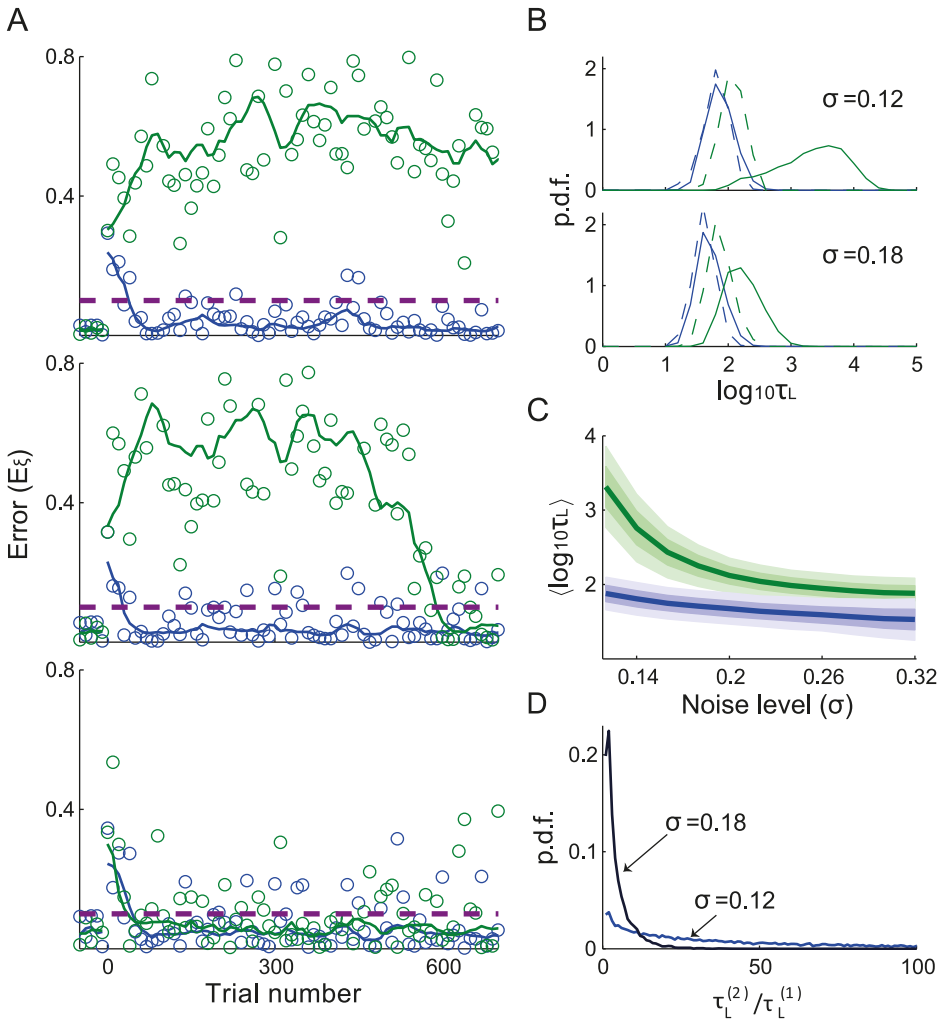

Figure 7. Delayed learning for two targets in opposite directions. A. Learning curves plotted against the number of trials for each of the targets, sampled every 10 trials. For the target that is learned first (resp. second) the curve is plotted in blue (resp. green). Top: $\sigma=0.1$. Middle: $\sigma=0.14$. Bottom panel: $\sigma=0.18$. **B.** Distribution of learning duration for two opposite targets for different noise levels. Solid lines: The probability density functions of $\log_{10}\tau_L^{(1)}$ (blue) and $\log_{10}\tau_L^{(2)}$ (green) for the two targets (solid lines) where $\tau_L^{(1)}$ (resp. $\tau_L^{(2)}$) is the learning duration for the target that is learned first (resp. second). Dashed lines: Distributions of $\log_{10}\tau_L^{(1)}$ and $\log_{10}\tau_L^{(2)}$ assuming that $\tau_L^{(1)}$ and $\tau_L^{(2)}$ are independent random variables. The distributions were estimated over 1,000 realizations of the noise. Simulations were long enough for the network to eventually adapt to both targets. Top: $\sigma=0.12$. Bottom: $\sigma=0.18$. **C.** The average and the SD of the distributions of $\log_{10}\tau_L^{(1)}$(blue) and $\log_{10}\tau_L^{(2)}$ (green) *vs.* the noise level. **D.** The distribution of the ratio $\frac{\tau_L^{(2)}}{\tau_L^{(1)}}$ for the two noise level values in **B**.

doi:10.1371/journal.pcbi.1003377.g007

targets affect learning the other one? We addressed these questions in numerical simulations, in which two targets were presented at an angular distance, $\Delta\theta$, at consecutive times. Similar results were obtained when the targets were presented in a random order with equal probability.

Delayed learning. The top panel of Figure 7A plots an example of the learning curves when the two targets are presented in opposite directions and the noise level is $\sigma=0.1$. Note that since this noise level and the target size are the same as in the bottom panel of Figure 2A, one might expect that learning the task would be fast. Remarkably, this is not the case here. The error for one of the targets decreases in fewer than 50 trials, beyond which it keeps fluctuating, most of the time below ϵ. The corresponding performance (see Eq. (26)) is 0.835 ± 0.005. This is in contrast to what happens for the other target, for which the error increases rapidly and keeps fluctuating for the whole duration of the simulation (1000 trials) around a mean that is much larger than ϵ. Therefore, in this example, the network is able to adapt in a reasonable amount of time to only one of the targets, in spite of the symmetry of the task with respect to target identity.

Increasing the noise has a dramatic effect, as shown in Figure 7A. For $\sigma=0.14$ (middle panel), the network is able to learn the task for both targets within 600 trials, but learning the second target is delayed. We term this effect throughout this paper: *delayed learning*. When increasing the noise level further ($\sigma=0.18$), the network adapts almost simultaneously to the two targets (bottom panel).

This effect of the noise in suppressing delayed learning is confirmed in Figure 7B, where the statistics of the logarithm of the learning durations over many realizations of the noise are

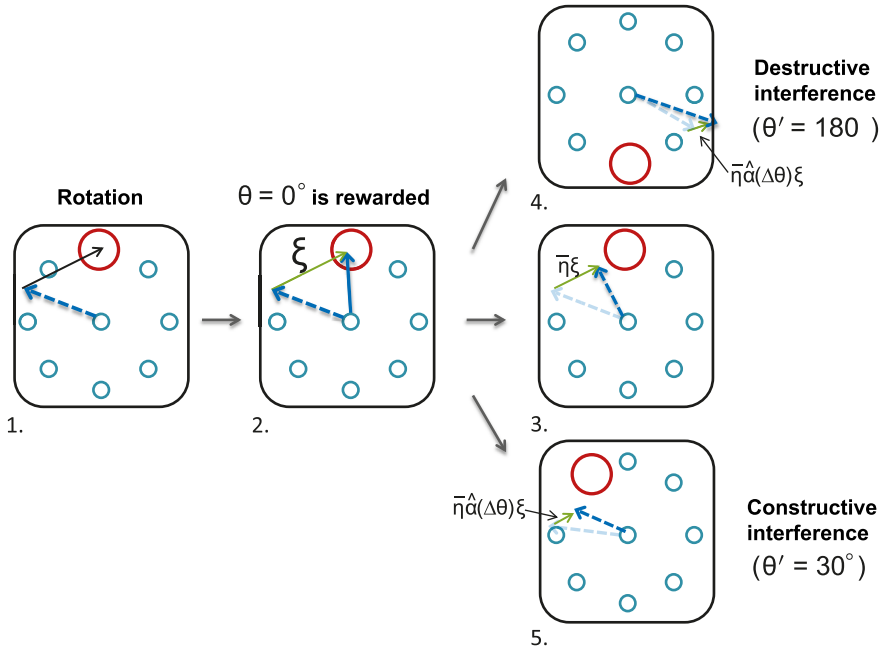

Figure 8. Geometric intuition for the destructive and constructive interferences. Following the perturbation, the cursor is rotated with respect to the output of the network, hence inducing a large noiseless error (black vector in panel 1.). The noise in the output layer (green vector in panel 2) helps the network to explore the 2D environment, until the cursor falls inside one of the targets (panel 2). This trial is rewarded and therefore the connectivity matrix is updated, affecting the output of the network for the next trials. This decreases the noiseless error, for the target for which the trial has been rewarded, as the rotated output of the network is now closer to it (by adding the vector $\bar{\eta}\xi$, panel 3). This update moves the rotated output *away* from the target in the opposite direction since the vector $\bar{\eta}\hat{\alpha}(\Delta\theta)\xi$ is away from it. This results in an increase in the error, referred to as *destructive interference*. The probability of a rewarded trial for this target is now substantially reduced, delaying learning for that target. A similar effect occurs when the two targets are sufficiently far apart. However, when they are close (panel 5) the interference becomes *constructive*, since after the update of the matrix, the rotated output gets closer to both targets. Note that the overlap, $\hat{\alpha}(\Delta\theta)$, depends on the width of the tuning curves (see Materials and Methods).
doi:10.1371/journal.pcbi.1003377.g008

depicted. The learning duration for the first (resp. the second) learned target is denoted by $\tau_L^{(1)}$ (resp. $\tau_L^{(2)}$). Obviously, the target for which adaptation occurs first depends on the specific realization of the noise. The distribution of $\log\tau_L^{(2)}$ (green) is shifted to the right with respect to the distribution of $\log\tau_L^{(1)}$ (blue), as for each realization $\tau_L^{(2)}>\tau_L^{(1)}$, by definition. As a consequence of delayed learning, this shift is larger than would be expected if the task had been learned independently for the two targets (dashed lines). For low noise level this shift is even larger (top panel). Figure 7C shows the averages of the distributions of $\log\tau_L^{(1)}$ and $\log\tau_L^{(2)}$ *vs.* σ. As it was the case for the average of $\log\tau_L$ for a single target (Figure 2C), these averages increase for low noise levels. However, the increase is faster for the second target.

The delayed learning effect is also clear in Figure 7D which plots the distribution of the ratio: $\tau_L^{(2)}/\tau_L^{(1)}$, for the same values of σ as in Figure 7B. For the highest noise level, in half of the realizations $\tau_L^{(2)}/\tau_L^{(1)}<2$. By contrast, for low noise level in more than half of the realizations the learning of the second target is at least 34 times longer than the first one. Overall, delayed learning is reduced when the noise level is increased.

Destructive and constructive interference. This delayed learning can be understood with a geometrical argument, as explained in Figure 8. When the network generates a rewarded trial for one of the targets, it affects the outcome of the second target. Hence, when the targets are in opposite directions, and if the tuning curves are sufficiently broad, this results in an increase in the error of the second target (see also Figure 7A). In other words, the learning processes for the two targets interfere *destructively*. As a result, the probability of generating a rewarded trial for the second target is reduced. Note that according to this argument if the targets are sufficiently close, the interference becomes *constructive*.

To further analyze the interference in adaptation to the two targets, we considered the correlations between the errors at consecutive presentations of the targets. For that purpose, we estimated the time dependent correlation coefficient ($CC(t)$) of the errors over different realizations (see *Materials and Methods*). A destructive interference corresponds to negative correlations, whereas a constructive interference corresponds to positive correlations. Figure 9A shows how the sign and the time course of the CC change with the angular distance, $\Delta\theta$. For the first few trials, usually none of the presentations of the targets are rewarded and, therefore, the matrix $\boldsymbol{W}$ does not change. Hence, during the first trials, $CC\simeq 0$. For a sufficiently large number of trials the network adapts to the two targets and $|CC(t)|$ reaches some stationary value.

The results in Figure 9A show that the temporal profiles of $CC(t)$ are qualitatively similar for $\Delta\theta=180^o$ and $\Delta\theta=80^o$, but in the latter case $CC(t)$ is less negative, indicating a reduction in the destructive interference. By decreasing the angle further to $\Delta\theta=60^0$ the shape of CC becomes biphasic. In the latter case the nature of the interference changes during adaptation from constructive to destructive. Finally, for sufficiently small $\Delta\theta$, the interference is always constructive. For the parameters in Figure 9A, this is already the case when $\Delta\theta=30^\circ$.

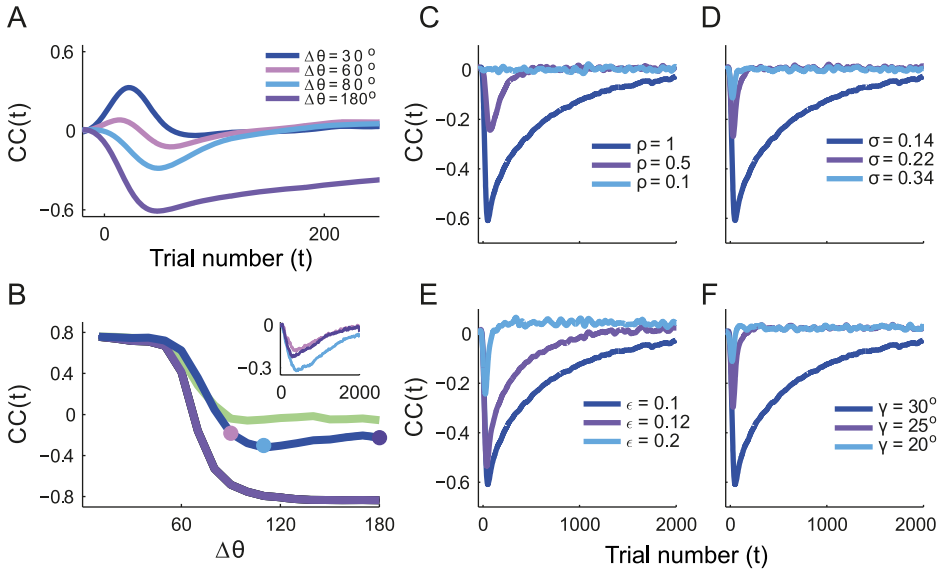

Figure 9. Destructive and constructive interferences are a function of the model parameters. The correlation coefficient, $CC(t)$, characterizes the strength and the nature of the interference during learning of the rotation task for two targets. **A.** $CC(t)$ for different values of the angular distance between the targets. The interference becomes constructive when $\Delta\theta$ decreases. **B.** The extremum of $CC(t)$ over t, CC^*, plotted against $\Delta\theta$ for different values of ρ. Purple: $\rho=1$. Blue: $\rho=0.4$. Green: $\rho=0.2$. The width of the curve was chosen to correspond to the SD of CC^*, estimated by bootstrap. Note the slight non-monotonicity for $\rho=0.4$. Inset: $CC(t)$ for $\Delta\theta=90^\circ$, $\Delta\theta=120^\circ$, $\Delta\theta=180^\circ$ for $\rho=0.4$ (same color code as for the dots on the main figure in this panel). Parameters: $\epsilon=0.1$, $\gamma=40^\circ$, $\sigma=0.14$. **C–F** $CC(t)$ is plotted for different values of σ (**C**), ϵ (**D**), ρ (**E**) and γ (**F**). In all these figures, $CC(t)$ was calculated over 1,000 repetitions. The result was low-pass filtered to suppress fast trial-to-trial fluctuations for the sake of clarity. Consequently, there is a causality artifact around $t=0$ and $CC\neq0$, although it should be. The standard errors estimated by bootstrap are small and are not plotted.
doi:10.1371/journal.pcbi.1003377.g009

Figure 9B plots the extremum of $CC(t)$, CC^*, against $\Delta\theta$, for different widths of the tuning curves. For broad and sharp tuning curves, CC^* varies monotonously with $\Delta\theta$ (Figure 9B, purple and green lines). For intermediate degrees of tuning (blue line), CC^* can display non-monotonous variations with $\Delta\theta$ (see also the inset in the figure). In fact, it reveals that the interference can vary non-monotonously with the angular distance, depending on the width of the tuning curves. This non-monotonicity can be grasped from the geometric intuition in Figure 8. The interference is more destructive when $\Delta\theta$ is large; however, as $\Delta\theta$ increases, $\hat{\alpha}(\Delta\theta)$ becomes smaller, making the interference less effective. A more rigorous proof is given in Material and Methods.

Similarly, the interference for fixed $\Delta\theta$ depends on ρ as the overlap, $\hat{\alpha}(\Delta\theta)$, becomes smaller when ρ decreases. This is depicted in Figure 9C, where we plot $CC(t)$ in the case of two targets in opposite directions, for three values of ρ. Decreasing the width of the tuning curves results in smaller values of $|CC^*|$. For very sharp tuning curves, interferences are minimal and $CC(t)$ remains very small during the whole learning process. In fact, in the limit $\rho\rightarrow0$, the adaptation process to each of the targets is independent.

Finally, Figure 9D displays $CC(t)$ for three values of noise level. The same qualitative behavior is observed in all these cases; however, CC^* is less negative and $CC(t)$ recovers faster when the noise is stronger. This is because increasing the noise decorrelates the adaptation process for the different targets, thus reducing the destructive interference. This is in line with the results displayed in Figure 7.

Destructive interferences are reduced by shaping the task or the reward. Increasing the target size (Figure 9E), as well as reducing the rotation angle (Figure 9F) reduces $|CC^*|$, and hence the destructive interference, when adapting for two targets in opposite directions. Therefore, we expect that shaping strategies which gradually manipulate these parameters can help overcome the delayed learning effect. Figure 10 shows that this is indeed the case, when changing the target size adaptively during the adaptation. The running average of the reward signal for each target was monitored *separately* and ϵ was decreased by $\Delta\epsilon$ only when both running averages reached a steady state. In this case, the network adapts to both targets quickly and simultaneously. Similarly, shaping the task by increasing the rotation angle progressively reduces the destructive interference and accelerates the learning (data not shown).

Finally, there is less interference if learning is performed with a reward which depends smoothly on the error (Eq. (1)). As depicted in Figure 10B, this results in a suppression in delayed learning. Increasing the smoothing parameter reduces $|CC^*|$. For instance, for the parameters of Figure 10B, $|CC^*|\simeq0.4$ for $T=10^{-1}$, whereas $|CC^*|\simeq0.8$ for $T=1.2\ 10^{-2}$. Similar results are found if the reward is binary but stochastic, with a probability that is a function of E_ξ (not shown).

Learning faster by learning more

How does the learning duration, *i.e.*, the time to learn the task for all the presented targets, vary with the number of targets? We simulated the learning of m targets, whose directions were *evenly distributed* between 0^o and 360^o. We took a small target size ($\epsilon=0.01$), so that up to 36 non-overlapping targets could be considered (for targets presented on a circle with radius 1).

Figure 11A plots the average time to learn the entire task in terms of the total number of target presentations for a fixed noise level and different values of tuning widths. It shows a non-monotonic dependency with the number of targets. This contrasts the monotonically increasing learning duration when targets are

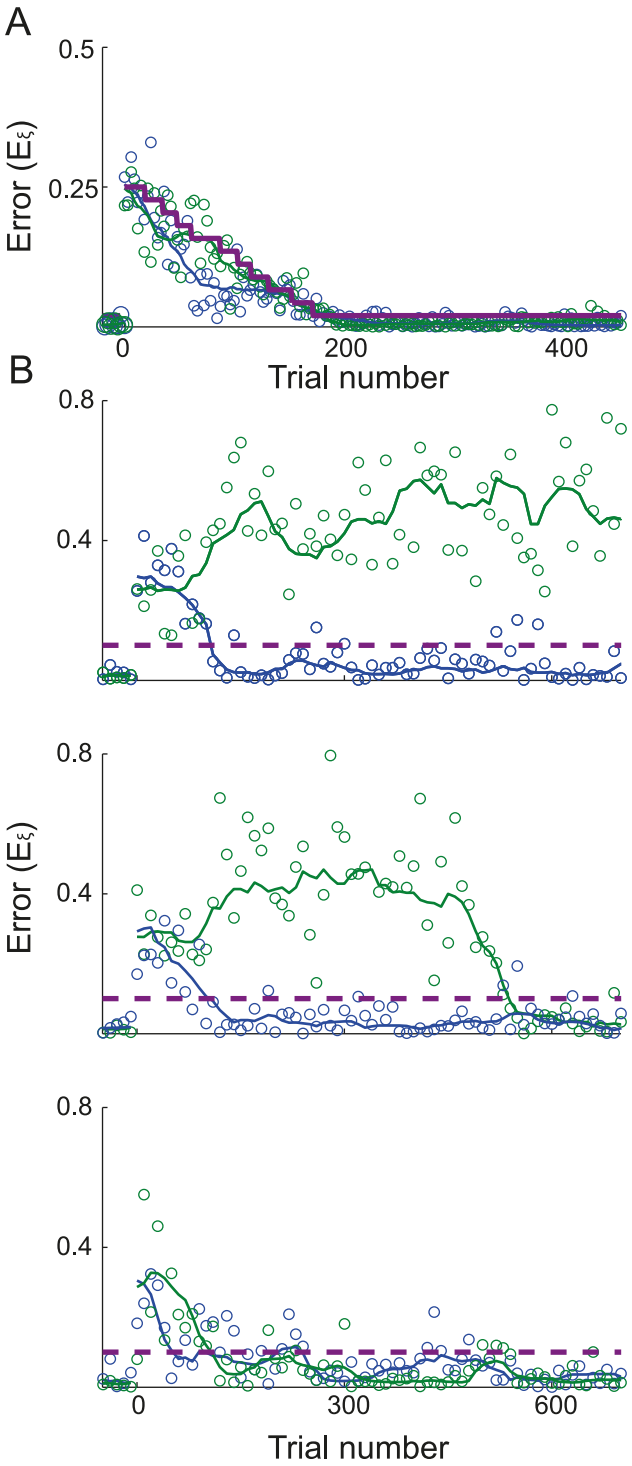

Figure 10. Shaping the task or the reward reduces the delayed learning effect. A. Learning curves for two targets in opposite directions. The task is shaped by reducing the target size. Parameters: $\epsilon_d = 0.02$; $\epsilon_0 = 0.25$; $\sigma = 0.05$. The running averages of the reward were monitored for the two targets separately. When both averages reached a steady state the target size was decreased by $\Delta\epsilon = 0.018$. The error was sampled every 3 trials. **B.** Adaptation with a smooth reward function, Eq. (1). Top: $T = 1.2 10^{-2}$. Middle: $T = 6 10^{-3}$. Bottom: $T = 10^{-1}$. Parameters: $\epsilon = 0.1$; $\sigma = 0.1$. The error was sampled every 10 trials.
doi:10.1371/journal.pcbi.1003377.g010

learned independently with the same noise level and target size (dashed line).

Narrow tuning curves. When the tuning curves are narrow (black and blue curves) and for small values of m, the overlap $\hat{\alpha}(\Delta\theta)$ is essentially zero; therefore, there is no interference and the network adapts independently to the different targets. An example is depicted in Figure 11B.1 for $\rho = 0.1$. In this figure, the *noiseless error* for all three targets is plotted against the number of *rewarded* trials. Independence is indicated by the fact that abrupt changes in the noiseless error for one of the targets do not affect the noiseless error for the other targets. The overlap only becomes significant when the targets are close enough, resulting in constructive interference (see also Figure 9A). In fact, when m increases, the adaptation for close targets interferes constructively, as depicted in Figure 11B.2 for $m = 6$. In this example, learning target 1 (see color coding in the figure) does not affect the learning of targets 3, 4 and 5 within the first 200 rewarded trials. It does, however, reduce the noiseless error for the closer targets, *i.e.*, 2 and 6. The constructive interference is also noticeable for the rest of the targets. This constructive interference between close targets facilitates adaptation and explains why the learning duration decreases for larger m, and the overall non-monotonicity of the learning duration with m.

Wide tuning curves. For wider tuning curves, interferences are already present for a small number of targets, but they can be *destructive* when the targets are far apart. For instance, for $\rho = 0.4$ and $m = 3$, improvements for one target result in an increased noiseless error, above the initial error, for the other targets (Figure 11B.3). However, as in this case ρ is not too large, adaptation is almost independent with $m = 2$ (green curve in Figure 11A). Similar to the case of narrow tuning curves, constructive interference between close targets emerges when m is increased. A representative example of adaptation with $m = 6$ and $\rho = 0.4$ is plotted in Figure 11B.4. Learning target 1 reduces the noiseless error for the two close targets, whereas the error for the other three targets, which are farther apart, becomes larger than their initial values. In this case, constructive interference among the close targets competes with destructive interference between targets that are far apart.

The drop in the learning duration when increasing m, both for wide and narrow tuning curves, implies that learning more targets might be faster than learning only a few. For instance, learning 6 targets for $\rho = 0.4$ is six times faster than learning only three of them (the 3 that are separated by 120^o).

Adaptation is in the close-to-far order when the tuning curves are broad. In Figure 11B.4 ($\rho = 0.4$) the network learned the task in a specific close-to-far order: after it had learned the first target, it learned the two closest targets (separated by $\pm 60^o$), and then the far targets (separated by $\pm 120^o$ and finally the 180^o target). Therefore, in this case the targets were learned in an *ordered* way. In contrast, in the example plotted in Figure 11B.2, the tuning curves are narrow ($\rho = 0.1$) and the learning of the targets is not ordered. This difference stems from the fact that broadening the tuning curves increases the amount of both destructive and constructive interference. As a result, by learning one target, the error of the closer targets is already reduced, whereas learning is delayed for the far targets. Increasing ρ thereby results in more ordered learning. To better characterize how the tuning width controls whether adaptation is ordered or not, we estimated the probability of this occurring as a function of ρ. Figure 11C depicts the results for $m = 6$. It shows that the fraction of the realizations for which learning is ordered increases monotonically with ρ.

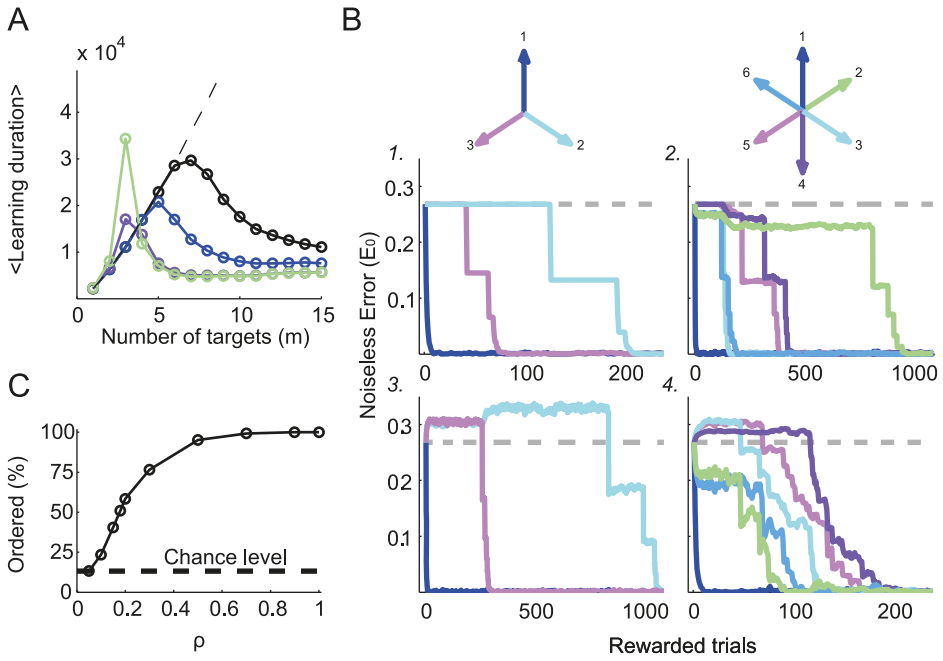

Figure 11. Adaptation to multiple targets. A. Average total number of target presentations required to learn the entire task *vs.* the number of presented targets, m. The targets are evenly distributed (between 0° to 360°). Black: $\rho=0.05$. Blue: $\rho=0.1$. Purple: $\rho=0.3$. Green: $\rho=0.4$. Dashed black line corresponds to learning the targets independently from the p.d.f. of τ_L, which was estimated from adapting to one target. **B.** Examples of the *noiseless* error during the learning, plotted *vs.* the number of *rewarded* trials. The target direction is color coded. Dashed gray lines: The initial noiseless error for $\gamma=30^\circ$. **B.1** and **B.2** are examples of the noiseless error for narrow tuning curves ($\rho=0.1$) in the case of 3 and 6 targets respectively. The plateau in the noiseless errors indicates that there is no interference between the targets. **B.3** and **B.4** are examples of the noiseless error for wider tuning curves ($\rho=0.4$) in the case of 3 and 6 targets respectively. The increase in the noiseless error above the initial error for some of the targets is the result of the destructive interference between far targets. **C.** The fraction of ordered realizations when $m=6$ as function of ρ. Chance level is 13.2%. An ordered realization is defined as learning the targets in a close-to-far order, as in the example in **B.4**. The statistics were calculated over 500 realizations. For all the results presented in this figure: $\sigma=0.14$.
doi:10.1371/journal.pcbi.1003377.g011

Generalization error for multiple targets. Figure 12A plots the generalization error after the network has adapted to 2 or 3 targets for $\rho=1$. The generalization is essentially one for all tested targets as soon as the network has adapted for three targets (green line). How does the generalization error depend on m and ρ? Figure 12B plots the noiseless performance (see Eq. (25)) averaged over all the test targets (denoted by P_t), for different values of m and ρ. For wide tuning curves, as in Figure 12A, learning the task with only 3 targets is sufficient for almost perfect performance on all the test targets (blue line, $P_t \simeq 1$). Therefore, there is no added value in adapting to more targets as far as generalization is concerned. However, as explained above, this can substantially accelerates learning. In fact, for the parameters used in Figure 12A the average learning duration is about 170 times shorter for $m=6$ than for $m=3$. When the tuning curves are narrower, the network only generalizes perfectly to all directions for large m (green and black lines in Fig 12B). Nevertheless, here it is also advantageous for the network to adapt to more targets than required for perfect generalization, since this can accelerate adaptation.

Discussion

We explored the reward-based component in adaptation processes in a setting in which a subject has to adapt reaching movements to a rotation when the only information available is the location of the target and a binary reward signal indicating success or failure on a trial [17]. The subject thus has to adapt to the perturbation by relying solely on the reward. In the framework of a simplified model of a neural network learning the task, we investigated the ways in which the adaptation dynamics depend on the noise level in the network, the target size, the size of the perturbation and the shape of the reward function. The key finding is that if the network has to adapt simultaneously to several target locations, constructive or destructive interferences between the different movements may occur. Such destructive interferences may result in a severe slowdown in the adaptation process, but this slowdown can be mitigated if the reward changes more gradually from a success to a failure around the target.

If the motor variability is not large enough with respect to the target size and the amount of perturbation (Figure 2), it takes a long time for the network to generate rewarded trials and to update its connectivity matrix. This results in slow adaptation and may be the reason why adaptation in the absence of visual feedback is notoriously difficult for subjects when the rotation angle is too large. For example, at the noise level and target size reported in [17], the probability to generate a rewarded trial in less than 10^8 trials for a rotation of 30° is essentially zero.

The time to adapt also depends on the size of the change in synaptic strength on each rewarded trial; i.e., on the learning rate parameter. We showed that perfect adaptation to one target (i.e. 100% performance in the absence of noise) is possible only when the (normalized) learning rate is smaller than 1. A high learning rate leads to decreased performance and eventually fully impedes adaptation (Figure 3). Therefore, the extent to which adaptation can be accelerated by choosing a large learning rate is limited.

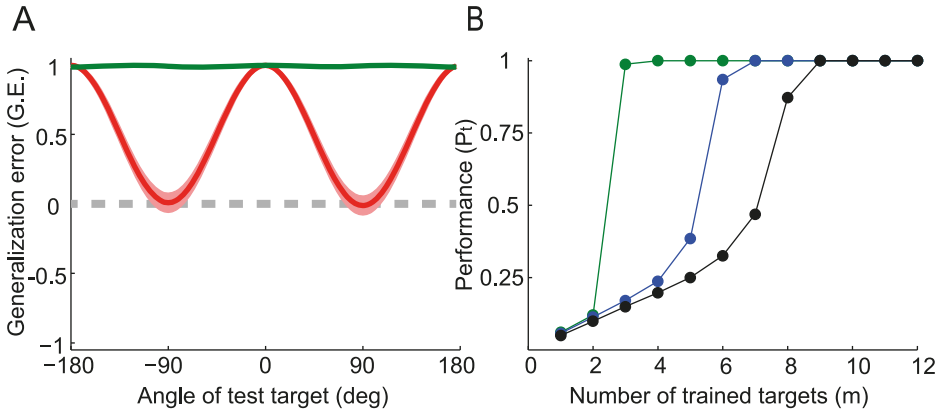

Figure 12. Generalization error ($G.E.$) and performance when adapting to multiple targets. A. The generalization error *vs.* the location of the test targets, estimated from simulations as in Figure 6. Shaded area represents one SD around the averages. Tuning width: $\rho = 1$. **B.** The noiseless performance (see Eq(25)), averaged over all the tested targets (P_t) is plotted *vs.* the number of trained targets. See Materials and Methods for details about how this quantity was estimated. Blue: $\rho = 1$. Green: $\rho = 0.1$. Black: $\rho = 0.05$. Dashed gray: zero $G.E.$ Parameters: $\epsilon = 0.01$, $\sigma = 0.14$.
doi:10.1371/journal.pcbi.1003377.g012

Adaptation is faster for large noise. On the other hand, if the noise is too large, final performance is impaired. Interestingly, motor areas display high variability at the early stages of learning, which becomes smaller afterward. This has been observed in reaching tasks in monkeys [39], as well as in song acquisition in songbirds [40]. Our study suggests that this change in noise level during learning can be functionally important to making a compromise between fast adaptation and good performance.

We showed that when adapting to multiple targets, learning the task for one target can impair performance on other targets due to destructive interference. As a result, the probability that the network will generate a rewarded trial for these targets decreases. Therefore, in this case the same noise level that allows exploration of one movement direction is insufficient when adapting to two or more targets, resulting in a delayed learning effect. Interestingly, when the network starts to adapt to the perturbation to the second target, it does not deteriorate the performance of the network on the first target that was already learned. This is because the network keeps generating rewarded trials for the first target and prevents the connectivity matrix from changing in the wrong direction for the first target.

We also showed that there are cases where the interference that occurs when multiple targets are presented is constructive. In fact, the strength and the nature of the interference depend on the similarities in the distance between the targets (the physical stimuli) and in the overlap of the tuning curves (the neural representations of the stimuli). Adding more targets creates constructive interference and therefore can accelerate adaptation.

Generality of the results

Models of sensorimotor control and learning frequently assume minimizing a squared error function. This is convenient because of analytical or computational simplicity [13,14]. However, it was shown that although these models can be a good approximation they tend to penalize large errors excessively [41]. In contrast, we chose to explore adaptation with a binary reward function, as used in experiments. Our results and predictions stem from the shape of the reward function. Specifically, they do not depend qualitatively on the specific choice of the distance error used, but are based primarily on the fact that the reward function varies sharply with the distance to the target center. The dynamics of the adaptation to more than one target depend on the overlap between the tuning curves of the input neurons. However, the precise shape of the tuning curves is not crucial and the results are unchanged if one replaces the Von Mises function we used here with any other tuning curve function, such as a cosine tuning curve (see e.g. Eq. 23).

As a matter of fact, the results we describe are the outcome of the following: 1) the same system is used to learn the task for several targets, leading to interference which depends on the way in which the targets differ physically as well as in their neuronal representation and 2) learning the task for one target can deteriorate performance on another target such that the information provided by the reward when attempting to learn the task for it becomes small, thereby delaying the learning. These two properties of the learning process are not specific to the simple model we investigated here.

In our model, the latter property stems from the fact that the reward varies sharply with the error. The learning rule we used is part of a general family of gradient-like reinforcement learning rules; i.e., learning rules that on average form a gradient ascent on the reward function [35–37]. In fact, learning with an on-line Gradient Ascent algorithm with a sigmoidal cost function can result in similar effects (Text S1; Figure S1). It might be claimed that plasticity also occurs when no reward is delivered [42]. Therefore, we also verified that the phenomenology of the model remains qualitatively the same when $R \in \{-1,1\}$ instead of using a $0-1$ reward function (unpublished data). Note that to avoid a drift of the output vector which occurs when $R \in \{-1,1\}$, the synaptic weights must be normalized in this case after each trial. Another extension of our model would be to use a reward prediction error instead of an instantaneous reward; e.g., by subtracting a running average of the reward from the instantaneous reward. Delayed learning also occurs with this type of learning rule (results not shown). In fact, previous works have argued that this modification does not affect most of the qualitative behavior of the algorithm [32,36]. However, it should be noted that in the case of multiple targets, computing the running average of the rewards over all targets is an additional source of interference, as shown recently in [35]. To avoid this, the running average of the reward needs to be monitored for each target *separately*.

We focused on the learning dynamics in a feed-forward network of linear neurons with only two layers. We chose this architecture for the sake of simplicity. However, we verified that similar

qualitative behaviors in terms of interference and delayed learning occur in a network model in which an intermediate layer consisting of nonlinear neurons was added, and in which a decoder provides the angle of reach movement instead of a vector (Text S1, Figure S2 and unpublished data). Note that in the framework of this more complex model, the noise can be unambiguously related to neuronal variability whereas in the simplified two-layer model considered in our paper, the noise is in the decoder.

One limitation of our work is that we did not model the trajectory of the movement and/or the muscle activation patterns needed to produce movements [43]. However, we expect that delayed learning and interferences also occur in a more detailed model of movement production, such as the one used in Legenstein et al. [34].

Relation to previous works and predictions

A reward-based component in a sensorimotor task was shown to be involved in adaptation to rotations even when detailed spatial information regarding the error was provided to the subject [18,19]. We investigated the ways in which neural possible mechanisms that reinforce successful actions affect adaptation dynamics. This type of reward-based mechanism was also studied in [17]. In this experiment, subjects adapted without visual feedback to a gradually increasing rotation of 1° every 40 trials, up to an 8° rotation. Our modeling results are in line with these experiments (Figure 4B). We thus predict that shaping the *reward* also accelerates adaptation.

Besides demonstrating that adaptation relying on rewards is possible by utilizing a gradual rotation paradigm, the Izawa and Shadmehr [17] results suggested that there is no change in the perceived sensory consequences of the motor commands; i.e., there should be no change in a "forward model". Therefore, in [17] adaptation was modeled by an action selection rule. Our model is similar to the latter, as we focused on the reward-based component during adaptation. However, our model differs in that it is value-free, whereas in [17] it involved value-based reinforcement learning. Nevertheless, our model can also account for the experimental results reported in [17] for one target (see Text S1, Figure S3). Moreover, it allowed us to investigate the generalization curve and possible interference during adaptation for multiple targets.

The learning algorithm. Reward modulated learning rules have been used in previous modeling studies of sensorimotor tasks, such as birdsong acquisition [10] and motor learning in primates [34]. Similar rules have also been implemented in models of decision making [32,35,44] and association tasks [45]. The reward modulated rule we used here is a special case of REINFORCE learning rules. As shown by Williams [36], REINFORCE learning rules are equivalent on average to a gradient ascent algorithm on the average reward function. In fact, the gradient ascent dynamics with the average reward function (Eq.(10), averaged over the different movement directions) can be computed analytically. However, for finite η the actual trajectories can deviate substantially from the gradient ascent trajectory. In particular, delayed learning and the reduction in learning duration with the number of targets occurs for finite η but these phenomena disappear when $\eta \to 0$ (unpublished data).

Shaping. Shaping strategies are used to teach subjects to perform operant conditioning tasks in a reasonable amount of time [22]. They were recently applied in the context of Reinforcement Learning by either increasing the complexity of the task [27,46] or by shaping the reward function [26,27,47]. In the context of our model we showed that adaptation to one target can be accelerated if the target size or the rotation angle are progressively changed. This also reduces destructive interferences, thereby accelerating adaptation to multiple targets as well. We also showed that reward shaping can efficiently suppress destructive interferences and accelerates adaptation without compromising on performance.

To the best of our knowledge there are only a few theoretical works that have addressed shaping strategies in computational models in neuroscience (see e.g. [28]). Fiete et al. [10] used an adaptive threshold for reinforcement that adapts to performance. This is equivalent to the adaptive target size used here (Figure 4A). Smooth reward functions have been used in previous models of sensorimotor learning [34,35], but the ways in which the shape of the reward function affects learning were not addressed.

Interference, delayed learning and generalization. The delayed learning effect exhibited by our network when it adapts to several targets is reminiscent of the slowing down that occurs in the model of birdsong learning in Fiete et al. [48]. In that model, a gradient ascent on a quadratic error function is performed by the network to learn a time dependent signal. The slowing down is due to destructive interferences in learning different temporal chunks of this signal. In fact, the presentation of multiple targets that involved a target in each trial, can be considered a discrete time dependent signal, and interferences when learning multiple targets can thus be seen as similar to interferences in different temporal chunks of the signal. However, in contrast to Fiete et al. [48], our network learns with a stochastic online learning rule, rather than a deterministic batch rule, and a different reward function is utilized.

Fiete et al. [48] suggested that to avoid interferences the avian brain exploits sparse neural representations. This solution is qualitatively similar to narrowing the tuning curves in our model. Similarly, Tanaka et al. [13] showed that narrow tuning curves can explain the independent learning of multiple targets in the context of a visuomotor rotation task with visual feedback. However, narrowing the tuning curves is not the only way to suppress destructive interferences, in that we showed here that they can also be suppressed by increasing the noise level, increasing the number of targets, and shaping the task or the reward.

Similarly to previous theoretical works on sensorimotor adaptation, we also showed that the shape of the generalization curve depends on the width of the tuning curves of the input neurons [4,13,14,49]. In [17] it was shown that generalization in a reward-based rotation task falls to half of its maximum value already at 10° apart from the adapted direction. However, generalization above 30° was not explored in this study. We therefore did not limit our model to a specific tuning width, as further experiments should be conducted to determine the generalization in the case of adaptation with rewards.

Negative generalization have been experimentally observed, both in adaptation to reaching movements under force-fields [4] and in visuomotor rotations with visual feedback [14]. In the latter study, the authors demonstrated that generalization curves are task-dependent, and showed how subjects negatively generalize the adaptation when targets that are far from the adapted target are presented. In fact, this study showed that generalization curves can even be non-monotonic. We predict here that this can also occur in the case of adaptation without sensory feedback.

As far as we can ascertain,, delayed learning in sensorimotor adaptation has not been reported before. For delayed learning to occur in our model, adaptation to one target needs to impair the performance on other targets and the reward must change abruptly around the target from a success to a failure. Under the

assumptions we made, the shape of generalization curves can hint at on-line interferences that can be expected during adaptation. Therefore, because negative generalization was reported in a visuomotor adaptation task when the subject receives a continuous error [14], one might expect to find on-line interferences as well when visual feedback is available. However, in this case the error function does not change abruptly with respect to the distance to the target, as subjects are aware of the cursor location. Hence, when subjects receive visual feedback, we do not expect that interferences will result in substantial delayed learning or that learning will accelerate when the number of targets is large. We verified this expectation in the case of a quadratic error [13]. In particular, the learning duration increases monotonically with the number of targets and saturates when this number is large (Text S1; Figure S4).

On the other hand, in the case of adaptation with binary rewards, we do expect that if there are angles for which generalization is negative, delayed learning will be noticeable, as the reward function changes abruptly from a success to a failure (Figure 10).

Conclusions and perspectives

The key finding of this theoretical work is that if a reward-modulated learning rule underlies adaptation, interferences are likely to be observed when learning multiple targets with a binary reward. It would be valuable to explore whether such effects occur in reward-based sensorimotor adaptation experiments with multiple sensory stimuli. We predict that for a binary reward function, destructive interferences will be observed if the neurons that encode the stimuli have broad tuning curves. These interferences are a dynamical counterpart of the generalization function and might result in a dramatic slowdown because of the abrupt change in the reward from success to failure around target size. We also predict that adding more targets should accelerate adaptation (Figure 11). From the learning curve of adaptation to one target, the rate and variability in which subjects adapt can be estimated. We predict that at parity of variability, subjects with larger learning rates will tend to display more destructive interferences and therefore slower adaptation to two targets (see Eq. (23)). By contrast,if the tuning curves are very narrow, destructive interferences are unlikely to be found. However, even in this case, when the stimuli are sufficiently close, constructive interferences should be observed. In this case as well, adding more targets should accelerate the adaptation.

Another prediction is that if adaptation is driven by reward modulated plasticity rules similar to the one we used here, smoothing the reward function should reduce interferences. In our model, this stems from the assumption of a reward modulated learning rule and not from the simplifying assumptions we made in constructing the model. Therefore, we suggest that testing this prediction could shed light on the synaptic mechanisms underlying adaptation tasks.

Finally, the location of the reward-based mechanism involved in adaptation could be the cortex-basal-ganglia network. As a matter of fact, there is evidence for the involvement of this network in pitch shift adaptation in songbirds. Although the neural correlates for adaptation in songbirds are unknown, when an auditory feedback is available to songbirds (by using miniature headphones [7]), the anterior frontal pathway, which is the avian homologue of the cortex-basal-ganglia network [50], is essential for adaptation based solely on binary rewards [15,16]. Thus, exploring the behavioral and neural differences in auditory feedback versus binary reward adaptations in pitch shift experiments in songbirds may help reveal the neural mechanisms for reward-based adaptation.

Materials and Methods

The task

We consider a motor reaching task (see Figure 1A) in which a subject manually controls the location of a cursor on a screen to bring it within a circular target of radius $\sqrt{\epsilon}$ [16]. The target location is characterized by a two dimensional vector $\hat{\boldsymbol{r}}$ of norm 1 (we assume that the target is always at distance 1 from the center of the screen) and direction θ. In a standard block of trials, the direction of motion of the cursor and the hand are the same. We assume that the subject is able to perform the task perfectly in this case. In a rotation block of trials a perturbation is introduced: the movement of the cursor on the screen is now rotated by an angle γ with respect to the hand movement. To overcome this perturbation the subject must move his hand in a direction $-\gamma$ with respect to the target. Here we focus on the case where there is no visual feedback (the cursor is not on the screen): the only information the subject receives about his performance is provided by a reward signal delivered by the experimentalist [17].

The network model

We consider a simplified computational model of a network performing this reaching task, see Figure 1B. The input layer of the network encodes the sensory information regarding the direction of the target, θ. It is composed of N directionally tuned neurons labeled by their preferred direction, $\theta_i = \frac{2\pi}{N} i$ $(i = 1,2...N)$. For simplicity, we assume that the shape of the tuning curves is the same for all neurons: upon presentation of a target in direction θ the activity of neuron i is $f(\theta_i - \theta)$. We take:

$$f(\theta_i - \theta) = C \exp\left(\frac{cos(\theta_i - \theta) - 1}{\rho}\right) \tag{2}$$

where ρ characterizes the width of the tuning curve and C is the peak response of a neuron. The width of the tuning curves at half of its maximal activity relative to the baseline (half bandwidth) in this case is:

$$\theta_{\frac{1}{2}} = \arccos\left(\rho \log\left(\frac{1 + e^{\frac{\rho}{2}}}{2}\right) - 1\right) \tag{3}$$

The second layer of the network encodes the location of the endpoint of the hand movement. It consists of two output units whose activities, r_1 and r_2, represent the two components of the hand position, $\boldsymbol{r}$. Upon presentation of a target in direction θ:

$$\boldsymbol{r}(\theta) = \frac{1}{N} \boldsymbol{W} \boldsymbol{F}(\theta) + \boldsymbol{\xi} \tag{4}$$

where $\boldsymbol{W} \in \boldsymbol{R}^{2 \times N}$ is the connectivity matrix between the two layers, $\boldsymbol{F}(\theta)$ denotes the N dimensional vector of the input layer with components $F_i(\theta) = f(\theta_i - \theta)$, and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is a Gaussian noise. The location of the cursor at the end of the movement is related to $\boldsymbol{r}$ by a 2×2 rotation matrix, $\boldsymbol{D}_\gamma$, of angle γ. Therefore, the squared distance between the endpoint location of the cursor and the center of the target is:

$$E_\xi = ||\boldsymbol{E}_\xi||^2 = ||\hat{\boldsymbol{r}} - \boldsymbol{D}_\gamma \boldsymbol{r}||^2 = ||\tilde{\boldsymbol{r}} - \boldsymbol{r}||^2 \tag{5}$$

where $\tilde{\boldsymbol{r}} = \boldsymbol{D}_\gamma^T \hat{\boldsymbol{r}}$. This quantity will be used to measure the error with which the network performs the reaching task. It is also useful

to define the *noiseless* error:

$$E_0 = ||\boldsymbol{E}_0||^2 = ||\hat{\boldsymbol{r}} - \boldsymbol{D}_\gamma \boldsymbol{y}||^2 = ||\tilde{\boldsymbol{r}} - \boldsymbol{y}||^2 \tag{6}$$

where $\boldsymbol{y} = \frac{1}{N}\boldsymbol{W}\boldsymbol{F}$. This quantity measures the error if the noise is suppressed.

Upon presentation of a target in a direction θ at trial t, the network performs the task and a reward R is delivered according to the outcome:

$$R = \begin{cases} 1 & E_\xi < \epsilon \\ 0 & otherwise \end{cases} \tag{7}$$

The matrix $\boldsymbol{W}$ is then modified according to a reward-modulated learning rule:

$$\boldsymbol{W}(t) = \boldsymbol{W}(t-1) + \eta R(t)\boldsymbol{\xi}(t)\boldsymbol{F}^T(\theta(t)) \tag{8}$$

where η is the learning rate. This learning rule can be derived in a REINFORCE framework [36].

We assume that at the beginning of learning ($t=0$), when there is no rotation, the network is able to perform the reaching task with zero noiseless error for all targets. When all the Fourier components of $f(\theta_i - \theta)$ are non-zero, this constraint fully determines $\boldsymbol{W}_j(0)$:

$$\boldsymbol{W}_j(0) = \frac{1}{|f_1|}\begin{pmatrix} \cos(\theta_j) \\ \sin(\theta_j) \end{pmatrix} \tag{9}$$

where f_1 is the first Fourier component of the tuning curves. To get Eq. 10, one needs to calculate the Fourier expansion of $\boldsymbol{W}_j(0)$ by using the constraint:

$$\boldsymbol{W}(0)\boldsymbol{F}(\theta) = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}$$

for each of the N possible target directions, θ. When some of the Fourier coefficients of the tuning curve function are zero, e.g. when using a cosine tuning curves, $\boldsymbol{W}$ is determined up to the Fourier coefficient that are irrelevant to the above constraint. This does not affect the learning dynamics.

Analysis of the model for adaptation to one target

Probability to generate a rewarded trial. The probability of generating a rewarded trial given the noiseless error at the end of the previous trial is:

$$\begin{aligned} p_1(E_0(t)) &= Pr(R=1|E_0(t)) = \frac{1}{2\pi\sigma^2}\int\int_{E_\xi(t)<\epsilon} d\xi e^{-\frac{||\xi||^2}{2\sigma^2}} \\ &= \frac{1}{\sigma^2} e^{-\frac{E_0}{2\sigma^2}} \int_0^{\sqrt{\epsilon}} e^{-\frac{r^2}{2\sigma^2}} I_0\left(\frac{r\sqrt{E_0}(t)}{\sigma^2}\right) r dr \end{aligned} \tag{10}$$

where $I_n(x)$ is the modified Bessel function of the first kind of order n [51]. The transition from the second to the third equation is done by a change of variables to polar coordinates, followed by the integration over the angle. Using this equation, we can calculate the probability to get the first reward in a given number of trials for an initial noiseless error, $E_0(t=0)$. This probability is given by a geometrical distribution with a parameter $p_1(E_0(0))$ (defined as p_1 for simplicity). When $E_0(0) > \epsilon$, the expectation value of this distribution, $1/p_1$, diverges for small values of ϵ and σ.

Learning dynamics in the limit $\epsilon \to 0$. In the limit $\epsilon \to 0$, the probability of a trial to be rewarded decreases and thus the number of trials between rewarded trials diverges (see Eq. 10). However, one can still characterize the dynamics in terms of the evolution of the error as a function of the number of *rewarded* trials. The condition that the network generates the k^{th} rewarded trial fully determines the noise:

$$\boldsymbol{\xi}(k) = \tilde{\boldsymbol{r}}(\theta) - \frac{1}{N}\boldsymbol{W}(k-1)\boldsymbol{F}(\theta) \equiv \boldsymbol{E}_0(k-1) \tag{11}$$

The connectivity matrix is then updated according to:

$$\boldsymbol{W}(k) = \boldsymbol{W}(k-1) + \eta \boldsymbol{E}_0(k-1)\boldsymbol{F}^T(\theta)$$

$$\boldsymbol{E}_0(k) = (1-\bar{\eta})\boldsymbol{E}_0(k-1)$$

where the normalized learning rate is defined by:

$$\bar{\eta} \equiv \eta\alpha \tag{12}$$

with $\alpha = \frac{1}{N}||\boldsymbol{F}(\theta)||^2$. Solving the above recursion, one finds:

$$\boldsymbol{\xi}(k) = (1-\bar{\eta})^k \boldsymbol{E}_0(0)$$

$$\boldsymbol{W}(k) = \boldsymbol{W}(0) + \frac{1-(1-\bar{\eta})^k}{\alpha}\boldsymbol{E}_0(0)\boldsymbol{F}^T(\theta)$$

The error and the squared Frobenius norm of $\boldsymbol{W}$ ($||\boldsymbol{W}||^2 = \sum_{ij} W_{ij}^2$) are then:

$$E_0(k) = 2(1-\cos\gamma)(1-\bar{\eta})^{2k}$$

$$||\boldsymbol{W}(k)||^2 = ||\boldsymbol{W}(0)||^2 + \frac{2N}{\alpha}(\cos\gamma - 1)(1-(1-\bar{\eta})^k)(1-\bar{\eta})^k$$

where we use the fact that $E_0(0) = 2(1-\cos\gamma)$.

The sequences $\boldsymbol{\xi}(k)$, $\boldsymbol{W}(k)$ and $E_0(k)$ converge when $k \to \infty$ if:

$$\bar{\eta} < 2 \tag{13}$$

Their limiting values are then:

$$\boldsymbol{\xi}(\infty) = 0$$

$$\boldsymbol{W}(\infty) = \boldsymbol{W}(0) + \frac{1}{\alpha}\boldsymbol{E}_0(0)\boldsymbol{F}^T(\theta_S)$$

$$E_0(\infty) = 0$$

$$||\boldsymbol{W}(\infty)||^2 = ||\boldsymbol{W}(0)||^2 \tag{14}$$

Therefore, after enough rewarded trials the noiseless error goes down to zero. Note that there is no need to normalize the connectivity matrix after each update in this case, since in the large k limit the norm of the matrix returns to the value it had at $k=0$.

The support of the noiseless error distribution is bounded. When ϵ is finite, the noiseless error after a rewarded trial is:

$$E_0(k)=||\boldsymbol{E}_0(k)||^2=||\boldsymbol{E}_0(k-1)-\bar{\eta}\boldsymbol{\xi}(k)||^2 \tag{15}$$

where $\boldsymbol{\xi}(k)$ is such that the constraint in Eq. (7) holds, *i.e.*, $||\boldsymbol{E}_0(k-1)-\boldsymbol{\xi}(k)||^2<\epsilon$. This constraint implies that $\boldsymbol{\xi}(k)$ can be written as:

$$\boldsymbol{\xi}(k)=\sqrt{E_0(k-1)}\hat{\boldsymbol{e}}_{\boldsymbol{E}_0}+\hat{\boldsymbol{\xi}}(k) \tag{16}$$

where $\hat{\boldsymbol{e}}_{\boldsymbol{E}_0}$ is the unit vector in the direction of $\boldsymbol{E}_0(k-1)$ and $\hat{\boldsymbol{\xi}}(k)$ is a vector with a maximal norm $\sqrt{\epsilon}$. Inserting Eq. (16) into Eq. (15) one finds:

$$\begin{aligned}E_0(k)=E_0(k-1)(1-\bar{\eta})^2+\bar{\eta}^2||\hat{\xi}(k)||^2-\\ 2\bar{\eta}(1-\bar{\eta})\sqrt{E_0(k-1)}||\hat{\xi}(k)||\mathrm{E}_0^T(k-1)\hat{\xi}(k)\end{aligned} \tag{17}$$

The noiseless error for a large number of trials is a random variable with a probability $P(E_0)$ on the support $(0,E_{max})$. For vector $\boldsymbol{y}$ to be close to the target, the maximum value of the noiseless error, E_{max}, needs to be as small as possible. To estimate E_{max}, we compute the realization of $\hat{\boldsymbol{\xi}}(k)$ which maximizes the noiseless error, Eq. (18), at each rewarded trial k.

When $\bar{\eta}<1$, $E_0(k)$ is maximal if $\boldsymbol{E}_0^T(k-1)\hat{\boldsymbol{\xi}}(k)=-1$ and $||\hat{\boldsymbol{\xi}}(k)||=\sqrt{\epsilon}$. One then gets:

$$\sqrt{E_0(k)}=\sqrt{E_0(k-1)}(1-\bar{\eta})+\bar{\eta}\sqrt{\epsilon}$$

Solving the recursion gives:

$$\sqrt{E_0(k)}=\sqrt{E_0(0)}(1-\bar{\eta})^k+\bar{\eta}\sqrt{\epsilon}\sum_{i=1}^{k}(1-\bar{\eta})^{i-1}$$

and therefore after a long time we get:

$$E_{max}(\bar{\eta}<1)=\epsilon$$

For $\bar{\eta}>1$, $E_0(k)$ in Eq(17) is maximal if $\boldsymbol{E}_0^T(k-1)\hat{\boldsymbol{\xi}}(k)=1$ and $||\hat{\boldsymbol{\xi}}(k)||=\sqrt{\epsilon}$. This leads to:

$$\sqrt{E_0(k)}=\sqrt{E_0(k-1)}(\bar{\eta}-1)+\bar{\eta}\sqrt{\epsilon}$$

Solving the recursion and taking the limit $k\to\infty$, one gets that for $1<\bar{\eta}<2$:

$$E_{max}(1<\bar{\eta}<2)=\frac{\epsilon}{(2/\bar{\eta}-1)^2}$$

and when $\bar{\eta}>2$:

$$E_{max}(\bar{\eta}>2)=\infty$$

To summarize:

$$E_{max}=\begin{cases}\epsilon & \bar{\eta}\le 1\\ \frac{\epsilon}{(2/\bar{\eta}-1)^2} & 1<\bar{\eta}<2\\ \infty & \bar{\eta}\ge 2\end{cases}$$

In particular, if $\bar{\eta}<1$ the noiseless error is guaranteed to always be smaller than ϵ at large time.

Generalization error after adaptation to one target. Let us assume that the network has adapted to the rotation of the target presented in direction θ. To measure the ability of the network to generalize to targets in other directions, we calculate the noiseless error for test target (E_{test}), presented in a direction $\theta'\neq\theta$ and define the generalization error by:

$$G.E.=1-E_{test}/E_0 \tag{18}$$

In the limit $\epsilon\to 0$ (assuming $\bar{\eta}<2$), E_{test} can be computed analytically, as function of $\Delta\theta=\theta-\theta'$. Using Eq. (15) one finds:

$$E_{test}=2(1-\cos\gamma)(1-2\hat{\alpha}(\Delta\theta)\cos\Delta\theta+\hat{\alpha}^2(\Delta\theta))$$

and

$$G.E.=2\hat{\alpha}(\Delta\theta)\cos\Delta\theta-\hat{\alpha}^2(\Delta\theta) \tag{19}$$

where $\hat{\alpha}(\Delta\theta)$ is:

$$\hat{\alpha}(\Delta\theta)=\frac{1}{N\alpha}\boldsymbol{F}^T(\theta)\boldsymbol{F}(\theta') \tag{20}$$

depends on θ and θ' only via $\Delta\theta$. Note that $\hat{\alpha}(0)=1$. Specifically, in the limit of large N and when using the tuning curve function in Eq. (2), one gets:

$$\hat{\alpha}(\Delta\theta)=\frac{I_0(x/\rho)}{I_1(2/\rho)} \tag{21}$$

where $x=\sqrt{2(1+\cos(\Delta\theta))}$.

Adaptation to two targets

How does a reward affect the next trial?. Here we consider the case where the network adapts to two targets in the direction θ and θ' presented in alternation. If a rewarded trial occurs for one of the targets, the connectivity matrix is updated, affecting the noiseless error on the next trial when the other target is presented.

This noiseless error can be computed in the limit $\epsilon\to 0$. It is a good estimate for the noiseless error in the beginning of the adaptation with finite ϵ, where the error is still big with respect to the target size. Let us assume that on trial k a target in direction θ is presented and that it is rewarded. This condition fully determines the realization of the noise on trial k, $\hat{\boldsymbol{\xi}}(k)$. The noiseless errors for the two targets following the rewarded trial, denoted $E_0^{\theta}(k)$ and $E_0^{\theta'}(k)$, can be determined analytically. One finds:

$$\begin{aligned} E_0^{\theta}(k) &= (1-\bar{\eta})^2 E_0(k-1) \\ E_0^{\theta'}(k) &= ||E_0^{\theta'}(k-1) - \bar{\eta}\hat{\alpha}(\Delta\theta)E_0^{\theta}(k-1)||^2 \\ &= E_0^{\theta'}(k-1) + \\ &\quad \sqrt{E_0^{\theta}}(k-1)\bar{\eta}\hat{\alpha}(\Delta\theta)(\bar{\eta}\hat{\alpha}(\Delta\theta)\sqrt{E_0^{\theta}}(k-1) - \\ &\quad 2\sqrt{E_0^{\theta'}}(k-1)\cos\Delta\theta) \end{aligned} \tag{22}$$

If $\bar{\eta}<2$, $E_0^{\theta}(k)<E_0^{\theta}(k-1)$, that is, the noiseless error for the target that has been rewarded decreases following the update of the connectivity matrix. For the other target (direction θ'), the effect of this update on the noiseless errors depends on the sign of the expression in parentheses in the second equation. If the two targets are in opposite directions, it is always positive and $E_0^{\theta'}(k)>E_0^{\theta'}(k-1)$. Thus, while the network performs better for one of the targets it performs worse for the other target. We term this situation *destructive* interference. On the other hand, if the targets are close such that the expression in parentheses is negative, $E_0^{\theta'}(k)<E_0^{\theta'}(k-1)$. In other words, if the network improves for one of the targets it also improves for the other target. We term this situation *constructive* interference.

In particular, for the first rewarded trial, using $E_0^{\theta}(0)=E_0^{\theta'}(0)$, we get:

$$E_0^{\theta'}(k+1)=E_0^{\theta'}(k)(1-Q(\bar{\eta},\rho,\Delta\theta))$$

where:

$$Q(\bar{\eta},\rho,\Delta\theta)=2\bar{\eta}\hat{\alpha}(\Delta\theta)\cos\Delta\theta-\bar{\eta}^2\hat{\alpha}^2(\Delta\theta) \tag{23}$$

We expect a constructive interference for $Q(\bar{\eta},\rho,\Delta\theta)>0$ and destructive interferences otherwise. Note that for $\bar{\eta}=1$ the interference function equals to the generalization error function (Eq. (20)). The transition between the constructive and destructive regimes is given by:

$$\cos(\Delta\theta)=\frac{\bar{\eta}\hat{\alpha}(\Delta\theta)}{2}$$

The quantity $Q(\bar{\eta},\rho,\Delta\theta)$) characterizes the strength of the interference. It can be a non-monotonous function of $\Delta\theta$. To show this, we calculate the derivative of $Q(\bar{\eta},\rho,\Delta\theta)$ with respect to $\Delta\theta$, using Eq. (22). This derivative changes sign when:

$$\cos(\Delta\theta)+\rho r\frac{I_0(r/\rho)}{I_1(r/\rho)}-\frac{I_0(r/\rho)}{I_0(2/\rho)}=0 \tag{24}$$

For instance, when $\bar{\eta}=0.3$ and $\rho<0.73$ the function $Q(\bar{\eta},\rho,\Delta\theta)$ depends non-monotonically on $\Delta\theta$. In other words, for sufficiently narrow tuning curves, the interference varies non-monotonically with the angular difference. However, this non-monotonicity effect can be very small: if the tuning curves are too narrow, $Q(\bar{\eta},\rho,\Delta\theta)$ quickly reaches zero when increasing $\Delta\theta$.

Numerical simulations

In the numerical simulations described in this paper, the input layer consists of $N=100$ neurons. We normalized the tuning curves (parameter C in Eq. (2)) such that α remains constant ($\alpha=0.36$) when changing ρ. This was done to guarantee that the time to learn one target does not depend on the tuning width.

Learning duration and final error. We define the final error of the network as the median of the error over the last 1,000 trials of the simulation for each realization. We then determine the learning duration, τ_L, as the trial number at which the filtered signal (median filter with a window length of 50 trials) crosses a threshold, defined to be 5% above the final error. In order to avoid boundary problems of the filter at time 0 (the discontinuity in the error when we induce the rotation), we calculate the error at $t<0$ while assuming that the cursor is already rotated (even though it did not). In the figures we plot the actual error before the rotation, which is small. Similar results were obtained using a linear filter.

Time dependent correlations between the errors for two targets. When the network adapts to a rotation for two targets presented in alternation in consecutive trials, the learning processes for the two targets interfere. This interference can be quantified by considering the correlations between the errors on two consecutive trials:

$$CC(t)=\frac{\langle(E_{\xi}(t)-\langle E_{\xi}(t)\rangle)(E_{\xi}(t+1)-\langle E_{\xi}(t+1)\rangle)\rangle}{\sqrt{Var(E_{\xi}(t))Var(E_{\xi}(t+1))}}$$

The brackets denote the average over repetitions of the adaptation process, which differ by the realization of the noise. Negative $CC(t)$ indicates that if the network improves for one target it deteriorates for the other target (destructive interference). Positive $CC(t)$ corresponds to constructive interference.

Performance and noiseless performance. We ran long simulations of 10^7 trials to estimate the performance and noiseless performance after the transient learning phase. The performance is given by:

$$\langle\Theta(\epsilon-E_{\xi}(t))\rangle_t \tag{25}$$

and the noiseless performance is given by:

$$\langle\Theta(\epsilon-E_0(t))\rangle_t \tag{26}$$

where $\Theta(x)$ is the Heaviside function and the average is over time, when the transient learning phase was excluded.

Supporting Information

Figure S1 Delayed learning effect with an on-line gradient ascent algorithm. **A.** Delayed learning in a reward function that varies abruptly with the error ($T=0.04$). **B.** The delayed learning is reduced for a smoother reward function. ($T=0.05$). **C.** The delayed learning almost disappears when the reward function is smoothed even further ($T=0.067$). Note the change of scale in the abscissa. Parameters: $\hat{\eta}=0.1$, $c=0.05$, $\rho=1$.
(EPS)

Figure S2 Delayed learning effect in a 30^o rotation for two targets in the 3-layer network. The reach angle (in degrees) is plotted as a function of the trial number. The shaded area corresponds to the target size. Initial conditions as explained in the text. Parameters: $\sigma=0.2$, $\epsilon=0.1$, $\eta=0.1$, $\rho=1$.
(EPS)

Figure S3 Gradual adaptation for an 8° rotation. **A.** Reach angle (in degrees) as a function of the trial number when the rotation angle is increased by 1° every 40 trials up to 8°. The shaded area corresponds to the target size ($\pm 3^{\circ}$ around the target

center). $\sigma = 0.06$. **B.** The generalization error, given as the change in reach angle. The learned target is at 0°. Circles : simulation results. For clarity, the results are displayed for test targets sampled every 2.5 degrees. Solid line: analytical results. Shaded area corresponds to the standard deviation in generalization error in numerical simulations estimated over 100 repetitions. Number of neurons in the input layer: $N = 500$. **C.** The shape of the tuning curves that was used in (**A**) and (**B**): $f(\theta_i - \theta) = C(a + \exp(\frac{cos(\theta_i - \theta) - 1}{\rho}))$, where C is a normalization constant (see Materials and Methods), $a = 0.14$, $\rho = 0.005$.
(EPS)

Figure S4 Learning duration when adapting to multiple targets varies monotonically with the number of learned targets when using a gradient descent on a quadratic error function. Total number of target presentations required to learn the entire task *vs.* the number of presented targets, *m*. The targets are evenly distributed (between 0° to 360°). Learning duration was calculated as the trial number at which learning curve crossed a threshold of $15 \quad 10^{-4}$. Color coded as in Figure 11 in the Results. Black: $\rho = 0.05$. Blue: $\rho = 0.1$. Purple: $\rho = 0.3$. Green: $\rho = 0.4$. Dashed black line corresponds to learning the targets independently. Compare with Figure 11 in main text.
(EPS)

Text S1 This document is a supporting text for the supplementary figures.
(PDF)

Acknowledgments

We thank Yonatan Loewenstein, Dana Barniv, Mehdi Khamassi, German Mato and Carl van Vreeswijk for fruitful discussions and critical reading of the manuscript of this paper. We also thank Itay Novick, David Perkel and Eilon Vaadia for illuminating discussions.

Author Contributions

Conceived and designed the experiments: RD DH AL. Performed the experiments: RD. Analyzed the data: RD DH AL. Contributed reagents/materials/analysis tools: RD DH AL. Wrote the paper: RD DH AL.

References

1. Pouget A, Snyder L (2000) Computational approaches to sensorimotor transformations. Nature Neuroscience 3: 1192–1198.
2. Piaget J, Cook M (1953) The origin of intelligence in the child. London: Routledge & Kegan Paul.
3. Krakauer J, Pine Z, Ghilardi M, Ghez C (2000) Learning of visuomotor transformations for vectorial planning of reaching trajectories. The Journal of Neuroscience 20: 8916–8924.
4. Thoroughman K, Shadmehr R (2000) Learning of action through adaptive combination of motor primitives. Nature 407: 742–747.
5. Chou I, Lisberger S, et al. (2002) Spatial generalization of learning in smooth pursuit eye movements: implications for the coordinate frame and sites of learning. The Journal of Neuroscience 22: 4728–4739.
6. Linkenhoker B, Knudsen E (2002) Incremental training increases the plasticity of the auditory space map in adult barn owls. Nature 419: 293–296.
7. Sober S, Brainard M (2009) Adult birdsong is actively maintained by error correction. Nature neuroscience 12: 927–931.
8. Houde J, Jordan M (1998) Sensorimotor adaptation in speech production. Science 279: 1213–1216.
9. Sober S, Brainard M (2012) Vocal learning is constrained by the statistics of sensorimotor experience. Proceedings of the National Academy of Sciences 109: 21099–21103.
10. Fiete I, Fee M, Seung H (2007) Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. Journal of Neurophysiology 98: 2038–2057.
11. Rokni U, Richardson A, Bizzi E, Seung H (2007) Motor learning with unstable neural representations. Neuron 54: 653–666.
12. Poggio T, Bizzi E (2004) Generalization in vision and motor control. Nature 431: 768–774.
13. Tanaka H, Sejnowski T, Krakauer J (2009) Adaptation to visuomotor rotation through interaction between posterior parietal and motor cortical areas. Journal of Neurophysiology 102: 2921–2932.
14. Taylor J, Hieber L, Ivry R (2013) Feedback-dependent generalization. Journal of Neurophysiology 109: 202–215.
15. Tumer E, Brainard M (2007) Performance variability enables adaptive plasticity of crystallized adult birdsong. Nature 450: 1240–1244.
16. Andalman A, Fee M (2009) A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. Proceedings of the National Academy of Sciences 106: 12518–12523.
17. Izawa J, Shadmehr R (2011) Learning from sensory and reward prediction errors during motor adaptation. PLoS computational biology 7: e1002012.
18. Shmuelof L, Huang V, Haith A, Delnicki R, Mazzoni P, et al. (2012) Overcoming motor forgetting through reinforcement of learned actions. The Journal of Neuroscience 32: 14617–14621.
19. Huang V, Haith A, Mazzoni P, Krakauer J (2011) Rethinking motor learning and savings in adaptation paradigms: model-free memory for successful actions combines with internal models. Neuron 70: 787–801.
20. Paz R, Boraud T, Natan C, Bergman H, Vaadia E (2003) Preparatory activity in motor cortex reflects learning of local visuomotor skills. Nature Neuroscience 6: 882–890.
21. Warren T, Tumer E, Charlesworth J, Brainard M (2011) Mechanisms and time course of vocal learning and consolidation in the adult songbird. Journal of Neurophysiology 106: 1806–1821.
22. Skinner B (1967) Science and human behavior. Free Press.
23. Lawrence D (1952) The transfer of a discrimination along a continuum. Journal of Comparative and Physiological Psychology 45: 511.
24. Terrace H (1963) Discrimination learning with and without errors. Journal of the Experimental Analysis of Behavior 6: 1.
25. Kangas B, Bergman J (2012) A novel touch-sensitive apparatus for behavioral studies in unrestrained squirrel monkeys. Journal of Neuroscience Methods 209: 331–6.
26. Ng A, Harada D, Russell S (1999) Policy invariance under reward transformations: Theory and application to reward shaping. In: Machine learning: proceedings of the Sixteenth International Conference (ICML'99). Morgan Kaufmann Pub, p. 278.
27. Randløv J (2000) Shaping in reinforcement learning by changing the physics of the problem. In: Proc. 17th International Conf. on Machine Learning. pp. 767–774.
28. Krueger K, Dayan P (2009) Flexible shaping: How learning in small steps helps. Cognition 110: 380–394.
29. Kerr J, Wickens J (2001) Dopamine d-1/d-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. Journal of Neurophysiology 85: 117–124.
30. Ding L, Perkel D (2004) Long-term potentiation in an avian basal ganglia nucleus essential for vocal learning. The Journal of Neuroscience 24: 488–494.
31. Reynolds J, Hyland B, Wickens J (2001) A cellular mechanism of reward-related learning. Nature 413: 67–70.
32. Loewenstein Y, Seung H (2006) Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proceedings of the National Academy of Sciences 103: 15224–15229.
33. Reynolds JN, Wickens JR (2002) Dopamine-dependent plasticity of corticostriatal synapses. Neural Networks 15: 507–521.
34. Legenstein R, Chase S, Schwartz A, Maass W (2010) A reward-modulated hebbian learning rule can explain experimentally observed network reorganization in a brain control task. The Journal of Neuroscience 30: 8400–8410.
35. Frémaux N, Sprekeler H, Gerstner W (2010) Functional requirements for reward-modulated spike-timing-dependent plasticity. The Journal of Neuroscience 30: 13326–13337.
36. Williams R (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning 8: 229–256.
37. Fiete I, Seung H (2006) Gradient learning in spiking neural networks by dynamic perturbation of conductances. Physical review letters 97: 48104.
38. Werfel J, Xie X, Seung H (2005) Learning curves for stochastic gradient descent in linear feedforward networks. Neural computation 17: 2699–2718.
39. Mandelblat-Cerf Y, Paz R, Vaadia E (2009) Trial-to-trial variability of single cells in motor cortices is dynamically modified during visuomotor adaptation. The Journal of Neuroscience 29: 15053–15062.
40. Ölveczky B, Otchy T, Goldberg J, Aronov D, Fee M (2011) Changes in the neural control of a complex motor sequence during learning. Journal of Neurophysiology 106: 386–397.
41. Körding K, Wolpert D (2004) The loss function of sensorimotor learning. Proceedings of the National Academy of Sciences of the United States of America 101: 9839–9842.
42. Schultz W (1998) Predictive reward signal of dopamine neurons. Journal of neurophysiology 80: 1–27.
43. Izawa J, Kondo T, Ito K (2004) Biological arm motion through reinforcement learning. Biological cybernetics 91: 10–22.

44. Law C, Gold J (2009) Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nature Neuroscience 12: 655–663.
45. Vladimirskiy B, Vasilaki E, Urbanczik R, Senn W (2009) Stimulus sampling as an exploration mechanism for fast reinforcement learning. Biological cybernetics 100: 319–330.
46. Taylor M, Stone P (2009) Transfer learning for reinforcement learning domains: A survey. The Journal of Machine Learning Research 10: 1633–1685.
47. Laud A, DęJong G (2003) The influence of reward on the speed of reinforcement learning: An analysis of shaping. In: Proc. 12th International Conf. on Machine Learning (ICML-2003). volume 20, p. 440.
48. Fiete I, Hahnloser R, Fee M, Seung H (2004) Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. Journal of Neurophysiology 92: 2274–2282.
49. Donchin O, Francis J, Shadmehr R (2003) Quantifying generalization from trial-by-trial behavior of adaptive systems that learn with basis functions: theory and experiments in human motor control. The Journal of neuroscience 23: 9032–9045.
50. Gale S, Perkel D (2010) Anatomy of a songbird basal ganglia circuit essential for vocal learning and plasticity. Journal of chemical neuroanatomy 39: 124–131.
51. Gradshteĭn I, Ryzhik I, Jeffrey A, Zwillinger D (2007) Table of integrals, series, and products. Elsevier academic press.