

Motor learning with unstable neural representations

U. Rokni¹, A. G. Richardson², E. Bizzi³ and H. S. Seung¹

¹Brain & cognitive sciences & Howard Hughes Medical Institute, MIT, ²Division of Health Sciences and Technology, MIT and Harvard Medical School,

³Department of Brain & Cognitive Sciences and McGovern Institute for Brain Research, MIT

It is often assumed that learning takes place by changing an otherwise stable neural representation. To test this assumption, we analyzed data of a recent experiment, which measured changes in the directional tuning of primate motor cortical neurons during reaching movements, performed in both familiar and novel environments (Padoa-Schioppa et al., J Neurophysiol. 2004). During the familiar task, tuning curves exhibited slow random drift. During learning of the novel task, random drift was accompanied by systematic learning-related changes. Our analysis suggests that motor learning is based on a neural representation which is surprisingly unstable.

To explain the observed instability of the neural representations we propose a theory which is based on two assumptions: (1) motor cortex is *redundant*, i.e. it uses more neurons than required to produce the desired sensorimotor transformation, (2) when practicing a task, synapses are changed by a stochastic gradient descent learning rule. We demonstrate that these two assumptions are sufficient to explain the observed instability by simulating a model of a redundant motor cortical network. Because our network has more neurons than necessary, a desired sensorimotor transformation can be realized by a continuum of configurations of synaptic strengths, which we term the *optimal manifold*. Changing synapses along the optimal manifold changes the neural representation, but not the sensorimotor transformation. Our second assumption, of stochastic gradient learning, produces such behaviorally irrelevant changes. These learning dynamics can be described as stochastically moving down an error landscape which has a valley of minima at the optimal manifold (see schematic figure below). At the late stages of learning, noise driven changes are channeled along this valley by the gradient term, and thus the neural representation drifts although performance is nearly fixed.

We show that our model accounts reasonably well for the observed changes in the neural representations, in both familiar and novel environments. Furthermore, we infer several properties of synaptic plasticity underlying motor learning: (1) signal-to-noise ratio is around 1, (2) the source of variability is local, i.e. at individual synapses or neurons, rather than global noise from the environment, (3) plasticity noise is additive, and (4) it takes plasticity noise at least thousands of trials to change synapses completely. Additionally, we show that contrary to common views, a cell's tuning properties may be only weakly related to its anatomical connections to the motor output. Finally, we discuss what other evidence there are for our theory and how it could be further tested with brain computer interface experiments.

