Обучение с подкреплением=учителем

Николай Ильич Базенков, к.т.н.

Институт проблем управления им. В.А. Трапезникова РАН

Летняя школа РАИИ, 5-18 июля 2021 г.

Обучение

Без учителя
(unsupervised)
Выделение признаков
Кластеризация

Хеббовское обучение
STDP

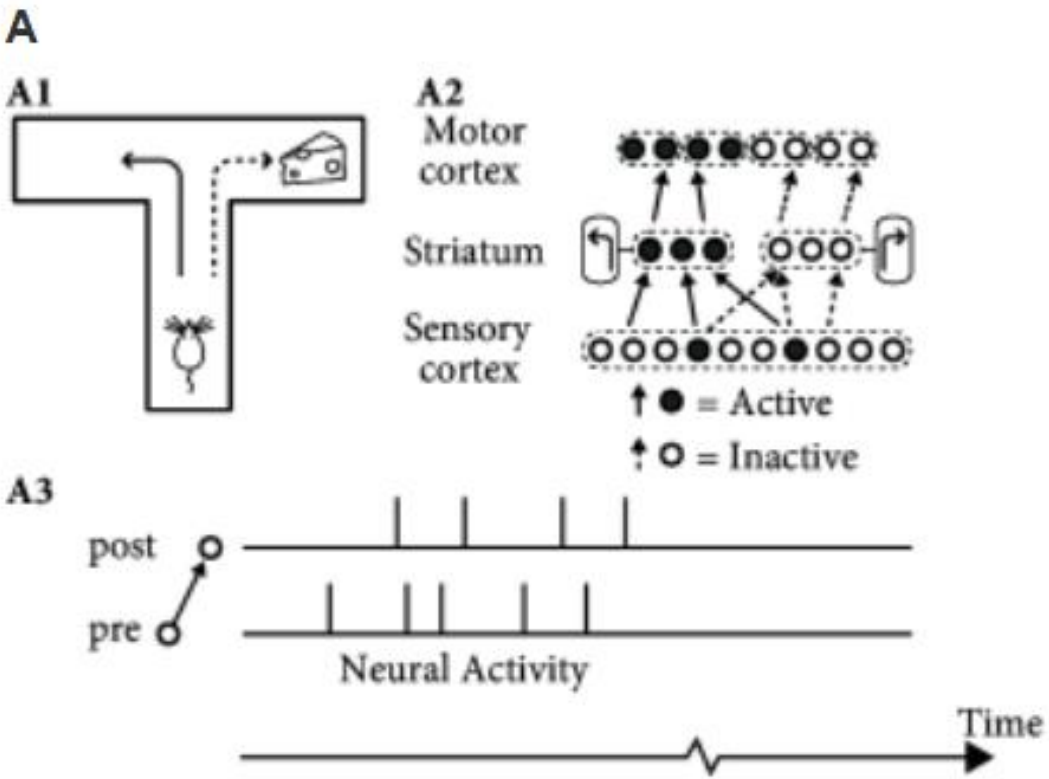
С учителем
(supervised)
Классификация
Регрессия

С подкреплением
(reinforcement learning)
Исследование мира
Оптимальное поведение

Обучение с подкреплением
(reward-based)

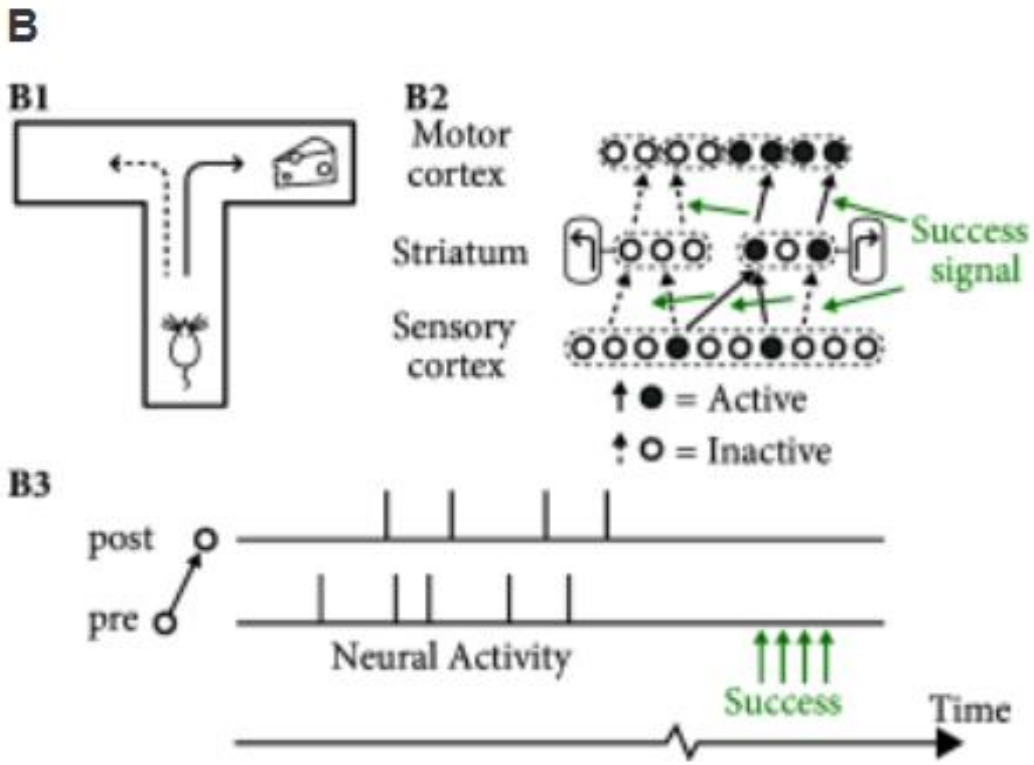
Вознаграждение и модуляция

Не вся активность одинаково полезна



Вознаграждение и модуляция

Выделять успешное поведение помогает нейромодуляция – воздействие нейротрансмиттера (допамина)

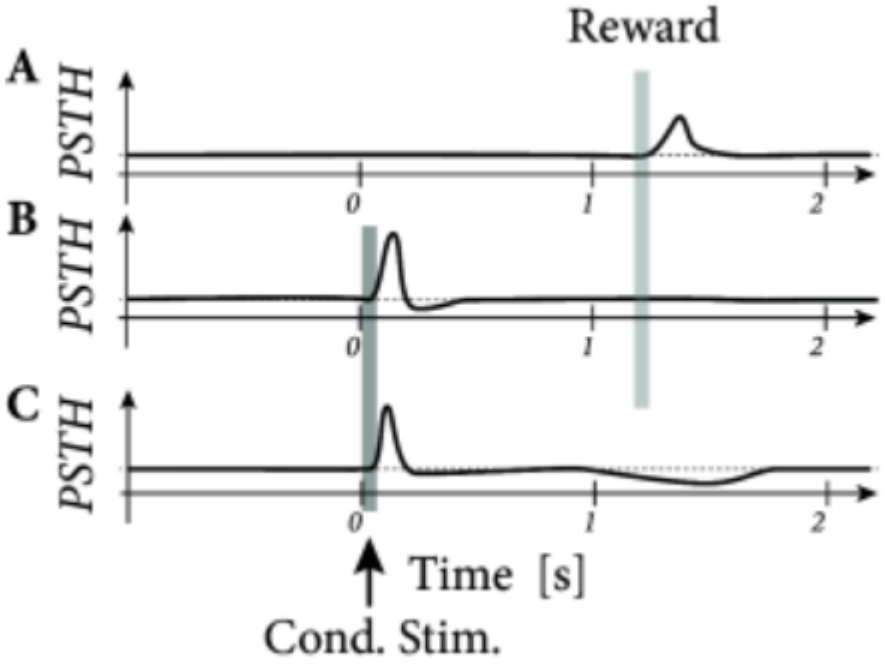


Вознаграждение и модуляция

Допаминовый нейрон (PSTH) :

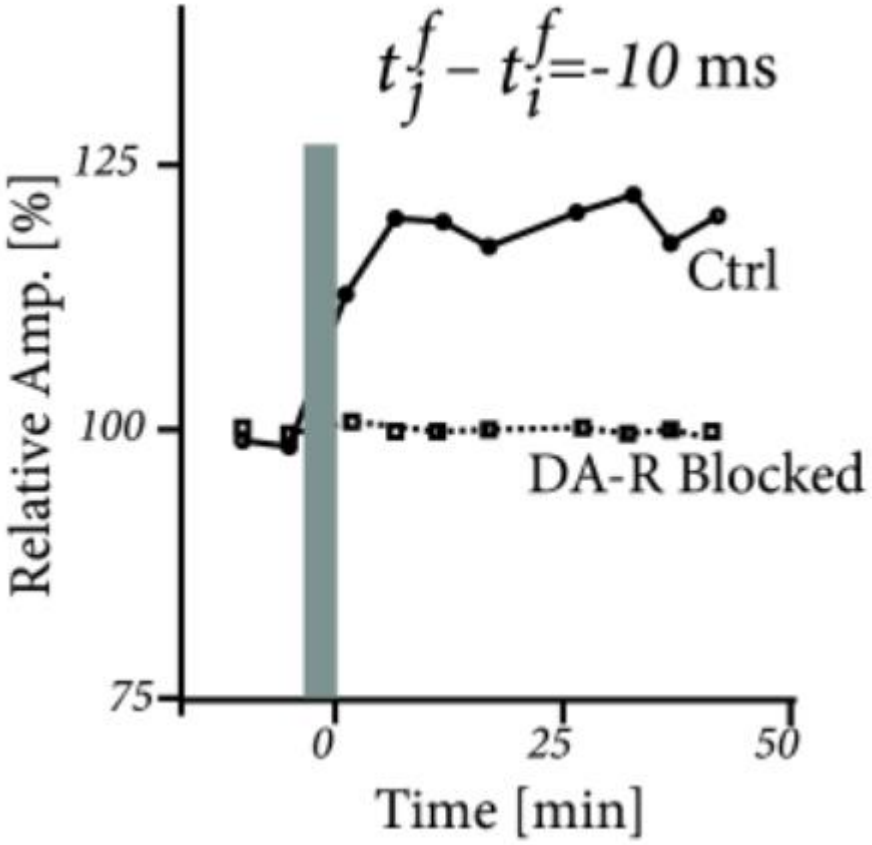
1. Активируется при получении вознаграждения
2. Активируется в ответ на условный стимул (свет, звук)
3. Затормаживается, если вознаграждение не поступает

Допамин = награда – ожидаемая награда



Роль допамина в STDP

При заблокированных допаминовых рецепторах нет обучения



Трехфакторное Хеббовское обучение

1. Синапс хранит историю активности (eligibility trace)

$$\tau_e \frac{\mathrm{d}}{\mathrm{d}\ t} e_{ij} = -e_{ij} + H\left(pre_j, post_i\right)$$

2. Вес меняется только в присутствии модулятора M

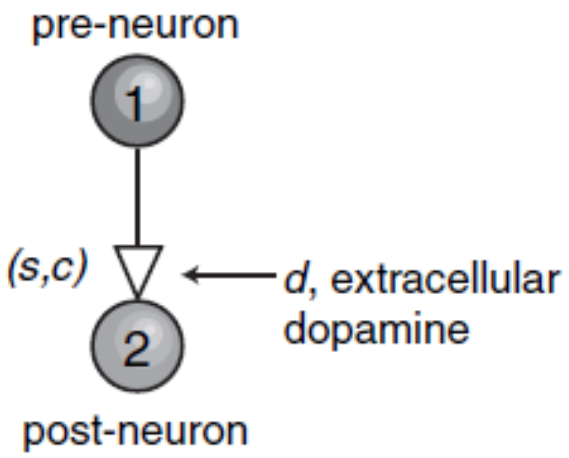
$$\frac{\mathrm{d}}{\mathrm{d}\ t} w_{ij} = M \cdot H\left(pre_j, post_i\right) e_{ij}$$

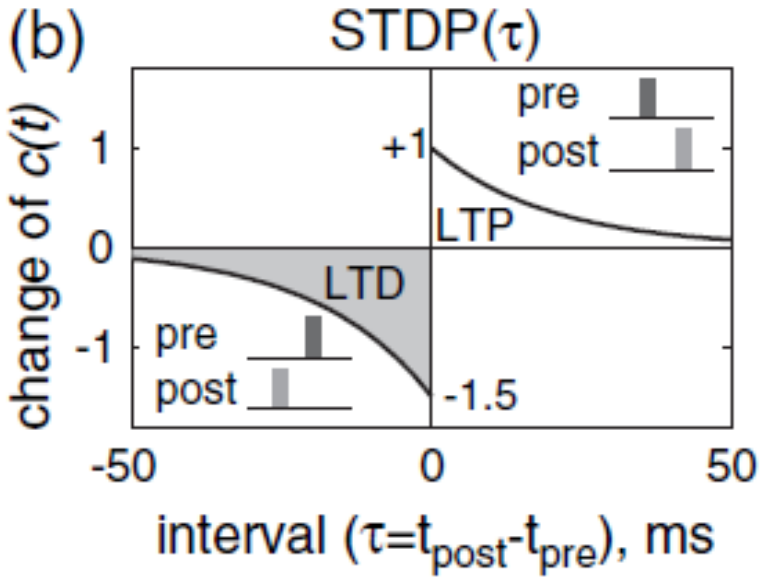
3. Модулятор выбрасывается, если награда превышает ожидаемую

$$M\left(t\right) = R\left(t\right) - \left\langle R \right\rangle$$

Допамин и STDP

Состояние синапса: сила (**s**) и недавняя история (eligibility trace) (**c**). STDP применяется к **c**, а не к **s**



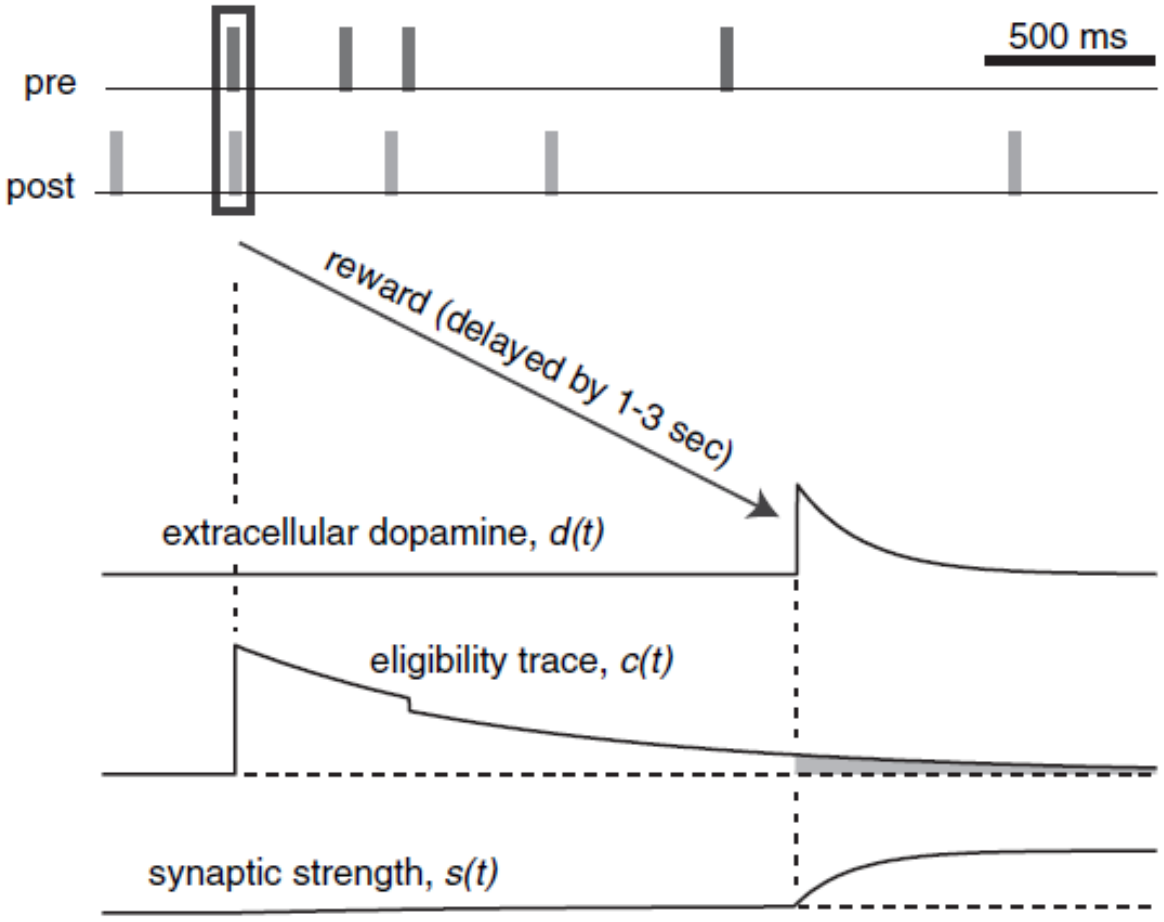


Eugene M. Izhikevich Solving the distal reward problem through linkage of STDP and dopamine signaling. Cerebral cortex 17, no. 10 (2007): 2443-2452.

https://brian2.readthedocs.io/en/stable/examples/frompapers.Izhikevich_2007.html

Допамин и STDP

Синапс усиливается только если получено вознаграждение $\mathbf{d_i(t)}$



Уравнения

Динамика синапса

$$\dot{c} = -c/\tau_c + \mathrm{STDP}(\tau)\delta(t - t_{\mathrm{pre/post}}), \quad \tau = t_{\mathrm{post}} - t_{\mathrm{pre}}$$

$$\dot{s} = cd.$$

$$\dot{d} = -d/\tau_d + \mathrm{DA}(t)$$

d(t) – количество допамина
STDP – функция усиления/ослабления на рисунке
DA(t) – поступление допамина от системы подкрепления

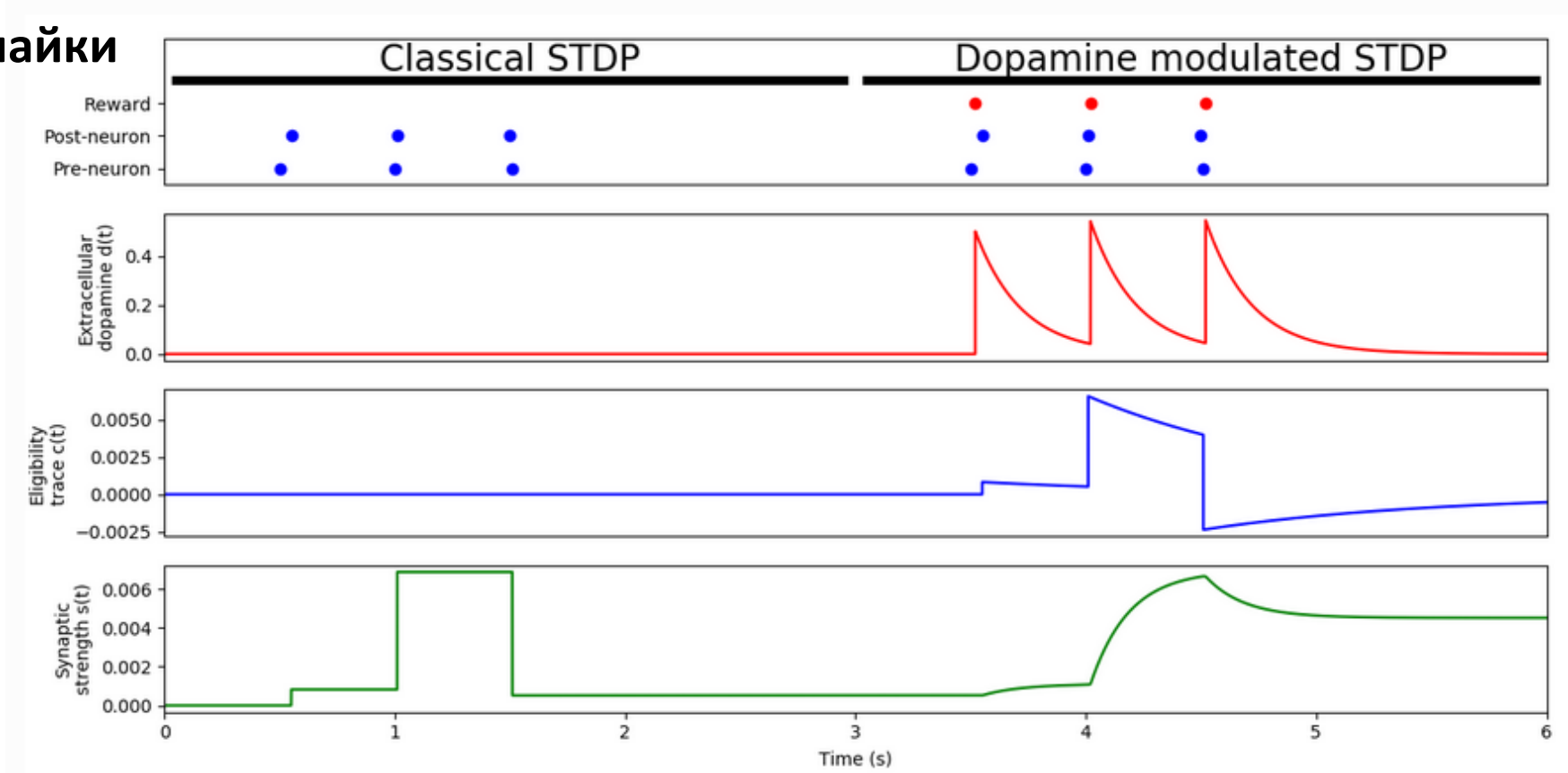
Сравнение с STDP

спайки

d

c

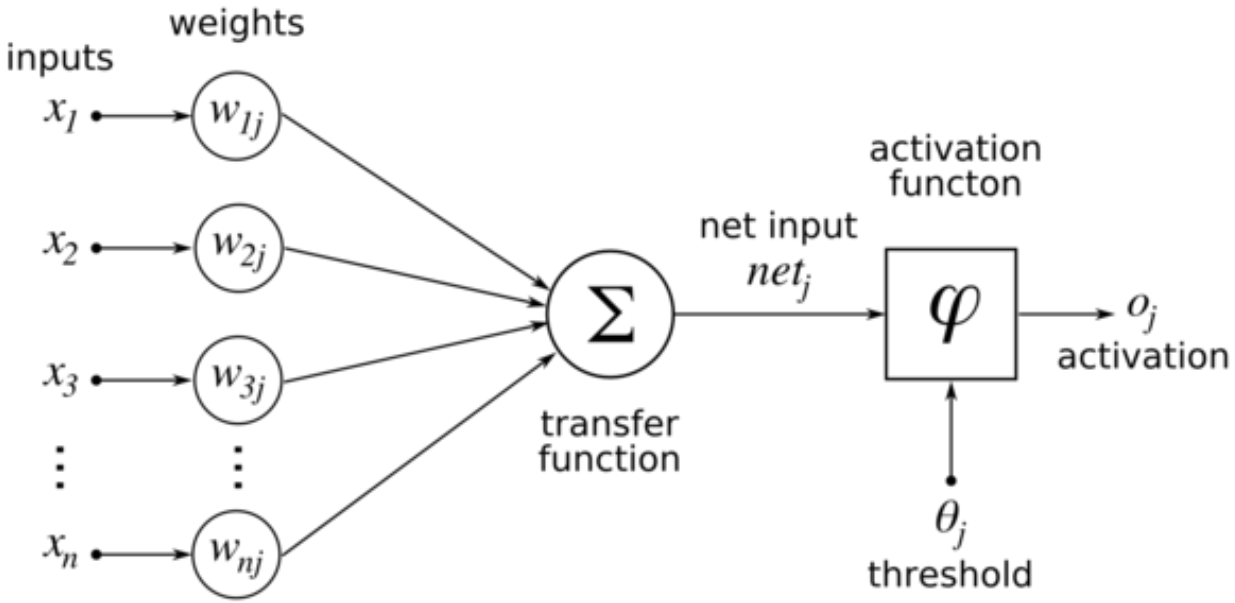
s



Обучение с учителем

1. Допамин – биологический аналог ошибки/вознаграждения в машинном обучении
2. Трехфакторное правило обучения – аналог обратного распространения ошибки
3. Проблема – дифференцируемость выхода нейрона

Обратное распространение ошибки



https://en.wikipedia.org/wiki/Backpropagation

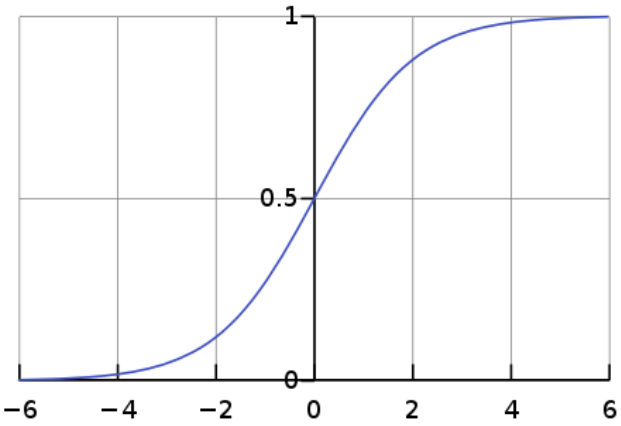
Обратное распространение ошибки

Выход нейрона

$$o_j = \varphi(\mathrm{net}_j) = \varphi\left(\sum_{k=1}^{n} w_{kj} o_k\right)$$

$$\varphi(z) = \frac{1}{1 + e^{-z}}$$

$$\frac{d\varphi(z)}{dz} = \varphi(z)(1 - \varphi(z))$$



Ошибка $E=L(o_j,o^*)$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial \mathrm{net}_j}\frac{\partial \mathrm{net}_j}{\partial w_{ij}}$$

Обучение

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

Спайкующий нейрон

Мембранный потенциал

$$\tau^{\text{mem}} \frac{dU_i}{dt} = (U^{\text{rest}} - U_i) + I_i^{\text{syn}}(t)$$

Синаптический ток

$$\frac{d}{dt} I_i^{\text{syn}}(t) = -\frac{I_i^{\text{syn}}(t)}{\tau^{\text{syn}}} + \sum_{j \in \text{pre}} w_{ij} S_j(t)$$

Спайки $S_j(t) = \sum_k \delta(t - t_j^k)$

Выход нейрона и ошибка

Ошибка между реальной серией спайков $S_i(t) = \sum_k \delta(t - t_i^k)$ и желаемой серией $\hat{S}_i$

$$L = \frac{1}{2} \int_{-\infty}^{t} ds \left[\left(\alpha * \hat{S}_i - \alpha * S_i \right)(s) \right]^2$$

α - оконная функция (экспонента, как в STDP)

Градиент ошибки

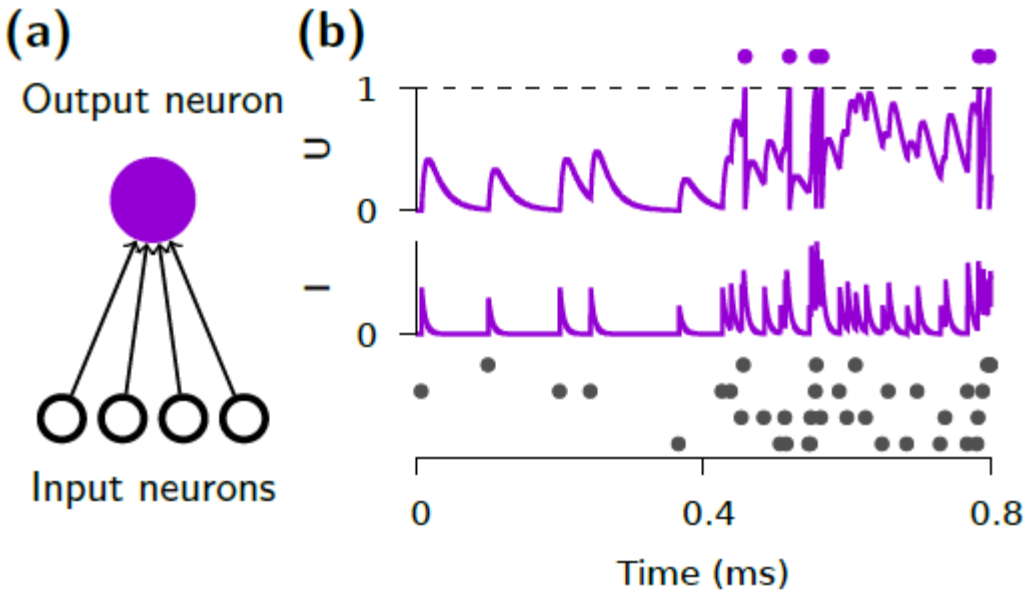
?

$$\frac{\partial L}{\partial w_{ij}} = - \int_{-\infty}^{t} ds \left[\left(\alpha * \hat{S}_i - \alpha * S_i \right)(s) \right] \left(\alpha * \frac{\partial S_i}{\partial w_{ij}} \right)(s)$$

Как найти градиент?

Будем считать выходом нейрона не серию спайков, а мембранный потенциал U_i

Но он тоже имеет разрывы!



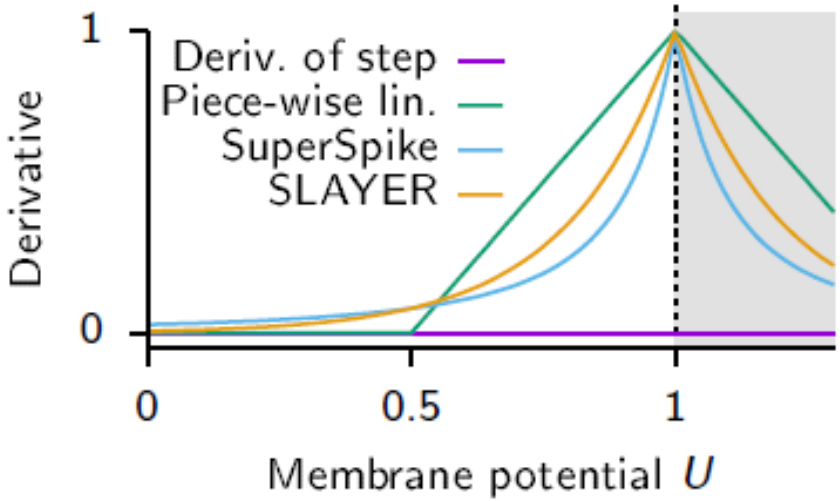
Как найти градиент?

Тогда будем использовать вместо спайков вспомогательную гладкую функцию активации

$$\sigma(U_i(t))$$

$$\frac{\partial S_i}{\partial w_{ij}} \rightarrow \sigma'(U_i)\frac{\partial U_i}{\partial w_{ij}}$$

?



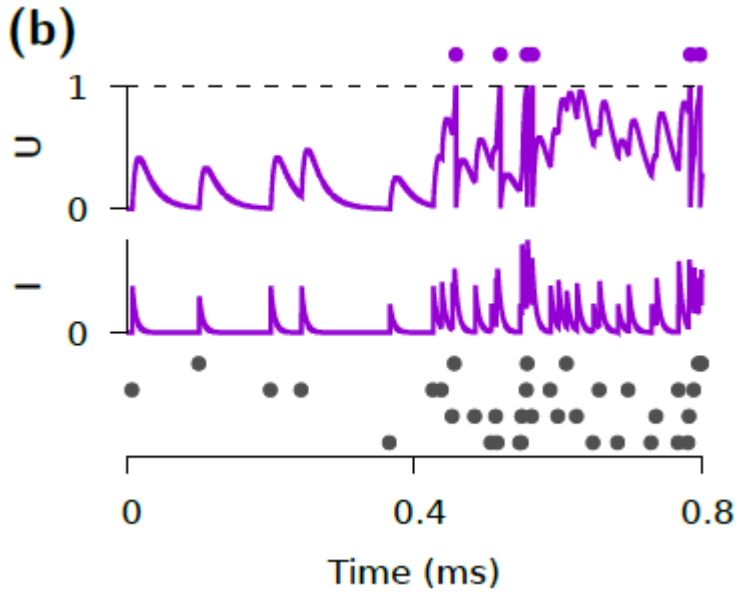
Как найти градиент?

Как теперь найти $\partial U_i / \partial w_{ij}$?

При условии, что спайки происходят достаточно редко, выполняется:

$$\frac{\partial U_i}{\partial w_{ij}} \approx (\epsilon * S_j(t))$$

ϵ - функция реакции мембраны (PSP), интегрирующая пресинаптические спайки и убывающая со временем



Пластичность

Super Spike learning rule:

$$\frac{\partial w_{ij}}{\partial t} = r \int_{-\infty}^{t} ds \quad \underbrace{e_i(s)}_{\text{Error signal}} \quad \underbrace{\alpha * \left(\underbrace{\sigma'(U_i(s))}_{\text{Post}} \underbrace{\left(\epsilon * S_j\right)(s)}_{\text{Pre}} \right)}_{\equiv \lambda_{ij}(s)}$$

Ошибка $e_i(s) \equiv \alpha * (\hat{S}_i - S_i)$

λ_{ij} - eligibility trace

ϵ - реакция мембраны (PSP) на серию спайков S_j

r – скорость обучения

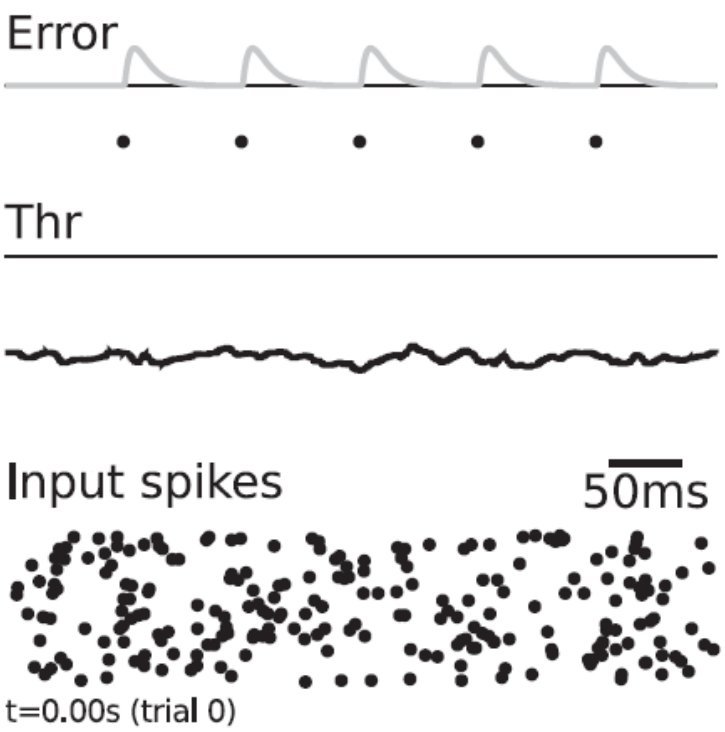
α – оконная функция

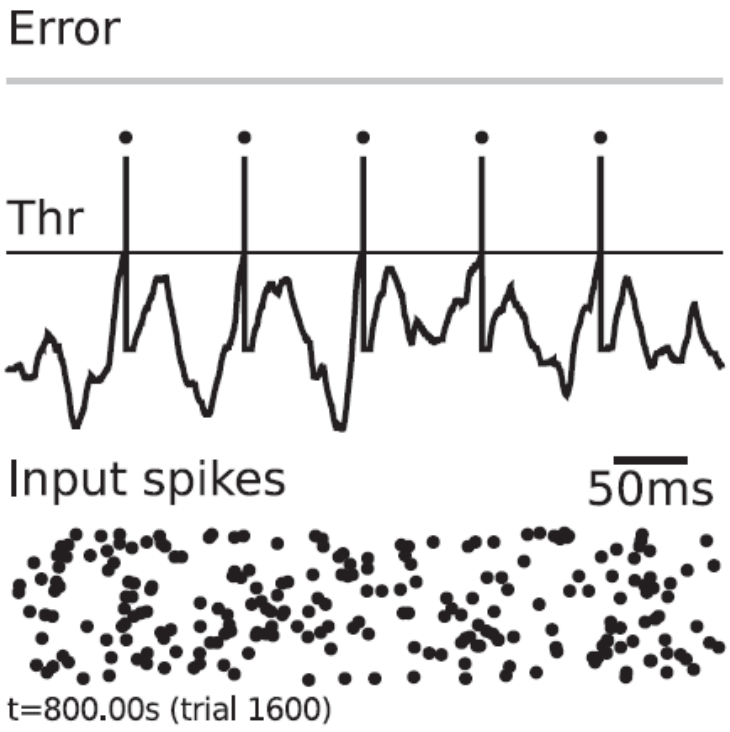
Свойства Super Spike правила

1. Пре- и пост-синаптическая активность умножается по Хеббу
2. Использует мембранный потенциал
3. Нелинейно из-за $\sigma'(U_i)$
4. Сохраняет eligibility trace для того, чтобы учесть отсроченное вознаграждение
5. Трехфакторное правило, где третий фактор (ошибка) специфичен для постсинаптического нейрона

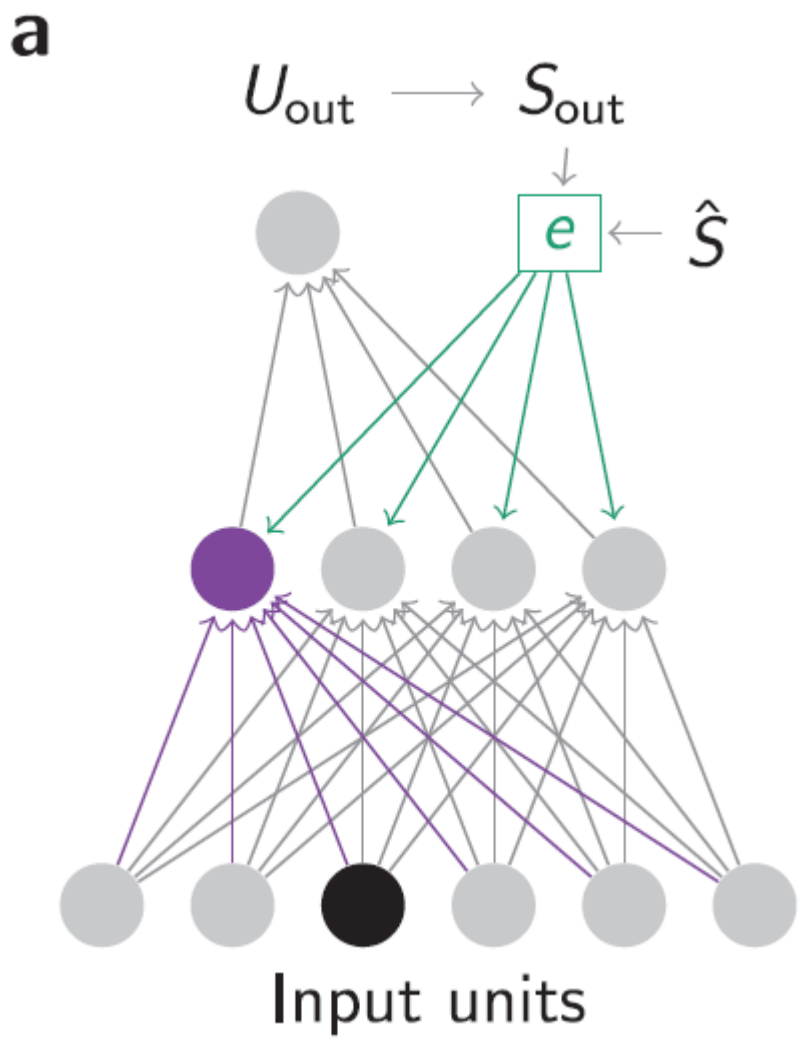
Обучение

Цель – обучить нейрон выдавать серию спайков

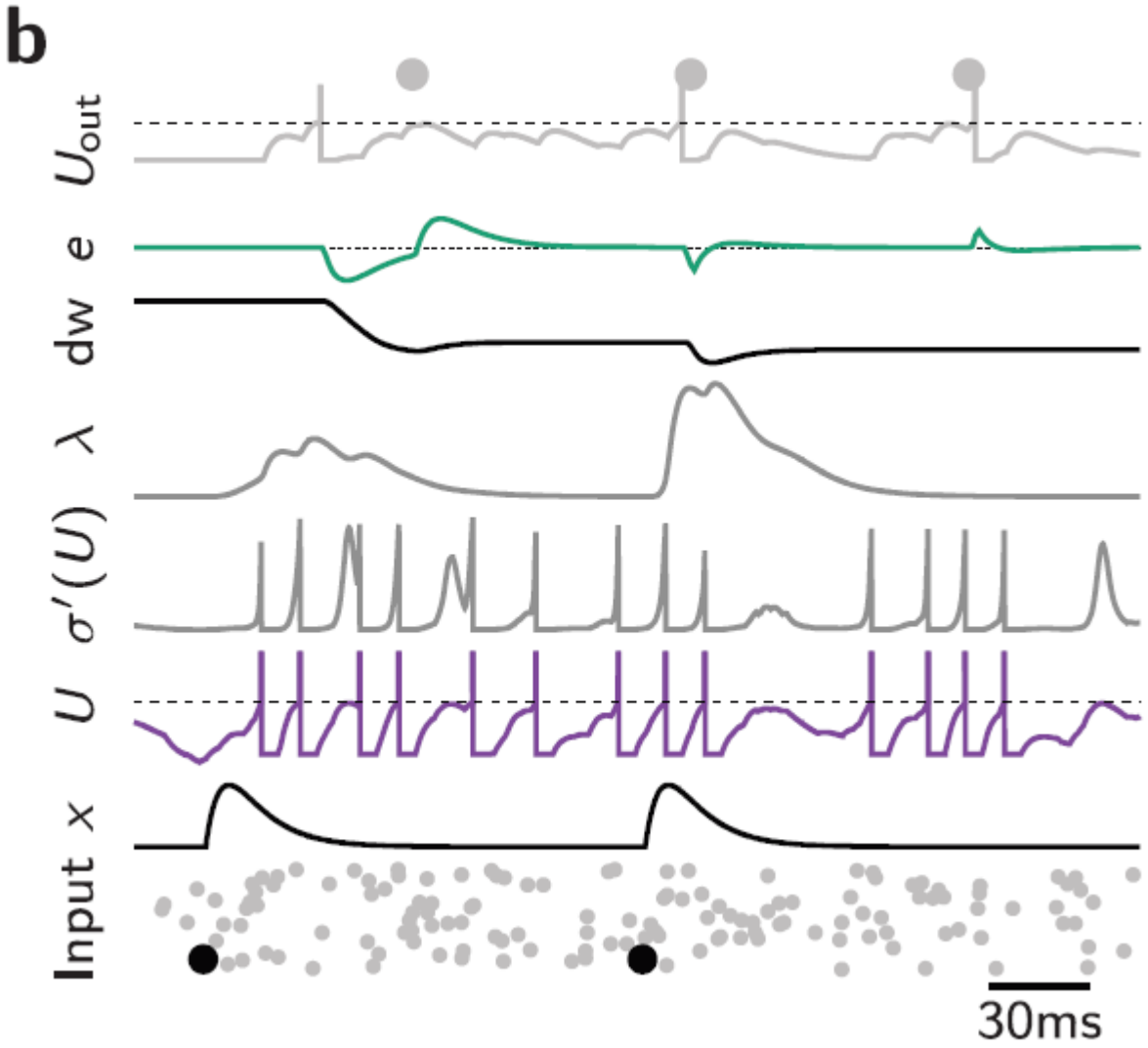




Обучение сети

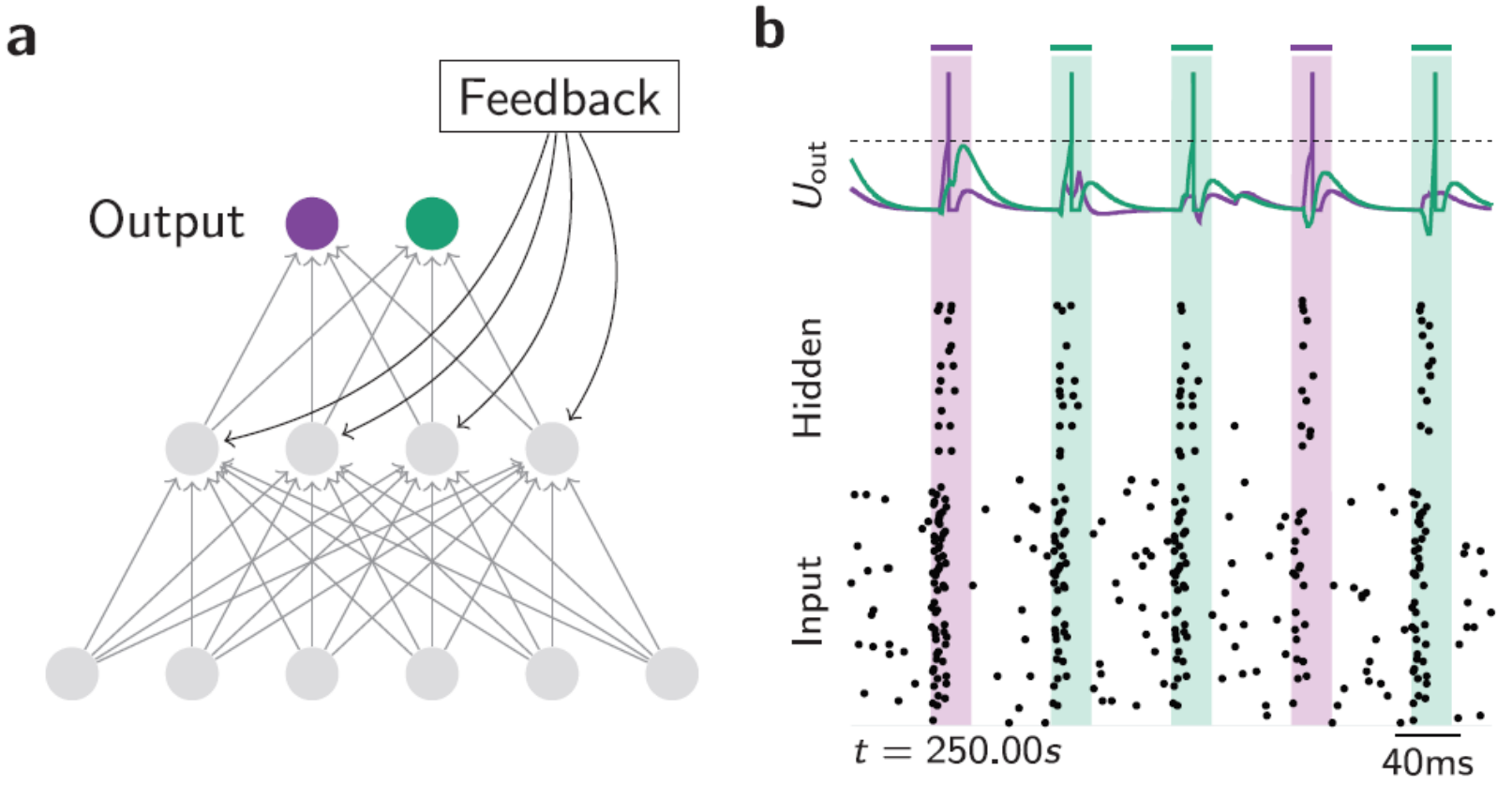


Обучение сети



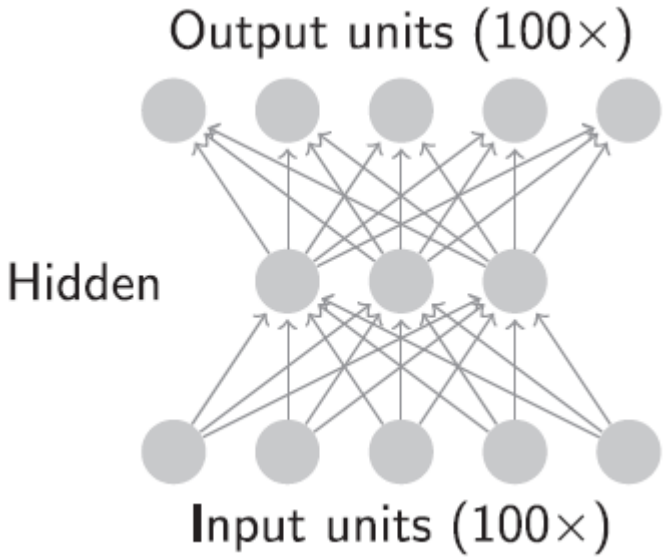
Классификация

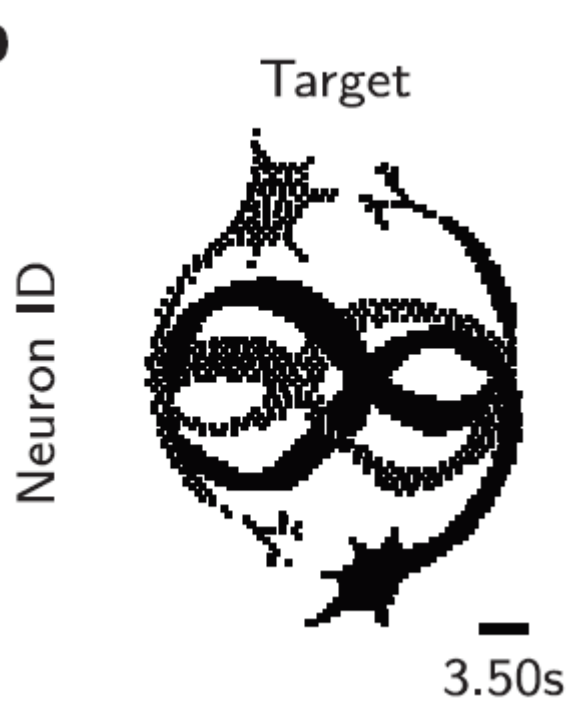
Цель – классифицировать два паттерна



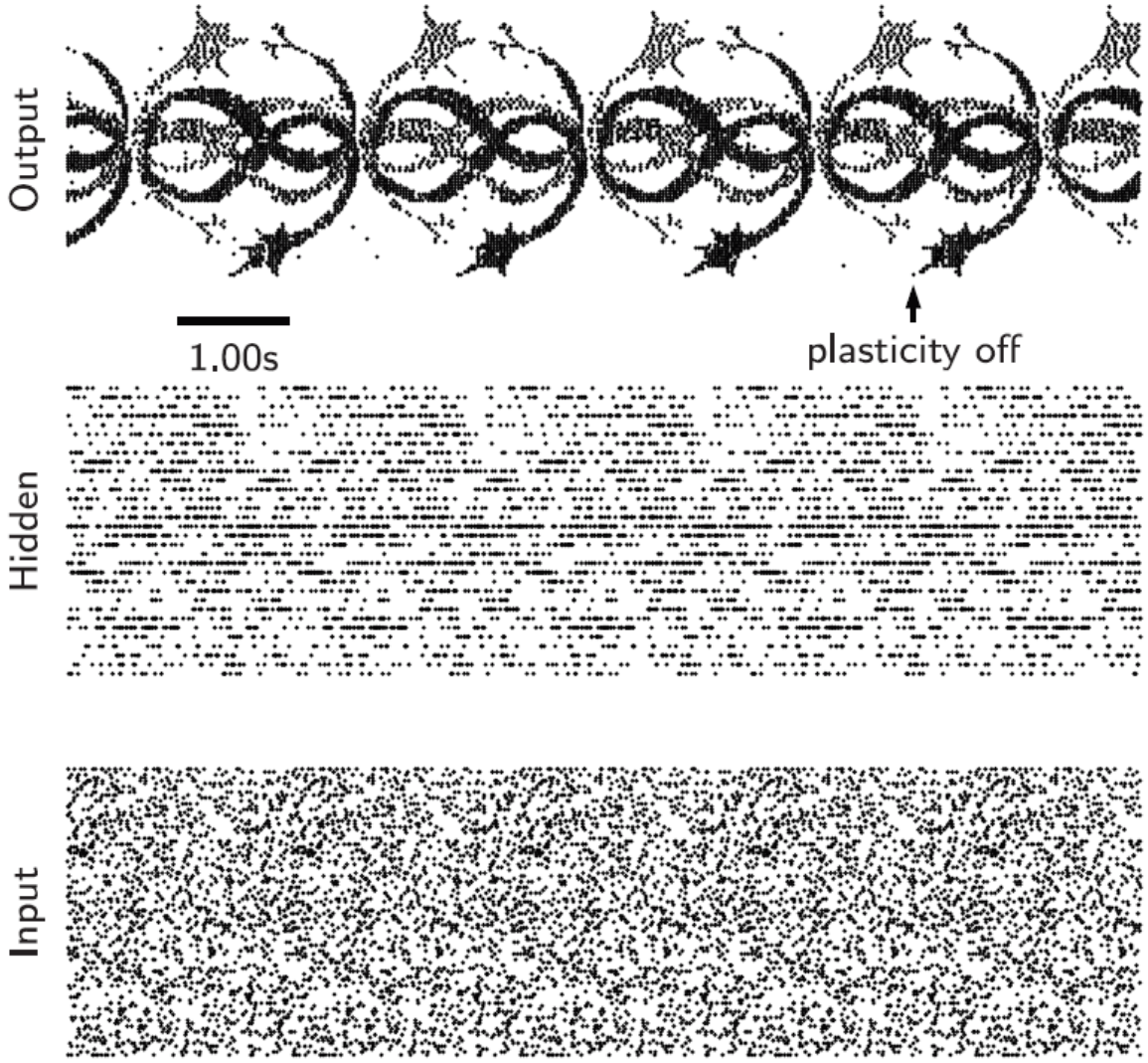
Классификация

Цель – воспроизвести сложный пространственно-временной паттерн спайков

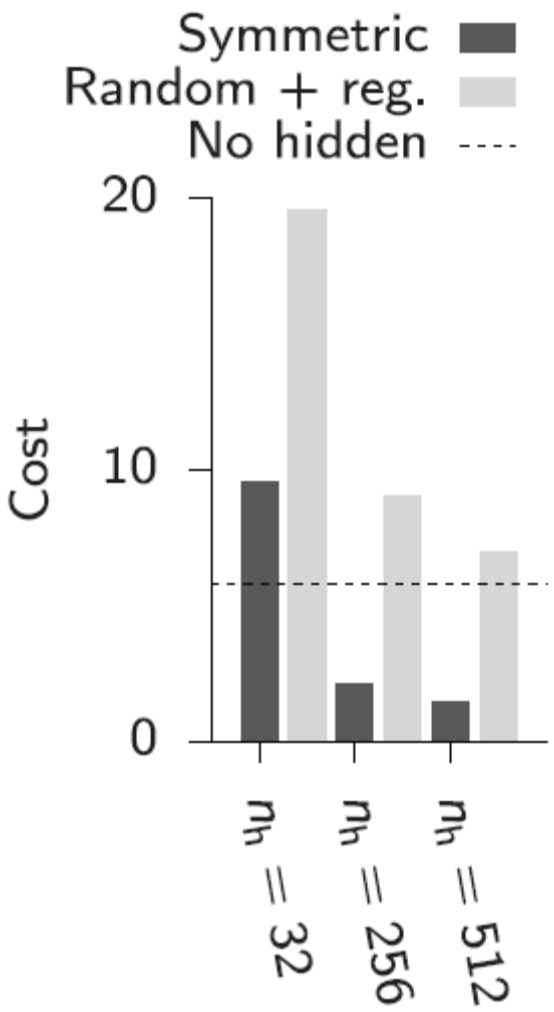


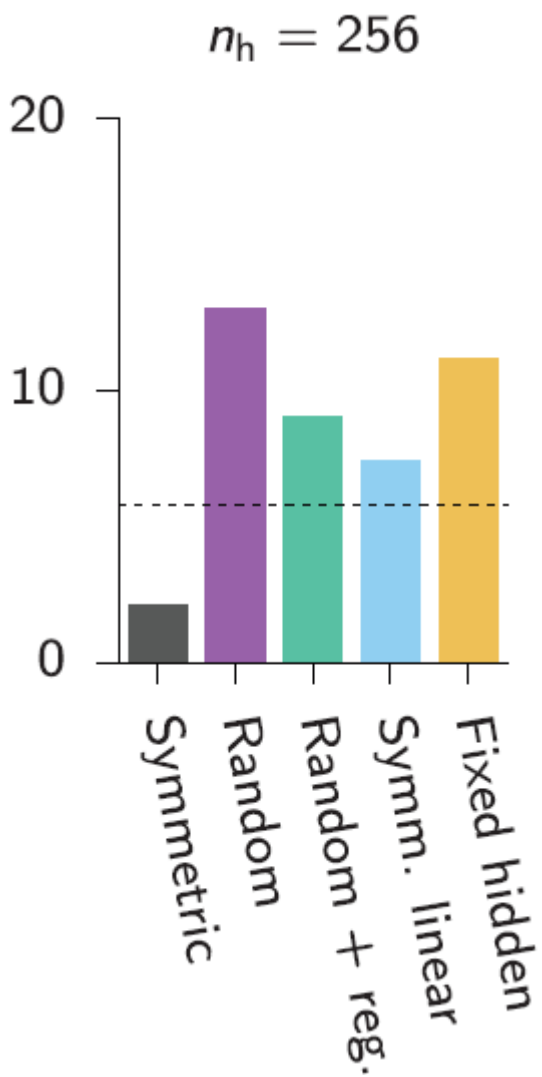


Сложная активность

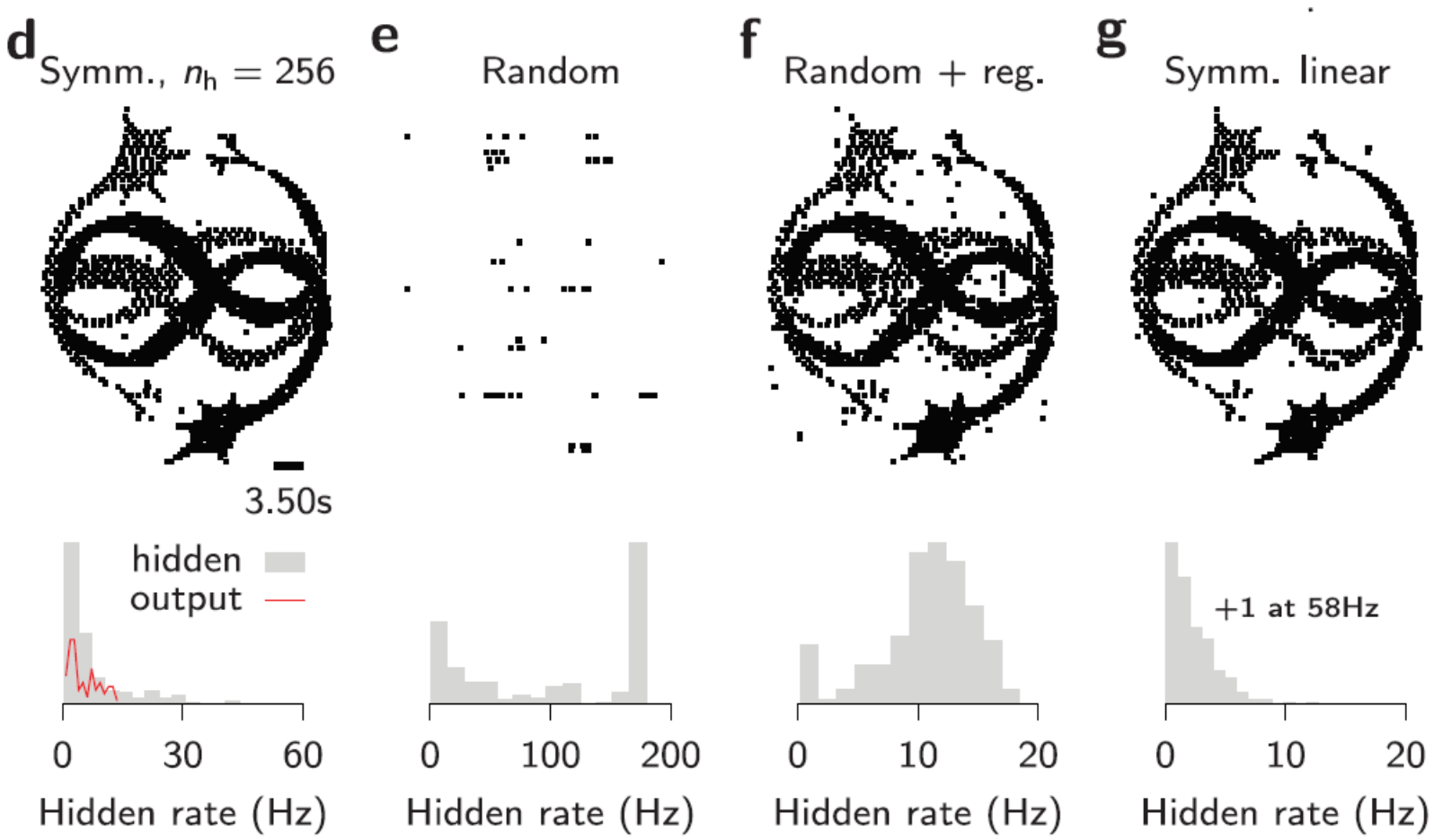


Обучение с разными обратными связями

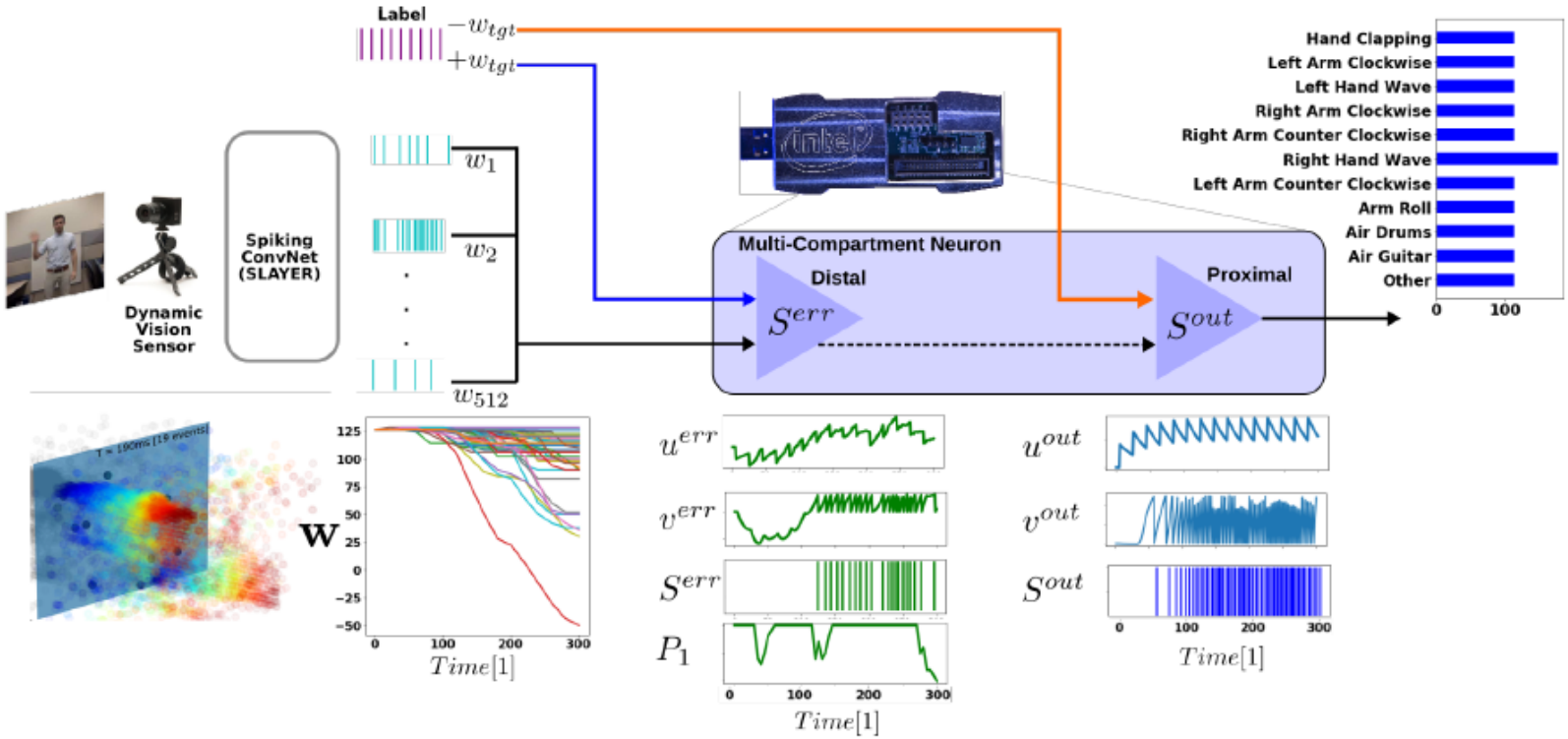




Обучение с разными обратными связями



Приложения – нейроморфные чипы



Stewart, K., Orchard, G., Shrestha, S. B., & Neftci, E. (2020, August). On-chip few-shot learning with surrogate gradient descent on a neuromorphic processor. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 223-227). IEEE.

Приложения – нейроморфные чипы

11-WAY FEW-SHOT CLASSIFICATION ON THE DVSGESTURE DATASET

Dataset	Learning Method	Shots	Train	Test
DVSGesture	Loihi Plasticity Rule	**1**	**100%**	**52.2%**
		5	**86.6%**	**56.8%**
		14	**72.2%**	**64.7%**
	SLAYER+Spiking	1	9.1%	20.1%
		5	40%	50.1%
		14	56.5%	61.8%
	SLAYER+Linear	1	<1%	41.8%
		5	38.2%	42.7%
		14	53.9%	51.8%

Stewart, K., Orchard, G., Shrestha, S. B., & Neftci, E. (2020, August). On-chip few-shot learning with surrogate gradient descent on a neuromorphic processor. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 223-227). IEEE.

Заключение

1. Трехфакторное правило обучения использует кроме совместной активности еще и наличие модулятора
2. Подкрепление допамином обеспечивает закрепление выигрышного поведения
3. Для спайковых нейронных сетей тоже существуют методы обратного распространения ошибки
4. Область приложений – нейроморфные чипы