برای تعریف کردن RSTD ابتدا به Reinforcement learning می پردازیم:

وقتی یک لرنینگ برای یک هدف خاصی طراحی نشده باشد اصطلاحاً unsupervised است. در یادگیری تقویتی سعی بر این است که از حالت unsupervised خارج شویم به طور کامل و بتوانیم agent مان را برای هدف خاصی آموزش دهیم.

Classical Conditioning: وقتی کاری که در نظر داریم را انجام داد جایزه‌ای به آن تعلق می‌گیرد و اگر نتوانست انجام دهد تنبیه می‌کنیم.

Operant Conditioning: فعالیت‌های نورونی یا شبکه‌های نورونی با دریافت جایزه یا تنبیه اعمال را یاد می‌گیرند. یادگیری تقویتی نه supervised است نهunsupervised

Unsupervised نیست به این دلیل که به واسطه عملی کهagent انجام می‌دهد یک فیدبک از محیط می‌گیرد.

Supervised نیست چون به agent نمی‌گوییم چی درست یا چی غلط است صرفاً سیگنال‌های جایزه یا تنبیه می فرستیم.

یک agent یادگیرنده داریم که در یک محیط تعامل انجام می‌دهد. این محیط حالت‌های مختلفی دارد. به عنوان agent ابتدا باید محیط را درک کنیم و بدانیم در چه state ای وجود داریم. امکان دارد fully observable یا partial observable باشد. Agent تصمیم بگیرد که عملیاتی را در قبال محیط انجام دهد این اکشن محیط را تحت تأثیر قرار میدهد. بر مبنای عملی که انجام داده محیط مجازات می‌کند یا جایزه می‌دهد. تمام حس‌های مثبت و منفی که ادراک می کنیم از اعمال و رفتار همان نتیجه فرایند یادگیری تقویتی است. مثلاً سیگنال reward در مغز ترشح می‌شود. هر حالت مثبت و منفی که احساس می‌کنیم به خاطر تغییر حالت شیمیایی در مغزمان است. در محیط Agent سعی می کند جایزه هایش را بیشینه کند.

فیدبک ها معمولاً با تاخیر میرسند ← سیستم credit assignment برای این‌که اگر یک پاداش رسید بدانیم برای کدام عمل بوده است.

Distal problem reward: جایزه خیلی فاصله‌ای با عمل ندارد.

- اگر زمان خیلی زیادی بین عمل و جایزه باشد ممکن است Action های دیگری هم انجام داده باشیم و ممکن است credit assignment سخت باشد که این جایزه یا تنبیه به خاطر کدام عمل بوده که در ازای آن تغییر ایجاد کنیم.
- باید تشخیص دهیم چه بخش‌هایی از شبکه عصبی درگیر این عمل بوده‌اند. که بتوانیم تضعیف یا تقویت انجام دهیم.
- کدام اسپایک‌ها بوده که در ازاش credit گرفتیم.

Dopamine: یکی از فرایندهای یادگیری تقویتی شناخته شده در مغز مبتنی بر یک neuro-modulator به اسم dopamine است. احساس رضایت در نتیجه دوپامین ایجاد می‌شود.

STDP دوفاز دارد :LTP :موقعی که سیناپس ها تقویت میشود؛ LTD: موقعی که سیناپس‌ها تضعیف شود.

وقتی دوپامین ترشح می‌شود غلظتش می تواند نحوه انجام STDP را تغییر دهد.

دوپامین می‌تواند یک سری از پترن‌های فایر شدن را reinforce کند و عموماً تأثیر دوپامین روی STDP با تأخیر یک تا دو ثانیه همراه است.

دوپامین همیشه در مغز ترشح می‌شود اگر مقدار آن از یک حدی پایین‌تر بیاید punishment داریم و اگر از یک مقدار بیشتر شود reward.

خاصیتی که دوپامین دارد که بخواهد الگوهای firing که با افزایش دوپامین همراه هستند را تقویت بکند و بقیه تضعیف شوند. بر مبنای یک فرایند سیناپتیکی به اسم synaptic eligibility traces یا synaptic tags است. این فرایند می‌گوید وقتی نورون‌ها فایر می‌شوند تا مدتی بعد از فایر اثر آن در نورون باقی می‌ماند موقعی که دوپامین بعد از تاخیر می‌آید سیناپس‌هایی که Eligibility trace شان بیشتر شده باشد تقویت می‌شوند. اگر دوپامین کم شود، ضعیف می‌شود. به نوعی می‌گوییم STDP را modulate می‌کند.

Reward modulated STDP (RSTDP): دوپامین در فرایند STDP روی دو چیز اثر می‌گذارد: ۱- changing STDP's window ۲- changing STDP's polarity

اگر بخواهیم STDP را به سمتی ببریم که تحت تأثیر دوپامین در مغز چه کاری باید انجام دهیم:

فرض کنیم وضعیت سیناپسها را با دو چیز نشان می دهیم:

-وزن سیناپس S

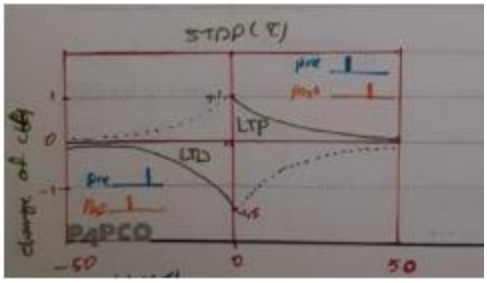
-یک آنزیم ثانویه c که قرار است نقش synaptic tags را اجرا کند.

روابط دینامیک این متغیرها:

$$\frac{dc}{dt} = -\frac{c}{\tau_c} + STDP(\tau)\delta(t - t_{pre/post})$$

$$\frac{ds}{dt} = cd$$

dدر واقع میزان دوپامینی است که به صورت خاجی از سلول به نورون رسیده. d اگر منفی باشد دوپامین کم شده و اگر مثبت باشد دوپامین افزایش پیدا کرده. اگر $d > 0$ باشد STDP حالت عادی اجرا می‌شود. اگر $d < 0$ باشد برعکس می‌شود.



اگر نورون اخیرا اسپایک زده باشد $-\frac{c}{\tau_c}$ به صورت exponentially به صفر میل می‌کند. هر زمان که یکی از نورون‌های post یا Pre فایر کرده باشد eligibility باید آپدیت شود. $STDP(\tau)\delta(t - t_{pre/post}$ توی لحظاتی که pre یا post فایر کرده باشد آپدیت می‌کنیم.

هر زمان که سیناپس جایزه یا تنبیه گرفت و c را در d ضرب می‌کنیم. اگر d مثبت باشد حالت عادی خواهد بود و اگر منفی باشد به anti STDP تبدیل می‌شود.

پس c میگوید بر مبنای STDP عادی وزن ما باید چقدر دچار LTD یا LTP شود.

یک عامل دوپامین هم هست که میتواند STDP polarity را نگه دارد یا برعکس کند که همان d است.

تفاوت RSTDP و STDP در آن است که در RSTDP می‌توانیم عملکرد نورون‌ها را با سیستم جایزه و تنبیه کنترل کنیم.