Supplementary information for "Long-term memory stabilized by noise-induced rehearsal" by Yi Wei and Alexei Koulakov

The validity of perturbation theory approximation

In our study we considered fluctuations induced by noise to be small. This means that noise was considered a small perturbation, i.e. within a perturbation theory. Here we discuss the validity of this approximation. Although we used this approximation (perturbation theory) to solve equations of our model, our mechanism may take place even when the equations cannot be solved using this method.

More precisely, smallness of the amplitude of noise was needed when we used Taylor expansion around the value u_i in equation (12). This equation does not include second order term, i.e. $F''(u_i)\delta u_i^2/2$. This approximation is valid when

$$|\delta u_i| \ll \left.\frac{F'(u)}{F''(u)}\right|_{u=F^{-1}(b)} \qquad \text{(S1)}$$

where b is given in equation (15). Because $\delta u_i \sim \xi_i$, this condition imposes a constrain on the noise amplitude ξ. To see this, solving equation (13), we get

$$\delta\mathbf{u}(t) = \frac{1}{\tau}\int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau}(1-g\mathbf{W})}\boldsymbol{\xi}(t') . \qquad \text{(S2)}$$

From this and equation (8) we find

$$\left\langle \delta u_i^2(t)\right\rangle = \frac{\xi^2}{\tau^2}\int_{-\infty}^{t} dt' \left[e^{-\frac{2(t-t')}{\tau}(1-g\mathbf{W})}\right]_{ii} . \qquad \text{(S3)}$$

As follows from this equation, this quantity averaged over neurons is

$$\overline{\left\langle \delta u^2(t)\right\rangle} = \frac{1}{N}\sum_i^N \left\langle \delta u_i^2(t)\right\rangle = \frac{1}{N}\frac{\xi^2}{\tau}\sum_i^N \frac{1}{2(1-gc_i)}. \qquad \text{(S4)}$$

Here c_i is the *i*-th eigenvalue of matrix **W.** Since there is only a small number of patterns, whose corresponding c_i are finite, and most of the eigenvalues c_i are close to zero, we have

$$\overline{\left\langle \delta u^2(t)\right\rangle} \approx \frac{\xi^2}{2\tau} . \qquad \text{(S5)}$$

Combining this with equation (22), we find the condition for perturbation calculation to be valid is

$$\frac{\xi}{\sqrt{\tau}} \ll \left.\frac{F'(u)}{F''(u)}\right|_{u=F^{-1}(b)} . \qquad \text{(S6)}$$

The amplitude of noise is therefore limited by $\xi^2 << \tau(F'(u)/F''(u))^2_{u=F^{-1}(b)} \sim \tau u^2 \sim \tau f^2/g^2$. Here, u is a typical value of membrane voltage and f is the typical value of the firing rates. Thus, the levels of noise have to be sufficiently low for the perturbation theory analysis to be valid.

For bistability, we need the value of noise larger than certain threshold. The detailed conditions for this criterion are described in section titled 'Conditions of bistability'. Thus, our analysis can be used when the level of noise is big enough for the bistability to exist, and small enough for the Taylor expansion in equation (12) to be valid. Can such a regime exist? Here, we will provide simple estimate for the existence of such a window of parameters. The perturbation theory is valid if noise is weak, i.e.

$$\xi^2 << \tau f^2/g^2. \qquad \text{(S7)}$$

As follows from the discussion in the paper, the bistability exists, if, approximately speaking, learning rate is sufficiently strong, i.e.

$$\gamma g^3 \xi^2 >> \frac{1}{A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2 . \qquad \text{(S8)}$$

Both conditions can be satisfied, if the amplitude of noise ξ^2 lies within the range defined as follows

$$\frac{1}{\gamma g^3 A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2 << \xi^2 << \frac{\tau f^2}{g^2} \qquad \text{(S9)}$$

This range exists if the boundaries for the rage differ in the correct direction, i.e.

$$\frac{1}{\gamma g^3 A_\pm}\left(\frac{\tau}{\tau_\pm}\right)^2 << \frac{\tau f^2}{g^2}. \qquad \text{(S10)}$$

which implies that

$$\frac{\tau}{\tau_\pm} << \gamma A_\pm f^2 g \qquad \text{(S11)}$$

Thus, if the learning rate $\gamma A_\pm$ is sufficiently big, both perturbation theory analysis (Taylor series expansion in equation (12)) is valid and bistability necessary for our mechanism exists. This occurred because the firing rate equations and, consequently, Taylor series expansion, do not depend on the learning rates. Thus, learning rates can be used as an independent parameter to reach the conditions of bistability. Also, the effects of noise can be big, even though we assume weak noise in the perturbation theory. This is because our assumption of the weakness of noise only includes the validity of Taylor series expansion. Thus, the overall impact of noise can be substantial despite its small amplitude.

The validity of the mean field approximation

In this section, we show that the mean-field calculation in section Materials and Methods is justified. In equations (9) and (11) we assumed that the learning rates are determined by the averages of the firing rates over the ensemble of noise. In reality, these equations should be used without such averaging. To derive our results, we therefore used an approximation that could be called the mean field approximation. At what condition can the instantaneous values of the pairwise products of the firing rates be replaced by their correlations? Below we will show that this condition is determined by the time scale of synaptic modifications. In particular, it is determined by the time constant of synaptic decay τ_0. It is this time scale that determines the duration of time over which the firing rates are averaged in equations (9) and (11). We will show that when the duration of STDP learning kernels $\tau_\pm$ [equation (10)] is much smaller than the forgetting time-scale, i.e. $\tau_\pm << \tau_0$, the mean field approximation can accurately describe the behavior of the network. Because, in reality, the STDP learning kernel lasts around 100ms, while the forgetting time scale extends over several weeks, $\tau_0 \sim 10^9$ ms, the variance of the deviations from the mean field values are small

$$\frac{\overline{(c-c_{MF})^2}}{c_{MF}^2} = \frac{\overline{\delta c^2}}{c_{MF}^2} \sim \frac{\tau_\pm}{\tau_0} \sim 10^{-7}. \qquad \text{(S12)}$$

This makes the mean field approximation a valid method.

We start from equation (9). Without averaging of noise, equation (9) has the following form

$$\tau_0 \frac{dW_{ij}}{dt} = -W_{ij} + \gamma \int_{-\infty}^{t} dt_1\, f_i(t_1) K(t_1 - t) f_j(t) + \gamma \int_{-\infty}^{t} dt_2\, f_i(t) K(t - t_2) f_j(t_2). \qquad \text{(S13)}$$

Let $c_a(t)$ be the strength of implicit pattern *a*. Using perturbation equation (12) and projecting both sides of (S13) with operator $\mathbf{P}^a$, we can find the equation for c_a. In the main text, where we used mean-field analysis, the equation for c_a is given in equation (19). The quantity described by that equation will be called $c_{MF}(t)$. Here we are interested in the difference between the mean field result and the result without averaging. To this end, we calculate the variance of the following quantity, $A(t) \equiv \tau_0 dc_a / dt + c_a$. To make notations simpler, we will omit the subscript *a* in the remaining part of this section and it is understood that our calculation is about a certain implicit patter *a* whose strength is *c*.

By projecting equation (S13) onto state a we obtain

$$\tau_0 dc_a / dt + c_a = A(t) \qquad \text{(S14)}$$

where $A(t)$ is given by

$$A(t) = \gamma g^2 \int_{-\infty}^{t} dt_1\, u(t_1) K(t_1 - t) u(t) + \gamma g^2 \int_{-\infty}^{t} dt_2\, u(t) K(t - t_2) u(t_2). \qquad \text{(S15)}$$

where

$$u(t) = N^{-1/2} \sum_i p_i^a \delta u_i(t) \qquad \text{(S16)}$$

is the projection of the membrane voltage defined in the main text onto state a. This quantity can be related to a Gaussian variable describing noise

$$u(t) = \frac{1}{\tau} \int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau}(1-gc)} \xi(t') \qquad \text{(S17)}$$

Here

$$\xi(t) = N^{-1/2} \sum_i p_i^a \xi_i(t) \qquad \text{(S18)}$$

Here, it is easy to see that $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \xi^2 \delta(t-t')$. It is also direct to show that $\langle u(t) \rangle = 0$ and

$$\langle u(t)u(t') \rangle = \frac{\xi^2}{2\tau} \frac{1}{1-gc} e^{-\frac{1-gc}{\tau}|t-t'|}. \qquad \text{(S19)}$$

From equation (19) we already know that

$$A_{MF} = \langle A(t) \rangle = \frac{1}{1-gc} \left(\frac{A'_+}{\frac{1-gc}{\tau} + \frac{1}{\tau_+}} + \frac{A'_-}{\frac{1-gc}{\tau} + \frac{1}{\tau_-}} \right) \qquad \text{(S20)}$$

To see how well we can approximate $A(t)$ by A, we need to calculate the variance of $A(t)$, i.e.

$$\langle (A(t) - A_{MF})(A(t') - A_{MF}) \rangle = \langle A(t)A(t') \rangle - A_{MF}^2 \qquad \text{(S21)}$$

In the calculation we use the fact, which follows from the properties of Gaussian white noise, that

$$\begin{aligned} &\langle u(t_1)u(t_2)u(t_3)u(t_4) \rangle \\ &= \langle u(t_1)u(t_2) \rangle \langle u(t_3)u(t_4) \rangle + \langle u(t_1)u(t_3) \rangle \langle u(t_2)u(t_4) \rangle \\ &\quad + \langle u(t_1)u(t_4) \rangle \langle u(t_2)u(t_3) \rangle. \end{aligned} \qquad \text{(S22)}$$

By straightforward calculations using equation (S17), we find that

$$\begin{aligned} &\frac{\langle (A(t) - A_{MF})(A(t') - A_{MF}) \rangle}{A_{MF}^2} \\ &= e^{-\frac{1-gc}{\tau}|t-t'|} \left(c_+ e^{-\frac{|t-t'|}{\tau_+}} + c_- e^{-\frac{|t-t'|}{\tau_-}} + c_0 e^{-\frac{1-gc}{\tau}|t-t'|} \right). \end{aligned} \qquad \text{(S23)}$$

Here c_+, c_- and c_0 are all functions of $A_\pm$, $\tau_\pm$, τ and g. To simplify the results, define the following three variables,

$$t_\pm = \frac{\tau_\pm}{\tau}(1-gc) \quad \text{and} \quad n = \left(\frac{A_+\tau_+}{1+t_+} + \frac{A_-\tau_-}{1+t_-} \right)^{-2}.$$

With $t_\pm$ and n, terms in (S18) can be written as

$$c_+ = n \left\{ t_+ \left(1 + \frac{2}{1+t_+} \right) \frac{A_+^2\tau_+^2}{t_+^2 - 1} + \frac{A_+\tau_+}{t_+^2 - 1} \frac{A_-\tau_-}{1+t_-} \left[2\left(1 + \frac{\tau_+}{\tau_+ + \tau_-} \right) t_+ + \frac{2\tau_-}{\tau_+ + \tau_-} t_+^2 \right] \right\}$$

$$c_- = c_+ \left(A_+ \leftrightarrow A_-, \tau_+ \leftrightarrow \tau_-, t_+ \leftrightarrow t_- \right)$$

$$c_0 = n \left\{ 2\frac{A_+^2\tau_+^2}{1-t_+^2} - 2\frac{A_+\tau_+}{t_+ - 1} \frac{A_-\tau_-}{1+t_-} \right\} + \left(A_+ \leftrightarrow A_-, \tau_+ \leftrightarrow \tau_-, t_+ \leftrightarrow t_- \right)$$

From our analysis in the main text we know that near the second stable point $t_\pm$ are both of order 1, i.e. $t_\pm \sim O(1)$. Therefore, c_+, c_- and c_0 are all of order 1.

From previous discussion, we can write $A(t) = A_{MF} + \delta A(t)$, such that $\langle \delta A(t) \rangle = 0$. To estimate $\delta A(t)$, notice the facts that τ is about a few milliseconds, $\tau_\pm$ are about a few hundred milliseconds and τ_0 is about a few weeks, i.e. $\tau \ll \tau_\pm \ll \tau_0$. By equation (S18), we have

$$\langle \delta A(t) \delta A(t') \rangle \sim \tau_\pm A_{MF}^2 \delta(t-t'). \qquad \text{(S24)}$$

We can now write $c(t) = c_{\mathrm{MF}}(t) + \delta c(t)$ such that $\tau_0 dc_{MF} / dt = -c_{MF} + A_{MF}$ and $\tau_0 d\delta c / dt = -\delta c(t) + \delta A(t)$. The first equation leads the mean-field solution that is presented in the main text.

The second equation describes the fluctuations around the mean field results. Solving the second equation, we get

$$\delta c(t) = \frac{1}{\tau_0}\int_{-\infty}^{t} dt' e^{-\frac{t-t'}{\tau_0}} \delta A(\mathrm{t}') \,. \qquad \text{(S25)}$$

From equation (S20), we find

$$\left\langle \delta c(t)\delta c(t') \right\rangle \sim \frac{\tau_{\pm} A_{MF}^2}{\tau_0} \,. \qquad \text{(S26)}$$

If we choose τ_0 to be two weeks, then $\tau_0 \sim 10^9 \mathrm{ms}$ and $\tau_{\pm}$ is a few hundred milliseconds, e.g. $\tau_{\pm} \sim 100\mathrm{ms}$, therefore by equation (S26) we have $\left\langle \delta c(t)\delta c(t') \right\rangle \to 0$. This proves that we can approximate $c(t)$ with $c_{MF}(t)$ and the mean-field calculations in the main text, e.g. equation (19), are indeed valid approximations.