Statistical physics of RL algorithms in high dimensions

Andrew Saxe

December 2021

1 Problem formulation

Suppose we have an environment defined by a teacher perceptron $y = \text{sgn}(\bar{w}x)$ with input dimension N. At each time step $t = 1, \cdots, T$ within an episode, we receive a high-dimensional observation $x_t \sim \mathcal{N}(0, 1/N)$. The agent takes an action based on the output of a student perceptron,

How does this depend on current environment/previous action, or will we extend to that case after solving this one?

$$\hat{y} = \text{sgn}(wx). \tag{1}$$

We could consider many reward structures, but here we will assume a sparse one: we receive a reward of 1 at the final time step only if we make the correct decision on every time point, and zero otherwise.

The parameters of the student perceptrons are updated according to the REINFORCE policy gradient algorithm, which performs gradient descent on the average reward $J(w) = \langle \sum_t r_t \rangle$ where the average is over a batch of size b. The gradient is

$$\nabla_w J = \left\langle \sum_{t=0}^{T-1} \nabla_w \log \pi(a_t|x_t) \left(\sum_{t'=t+1}^{T} r_{t'} \right) \right\rangle \tag{2}$$

Because of our reward definition, the reward sum is either 0 or 1 for all times in a trial, depending on whether all classifications are correct. Thus we have

$$\nabla_w J = \left\langle \sum_{t=0}^{T-1} \nabla_w \log \pi(a_t|x_t) | y_t = \hat{y}_t \forall t \right\rangle p(correct). \tag{3}$$

Here the $\nabla_w \log \pi(a_t|x_t)$ term is approximately a Hebbian update between the output of the student (or teacher) and the input, given that the student was rewarded (i.e., correct on all T inputs),

$$\begin{aligned}
\nabla_w J &= \left\langle \sum_{t=0}^{T-1} \hat{y}_t x_t | y_t = \hat{y}_t \forall t \right\rangle p(correct) && (4)\\
&= T \langle yx | correct \rangle p(correct) && (5)\\
&= \frac{\sqrt{2}T}{\sqrt{\pi N}} \left(1 - \frac{\theta}{\pi}\right)^T w^H && (6)
\end{aligned}$$

where θ is the angle between w and $\bar{w}$, and w^H is a unit vector pointed half way between w and $\bar{w}$, that is,

Do we not contrain w to have normalization N in the thermodunamic limit?

$$w^H = \frac{\frac{w}{||w||} + \frac{\bar{w}}{||\bar{w}||}}{\left\|\frac{w}{||w||} + \frac{\bar{w}}{||\bar{w}||}\right\|}. \tag{7}$$

Now suppose that $||w(t)|| = b(t)$. Then we have the update

$$\begin{aligned} w(t+1) &= w(t) + \lambda \nabla_w J && (8)\\ \bar{w}^T w(t+1) &= b(t)\cos(\theta(t)) + \lambda \frac{\sqrt{2}T}{\sqrt{\pi N}} \left(1 - \frac{\theta(t)}{\pi}\right)^T \bar{w}^T w^H(t). && (9) \end{aligned}$$

The term

$$\begin{aligned} \bar{w}^T w^H(t) &= \bar{w}^T \frac{\frac{w}{||w||} + \frac{\bar{w}}{||\bar{w}||}}{\left\|\frac{w}{||w||} + \frac{\bar{w}}{||\bar{w}||}\right\|} && (10)\\ &= \frac{\cos(\theta(t)) + 1}{\left\|\frac{w}{||w||} + \frac{\bar{w}}{||\bar{w}||}\right\|} && (11)\\ &= \frac{\cos(\theta(t)) + 1}{\sqrt{2 + 2\cos(\theta(t))}} && (12)\\ &= \frac{1}{\sqrt{2}} \sqrt{1 + \cos(\theta(t))} && (13) \end{aligned}$$

Hence

$$b(t+1)\cos(\theta(t+1)) = b(t)\cos(\theta(t)) + \lambda \frac{T}{\sqrt{\pi N}} \left(1 - \frac{\theta(t)}{\pi}\right)^T \sqrt{1 + \cos(\theta(t))}.$$

After the gradient update, the norm is

$$\begin{aligned} ||w(t+1)||^2 &= ||w(t)||^2 + 2\lambda \nabla_w J^T w(t) + \lambda^2 ||\nabla_w J||^2 && (14)\\ b(t+1)^2 &= b(t)^2 + 2\lambda \frac{T}{\sqrt{\pi N}} \left(1 - \frac{\theta(t)}{\pi}\right)^T \sqrt{1 + \cos(\theta(t))} && (15)\\ &\quad + \lambda^2 \left[\frac{T}{\sqrt{\pi N}} \left(1 - \frac{\theta(t)}{\pi}\right)^T\right]^2. && (16) \end{aligned}$$

Now let $q(t) = \cos(\theta(t))$. Then we have the updates

$$\begin{aligned} b(t+1) &= \left(b(t)^2 + 2\lambda \frac{T}{\sqrt{\pi N}} \left(1 - \frac{\operatorname{acos}(q(t))}{\pi}\right)^T \sqrt{1 + q(t)} \right.\\ &\quad \left. + \lambda^2 \left[\frac{T}{\sqrt{\pi N}} \left(1 - \frac{\operatorname{acos}(q(t))}{\pi}\right)^T\right]^2\right)^{1/2}\\ q(t+1) &= \left(b(t)q(t) + \lambda \frac{T}{\sqrt{\pi N}} \left(1 - \frac{\operatorname{acos}(q(t))}{\pi}\right)^T \sqrt{1 + q(t)}\right) / b(t+1) \end{aligned}$$

Now we pass to the high dimensional limit in which the number of examples P and input dimension N go to infinity, but their ratio $\alpha = P/N$ is finite. Hence our discrete time step $t = \alpha N$.

(We note that this Hebbian assumption could be replaced with a more complex model, like logistic regression.)