Normalization of input patterns in an associative network

Andreas Liu and Wade G. Regehr

Department of Neurobiology, Harvard Medical School, Boston, Massachusetts

Submitted 19 September 2013; accepted in final form 11 November 2013

Liu A, Regehr WG. Normalization of input patterns in an associative network. J Neurophysiol 111: 544–551, 2014. First published November 13, 2013; doi:10.1152/jn.00678.2013.—Numerous brain structures have a cerebellum-like architecture in which inputs converge onto a large number of granule cells that converge onto principal cells. Plasticity at granule cell-to-principal cell synapses is thought to allow these structures to associate spatially distributed patterns of granule cell activity with appropriate principal cell responses. Storing large sets of associations requires the patterns involved to be normalized, i.e., to have similar total amounts of granule cell activity. Using a general model of associative learning, we describe two ways in which granule cells can be configured to promote normalization. First, we show how heterogeneity in firing thresholds across granule cells can restrict pattern-to-pattern variation in total activity while also limiting spatial overlap between patterns. These effects combine to allow fast and flexible learning. Second, we show that the perceptron learning rule selectively silences those synapses that contribute most to pattern-to-pattern variation in the total input to a principal cell. This provides a simple functional interpretation for the experimental observation that many granule cell-to-Purkinje cell synapses in the cerebellum are silent. Since our model is quite general, these principles may apply to a wide range of associative circuits.

several brain structures share a distinctive circuit motif consisting of a mass of granule cells that converge and form synapses onto large principal cells. Among these structures are the cerebellum and dorsal cochlear nucleus in mammals, the electrosensory lateral line lobe of weakly electric fish, and the insect mushroom body. They appear to serve extremely diverse functions, including motor control, sound localization, olfactory learning, and cancellation of responses to predictable modulations of self-generated electric fields (Bell et al. 2008; Bol et al. 2011; Farris 2011; Oertel and Young 2004). However, the common architecture of these circuits suggests that they perform a common computation.

In the cerebellum, granule cells form parallel fiber synapses onto Purkinje cells, which are the only output cells of the cerebellar cortex. The classic proposal for describing computation in the cerebellum is the Marr-Albus-Ito hypothesis, which states that plasticity at parallel fiber synapses allows Purkinje cells to associate spatially distributed patterns of granule cell activity with appropriate responses (Albus 1971; Ito 1984; Marr 1969). When many such associations must be stored at once, the patterns involved must meet two requirements. First, to minimize interference between superposed associations, different patterns should show activity in minimally overlapping sets of granule cells (Kanerva 1988; Willshaw et al. 1969). Second, the total amount of granule cell activity should be similar across patterns, otherwise patterns with high overall activity will drive inappropriately large principal cell responses.

The latter requirement led early theorists of the cerebellum to suggest that a major function of Golgi cells and other inhibitory interneurons is to insulate spike generation in Purkinje cells from the effect of variation in total granule cell activity (Albus 1971; Marr 1969). More recent theoretical work has examined the consequences of this variation by representing the principal cell as a perceptron that is trained to associate specific input patterns with specific responses (Legenstein and Maass 2008). In this formalism, each pattern is a snapshot of activity in the granule cell population, and the summed activity over all granule cells is called the L1 norm of the pattern. Reflecting the fact that granule cells are excitatory, all synaptic weights are constrained to be nonnegative. In the resulting model, the number of binary classifications of input patterns that the principal cell can perform is greatly reduced if the patterns have different L1 norms (Legenstein and Maass 2008). Conversely, uniformity of L1 norms is a sufficient condition for all binary classifications to be possible in the model, provided that patterns do not outnumber input units (Legenstein and Maass 2008). These results show the importance of normalizing input patterns, i.e., of making L1 norms as uniform as possible before signals reach the principal cell.

How might a circuit with a cerebellum-like architecture accomplish this? The classic proposal is that inhibition from Golgi cells could keep the overall level of granule cell activity roughly constant across patterns (Albus 1971; Brickley et al. 1996; Hamann et al. 2002; Marr 1969). This form of normalization is likely to be incomplete, and dynamic modulation of such inhibition is likely to be delayed relative to the granule cell response by the need to traverse an extra synapse (Crowley et al. 2009). Here, we investigate the possibility that granule cells themselves could also be configured to promote normalization. We use a model associative network to show how this can occur at two stages of signal processing. First, we examine how the input-output functions computed by individual granule cells determine the statistics of the patterns presented to the principal cell. We show how assigning different firing thresholds to different granule cells facilitates the storage of large sets of associations by allowing patterns to have similar L1 norms but also little spatial overlap. Second, we show that the perceptron learning rule selectively silences those synaptic connections that contribute most to pattern-to-pattern variation in the total input to the principal cell. This form of normalization may be an important functional consequence of the widespread silencing observed at parallel fiber synapses in the cerebellum (Ekerot and Jornwall 2001, 2003; Isope and Barbou 2002). Since our model embodies only a few very general features of cerebellum-like structures, the design principles
described here may apply to a wide range of associative circuits.

METHODS

The model circuit and training procedure are described in RESULTS. Here, we give a more explicit account of certain details.

Generation of Granule Cell Activity Patterns

The granule cell activity patterns used in the simulations are generated from random numbers as described in Fig. 1B and the accompanying text. The simulations in Fig. 4 are based on granule cell patterns produced by the Ramp network (Fig. 2A). Using patterns produced by the Step network (Fig. 2A) leads to similar results. For some simulations in Fig. 4E, granule cell patterns produced by the Ramp network are modified to have an L1 norm of 50. This is half of the largest possible value, as the activity of a granule cell is always a number between 0 and 1. The normalization is carried out as follows. If the L1 norm of the pattern exceeds 50, the activity of the rth granule cell \(a_r\) is lowered by:

\[
\Delta a_r = \frac{a_r}{|P|_1} (50 - |P|_1),
\]

where \(|P|_1\) is the L1 norm of the pattern. Similarly, if the L1 norm of the pattern is <50, the \(a_r\) is raised by:

\[
\Delta a_r = (1 - a_r) / (100 - |P|_1) (50 - |P|_1).
\]

This ensures that the activity of each granule cell remains between 0 and 1.

Training Procedure

At the start of each simulation, synaptic weights are uniformly randomly distributed between 0 and 1 (Fig. 4A). In all five networks studied, the activity of each granule cell is a number between 0 and 1, distributed symmetrically around 0.5. Thus the average pattern induces a total synaptic drive of 25 before training. The goal of training is to associate each pattern with a randomly assigned target level of synaptic drive. Targets are (possibly overlapping) intervals of width 2.5 with centers that are uniformly randomly distributed between 20 and 30. A “training step” is a complete cycle in which each pattern is presented once. In all simulations, the learning rate (\(\eta\)) is 0.005. Reducing this value does not increase the probability of convergence or change the relative performance of the five networks according to our criteria.

Because the activity of each cell in the Step and Steps networks is all or none, the learning rule makes larger individual adjustments to synaptic weights in these networks than in the networks with graded activity. This could make it inappropriate to compare networks using convergence times: in particular, it could be a trivial reason for the relatively fast convergence seen in the Step and Steps networks (Fig. 2C). To examine this possibility, we monitored how synaptic weights changed during a typical training step in which each pattern was presented in turn and synapses underwent multiple adjustments. We considered the path length traveled by a typical synaptic weight during the training step (the sum of the absolute values of the adjustments) as well as the net change in synaptic weight (the absolute value of the sum of the adjustments). We found that in the Step and Steps networks, path lengths were shorter but net changes larger than in the networks with graded activity. This indicates that Step and Steps converged faster not because individual weight adjustments were larger but because weights in these networks approached their final values in a more directed fashion.

RESULTS

Description of the Model

In cerebellum-like structures, the granule cells form an intermediate layer that integrates and transforms input from other brain regions before it reaches the principal cells. Our highly simplified model is designed to test the effect of various possible transformations on associative learning. The model circuit contains 64 afferents, 100 granule cells, and 1 principal cell. Each granule cell sums inputs from 4
We investigated how granule cells compute the same functions, which is the mean of all the step functions. The last 3 assignments are related in the same way as the 1st 3. Right: sample granule cell patterns resulting from transformations of the same afferent pattern. B: a set of afferent patterns was fed into a range of networks, including the networks described in A. The resulting sets of granule cell patterns had different statistical properties. Abscissa: coefficient of variation (c.v.) of the distribution of L1 norms. Ordinate: average Euclidean distance between 2 patterns normalized to the value for the Step network. The upper and lower curves show results from heterogeneous and homogeneous networks, respectively. Curves are parameterized by the slope of the input-output functions in the network. Colored markers represent the same networks as in A. C: cumulative histograms of convergence times in 300 trials for training sets containing 15, 20, or 25 patterns. The same afferent patterns were presented to the 5 networks, and the resulting granule cell patterns were used to train the principal cell. D: probability of convergence after 10,000 training steps as a function of the number of patterns to be stored. E: final weights for the Steps (red) and Ramp (black) networks after training with 5 or 25 patterns. Results from 3 pairs of trials are overlaid. Graphs have different scales, allowing correlations to be compared. F: left: summary plot of synaptic weight correlations after training as in B. This arrangement is based on the convergence ratio of mossy fibers to granule cells in the cerebellum as well as the electrotonic compactness of cerebellar granule cells (Mitchell and Silver 2003). Connections from afferents to granule cells are drawn randomly with no restriction on overlap. The likelihood of two granule cells having no overlap. The likelihood of two granule cells having no overlap. The likelihood of two granule cells having no overlap. The likelihood of two granule cells having no overlap.

Several simplifications in the model should be noted at the outset. First, we chose to represent a relatively small subset of the possible granule cells since the number of granule cells in the model has little effect on the distribution of L1 norms of granule cell patterns (see below). This is because the granule cells are not activated independently but are driven by a common set of afferents. Any nonuniformity in L1 norms is largely inherited from these afferents and cannot be smoothed over merely by adding granule cells to the model. Second, synapses from afferents to granule cells have fixed and uniform weights in the model, although their biological counterparts are heterogeneous (DiGregorio et al. 2002; Sargent et al. 2005; Xu-Friedman and Regehr 2003) and modifiable (D’Angelo et al. 1999; Nieus et al. 2006). Finally, to maintain the focus on the role of granule cells in normalization, inhibitory interneurons such as Golgi cells and stellate cells are not represented explicitly.

In the model, associative learning operates on patterns of granule cell activity, which we represent as vectors with 100 entries that specify the output level of each cell. To define a granule cell pattern, we 1st create an afferent pattern by drawing the activity of each afferent randomly from a uniform distribution (Fig. 1B, left). For each granule cell, the summed input from its 4 afferents across all patterns thus forms a bell-shaped distribution (Fig. 1B, middle). In what follows, we shall define various granule cell input-output functions with reference to this distribution. These functions determine how afferent patterns are transformed...
into the granule cell patterns that are presented to the principal cell.

Each granule cell pattern induces a synaptic drive, defined as the dot product of the pattern with the vector of synaptic weights from granule cells onto the principal cell. These weights are adjustable, allowing the network to store associations between arbitrary patterns of granule cell activity and arbitrary levels of principal cell output. The presumed common task of cerebellum-like structures in vivo is to store continuous sequences of such associations. In the model, the goal of training is to associate each pattern in a set with a synaptic drive that lies within an error margin of the randomly assigned target of that pattern (Fig. 1C). Training is governed by the perceptron learning rule (Hertz et al. 1991) with synaptic weights constrained to be nonnegative. Thus each pattern is presented in turn, and each weight $w_j$ is updated according to the rule:

$$\Delta w_j = \pm \eta a_i,$$

where the sign is positive if the synaptic drive of that pattern is too low and negative if too high. Any negative weights are adjusted back to 0. The synaptic drive of the next pattern is computed using the updated weights, and the process is repeated for each pattern in turn. The entire cycle is repeated until the process converges, i.e., each pattern is mapped to its assigned target.

**Normalization Through Cellular Heterogeneity**

How might the granule cell layer transform afferent patterns to facilitate associative learning? Such a transformation should satisfy two demands that appear to conflict. On the one hand, it has long been recognized that associations involving spatially distributed patterns are most easily stored if different patterns show activity in minimally overlapping sets of granule cells (Kanerva 1988; Willshaw et al. 1969). A natural measure of overlap is the Euclidean distance between two points in high-dimensional space where each point represents a pattern. Ideally, patterns used in associative learning should be separated by large Euclidean distances as it is easier to discriminate between such dissimilar patterns. On the other hand, they should have similar L1 norms, i.e., the summed activity over all granule cells should be similar across patterns. This makes it easier for the principal cell to produce learned responses as its output is no longer controlled by variations in aggregate granule cell activity. In this section, we look for a transformation that makes L1 norms similar but Euclidean distances large.

We first considered homogeneous networks in which all granule cells have the same input-output function. In one such network, called Step, each granule cell computes a step function for which the threshold is the mean of the bell-shaped input distribution (Fig. 2A). In another network, called Ramp, the function includes a ramp with endpoints lying 1 SD above and below the mean (Fig. 2A). Cells in Ramp produce graded responses, whereas those in Step respond maximally or not at all to a given afferent pattern (Fig. 2A). When a set of afferent patterns is fed into both networks, the resulting granule cell patterns in Step show greater Euclidean separation and less uniform L1 norms than those in Ramp (Fig. 2B). More generally, increasing the steepness of the input-output function reduces overlap between patterns as measured by Euclidean distances but also widens the distribution of L1 norms (Fig. 2B, lower curve).

To escape this tradeoff, we were led to consider networks with heterogeneous granule cell layers. In Steps, different granule cells compute different step functions for which thresholds are uniformly distributed within 1 SD of the mean (Fig. 2A). Since the mean input-output function is the same in Steps and Ramp, these networks transform a given afferent pattern into granule cell patterns with nearly identical L1 norms. Thus the distribution of L1 norms is similar in the two networks. However, granule cell patterns in Steps show greater Euclidean separation than those in Ramp (Fig. 2B). More generally, a heterogeneous network produces patterns with less overlap than a homogeneous network with the same distribution of L1 norms (Fig. 2B, upper and lower curves). This is because the mean of the input-output functions determines the distribution of L1 norms, whereas the shapes of the individual functions largely determine the typical Euclidean distance between patterns. Assigning different thresholds to different cells allows the individual functions to be steeper than the mean, making L1 norms similar but Euclidean distances large.

To test the relevance of this analysis, we performed parallel simulations in which identical afferent patterns were presented to the five networks shown in Fig. 2A. Each of the five resulting sets of granule cell patterns was then used for training as described in the previous section. We compared outcomes in two ways: 1) speed and probability of convergence; and 2) elasticity of final synaptic weights.

**Speed and probability of convergence.** Whereas some sets of patterns led to fast convergence, other sets of the same size could not be stored even after several thousand training steps, so there was no obvious measure of the capacity of a network. We therefore compared networks using the distribution of times to convergence as well as the probability that a training run would eventually converge, which was always stable after 10,000 steps (Fig. 2, C and D). The results can be summarized in 2 observations. First, networks with steeper input-output functions converged faster and with higher probability. This is because patterns overlap less in these networks, creating less interference and allowing synaptic weights to approach their final values with fewer deflections (see METHODS). Second, as predicted by our analysis, heterogeneous networks (e.g., Steps)
converged faster and with higher probability than homogeneous networks with the same mean or median input-output function (e.g., Ramp or Step). This held both when the functions were steep, as in the first three networks in Fig. 2A, and when they were shallower, as in the last three networks.

Elasticity of final synaptic weights. Another way to measure performance is to compare final states reached by the network from different initial states chosen at random. The synaptic weights in a network define a point in a weight space of 100 dimensions, and the learning rule directs a search through this space for a point that solves the problem of storing a given set of associations. The difficulty of the problem can be measured by the degree of similarity between solutions reached from unrelated starting points: the closer these endpoints tend to be, the more restricted is the class of solutions. To probe the weight space in our model, we ran each simulation twice, starting from different sets of 100 randomly chosen synaptic weights. Starting weights for the 2 trials were uncorrelated (Fig. 2E, left). After a set of patterns was stored, the synaptic weights were compared again (Fig. 2E, middle and right). As the number of patterns was increased, the final weights became more correlated as solutions became scarcer. Beyond 30 patterns, all the networks were overloaded, and most trials did not converge.

Up to 30 patterns, we considered only trials that did converge and found that networks with faster convergence also had less correlated sets of final weights. Thus both criteria gave the same ranking to the 5 networks (Fig. 2F). In particular, final weights were less correlated in heterogeneous networks than in homogeneous networks with the same mean or median input-output function. We conclude that assigning different thresholds to different granule cells expands the class of solutions to the learning problem.

Since this procedure captures only solutions accessible to the learning rule, we cannot exclude the possibility that a bias toward certain preferred solutions might affect the observed correlations between sets of final weights. However, the strong dependence of these correlations on the number of patterns to be stored argues that they do measure the difficulty of the learning problem.

Our criteria agree in identifying Steps as the most effective design. Why is Steps more effective than Step, where the patterns overlap less? Our analysis suggests that the greater uniformity of L1 norms in Steps can compensate for the extra overlap. This can be seen in a simulation where a set of associations is first stored, after which one new association must be learned without disturbing the existing mapping of the other patterns to their targets. While the new association is being learned, we track the error, i.e., the difference between the target synaptic drive of the new pattern and its actual synaptic drive during each training step (Fig. 3). The smaller initial error in Steps and Ramp reflects the greater uniformity of L1 norms in these networks compared with Step. The faster approach to the target in Step and Steps reflects the larger Euclidean distances between patterns in these networks compared with Ramp, resulting in less interference between the weight adjustments needed to maintain the stored associations and those needed to learn the new one. Of the three networks, only Steps enjoys both a small initial error and a fast elimination of that error.

Normalization Through Synaptic Silencing

In the cerebellum, it has been observed that many synapses from granule cells to Purkinje cells are silent (Ekerot and Jorntell 2001, 2003; Iospe and Barbour 2002). Theoretical work has also shown that the optimal distribution of synaptic weights in a perceptron with nonnegative weights contains a large fraction of zeroes, i.e., silent synapses (Brunel et al. 2004). In this section, we reproduce the phenomenon in a simple model, identify a criterion that determines whether a given synapse in the model will become silent during training, and show how the process contributes to normalization.

We performed simulations using granule cell patterns produced by the Ramp network (Fig. 2A). Using patterns produced by the Step network (Fig. 2A) led to similar results. Each simulation began with uniformly distributed synaptic weights (Fig. 4A). To maintain a rough balance between upward and downward weight adjustments, target values were assigned so that the mean synaptic drive across patterns remained constant during training (see Methods). When 15 patterns were presented, nearly half of the synapses became silent, whereas others developed large weights to compensate (Fig. 4A). More generally, the fraction of synapses suppressed was almost linear in the number of patterns (Fig. 4B).

Which property of a synapse determines whether it will be silenced? For each granule cell, we define the “conformity” of the cell as the correlation, computed across all patterns in the training set, between the activity of the cell and the L1 norm of the cell. This property measures the extent to which the activity of a given cell resembles that of the population. In particular, cells with high conformity are relatively active in patterns for which L1 norms are large. We reasoned that synapses belonging to these cells are most likely to be silenced during training. This is because patterns with large L1 norms tend to drive the principal cell too strongly, causing synaptic weights to undergo decrements. On average, these decrements are greatest for synapses belonging to granule cells with high conformity (since the perceptron learning rule specifies that the size of any weight adjustment is proportional to the activity of the cell in the current pattern). Over many training steps, the weights of these synapses should tend toward 0. This outcome was seen in our simulations, where the likelihood that a given synapse would be silenced rose steeply with the conformity of the presynaptic cell (Fig. 4C).

This selectivity has important consequences for associative learning because the cells with high conformity contribute disproportionately to widening the distribution of L1 norms. This can be seen by isolating the input from the subpopulation consisting of cells for which synapses are silenced. L1 norms taken over this subpopulation are very broadly distributed (Fig. 4D) so the original distribution of L1 norms (Fig. 4D) becomes narrower when input from these cells is discarded. Thus silencing these synapses leads to partial normalization as total activity across the granule cells with surviving synapses is relatively uniform from one pattern to the next.

We also observed that synaptic silencing in the model selectively affects correlated pairs of cells (Fig. 4E, top). This is not surprising as cells that are correlated with the population also tend to be correlated with each other. To distinguish the
effect of the learning rule on cell-to-cell correlations from its effect on cell-to-population correlations, we created a set of patterns that were already normalized. The input-output function of the Ramp network was first used to generate granule cell patterns as before, and these patterns were then modified to have identical L1 norms (see METHODS). This preserved cell-to-cell correlations while making all cell-to-population correlations undefined (since normalizing the patterns caused the population responses to have 0 variance). Associations based on the modified patterns were much easier to store: in particular, training with 40 normalized patterns silenced the same number of synapses as did the previously described training with 15 unmodified patterns. We found that when patterns were normalized before training, synaptic silencing was not selective against correlated pairs of inputs (Fig. 4E, bottom). In other words, silencing leads to decorrelation only when the patterns presented have nonuniform L1 norms as in the original simulations. We therefore interpret this decorrelation as a byproduct of suppressing input channels that interfere with normalization.

**DISCUSSION**

In this study, we asked how a cerebellum-like circuit can be configured to solve a basic problem in associative learning: how to store multiple associations involving input patterns with different overall levels of spiking activity. We addressed this problem in a model with general features that are shared by many associative circuits. The behavior of the model suggests design principles for two stages of signal processing in these circuits.

The first of these stages is the integration of inputs by granule cells. We asked how response properties of these cells could help or hinder associative learning. From experimental measurements, it is known that the relationship between the firing rate of a granule cell and the firing rates of its presynaptic inputs can be described by a saturating nonlinearity with a nonzero threshold (Crowley et al. 2009; Gabbiani et al. 1994; Mitchell and Silver 2003; Rothman et al. 2009). The key parameters that specify a function of this kind are the gain (the slope of the rising segment) and the threshold (the location of this segment on the input axis). The behavior of our model
leads to the prediction that the granule cells projecting to a given principal cell should have different thresholds but similar gains. This is because the resulting granule cell activity patterns have low overlap (in the Euclidean sense) but high uniformity (with regard to L1 norms). Associations based on patterns with these properties are relatively easy to store in large numbers.

What could lead to variation in granule cell thresholds? The granule cell population is known to include many subtypes distinguished by markers of gene expression (Consalez and Hawkes 2012). Although members of a subtype tend to reside in compact subregions, the length and orientation of the parallel fibers ensures that each Purkinje cell encounters granule cell axons originating from several subregions. Further work may reveal whether different subtypes show different densities of ion channels associated with a persistent sodium current and a fast-activating potassium A-current, both of which are known to affect the spike threshold of the granule cell (D’Angelo et al. 1998). Variation may also arise from long-term, synthetically induced changes in input resistance and spike threshold, which have been observed in cerebellar granule cells (Armano et al. 2000). Finally, granule cells may experience different amounts of inhibitory feedback from Golgi cells. This feedback has a fast component that shifts the threshold of a granule cell without changing its gain (Crowley et al. 2009). Since each Golgi cell is driven by many mossy fibers and granule cells and inhibits many granule cells in turn (Eccles et al. 1964; Ito 2006; Kanichay and Silver 2008), it has been proposed that Golgi cells could act as “thermostats” to regulate the overall level of granule cell activity (Albus 1971; Brickley et al. 1996; Hamann et al. 2002; Marr 1969). Our results suggest a refinement of this picture in which unevenly distributed Golgi cell feedback leads to granule cells having similar gain but different thresholds.

The second stage of signal processing examined in this study is synaptic transmission from granule cells to principal cells. We observed that when input patterns have nonuniform L1 norms, the learning rule is highly selective in suppressing synapses that carry correlated activity. Our model thus predicts that for a given Purkinje cell, activity in the incident parallel fibers for which synapses are silent should be more correlated than in the incident parallel fibers as a whole. Classically, suppressing correlations is interpreted as a way to reduce redundancy, allowing information to be compressed into fewer channels. However, in many contexts, it is not obvious why compression is necessary or useful (Barlow 2001). In the case of the cerebellum, it seems curious that input patterns should undergo an enormous divergence into the granule cell layer only to be compressed at the next synapse into a fraction of the available channels. The behavior of our model suggests a rationale for this compression: in blocking correlated inputs, the learning rule preserves a subset of channels over which L1 norms are relatively uniform. This uniformity makes the learning problem tractable for the synapses that remain.

ACKNOWLEDGMENTS

We thank Richard Born, Markus Meister, and Rachel Wilson for their comments on an early version of this paper.

REFERENCES


ACKNOWLEDGMENTS

We thank Richard Born, Markus Meister, and Rachel Wilson for their comments on an early version of this paper.


