Stanford CS229 ps4 sol

Shuifan Wang

August 2025

1 Neural Networks: MNIST image classification

(a) Let s denote the softmax score vector, then

$$\frac{\partial L(s,y)}{\partial x_j} = \sum_i (-\frac{y_i}{s_i}) s_i(\delta_{ij} - s_j) = -y_j + (\sum_i y_i) s_j = s_j - y_j.$$

Generally,

$$\frac{\partial L(s,y)}{\partial \mathbf{x}} = (\text{diag}(\mathbf{s}) - \mathbf{s}\mathbf{s}^T)\mathbf{g},$$

where $\mathbf{g}$ denotes the gradient w.r.t $\mathbf{s}$.
(b) Below is the training curve.

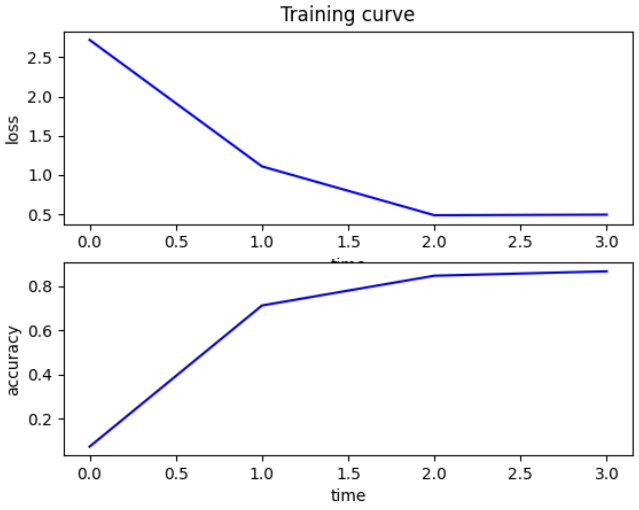

Figure 1: Problem 1

2 Off Policy Evaluation And Causal Inference

(a)

$$\begin{aligned}
\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) =& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) \\
=& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{p(a|s,\pi_1)}{p(a|s)} R(s,a) \\
=& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} \frac{p(a|s)}{p(a|s)} R(s,a) \\
=& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a),
\end{aligned}$$

since $\hat{\pi}_0 = \pi_0$.
(b) It is obvious since we can derive using the result of (a) and notice that

$$\begin{aligned}
\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} =& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\pi_0(s,a)} \\
=& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{p(a|s,\pi_1)}{p(a|s)} \\
=& \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} \frac{p(a|s)}{p(a|s)} \\
=& 1.
\end{aligned}$$

(c) If we only have a single data element in our observational dataset, then from weighted importance sampling estimate, we have

$$\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) / \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} = R_1(s,a).$$

However, from importance estimate, we have

$$\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) = \frac{\pi_1(s,a)}{\pi_0(s,a)} R_1(s,a) \neq R_1(s,a),$$

for the reason that $\pi_1 \neq \pi_0$ most of the time.
(d)

i. If $\pi_0 = \hat{\pi}_0$, then

$$\begin{aligned}
&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}}\left((\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a)) + \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}(R(s,a) - \hat{R}(s,a))\right)\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a) + \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}}\left((\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a)) - \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}\hat{R}(s,a)\right)\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a) - \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} \hat{R}(s,a) + \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}}(\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a))\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a) - \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} \hat{R}(s,a) + \mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} \hat{R}(s,a)\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a).
\end{aligned}$$

The second last equation is due to the fact that we simply sample from the (inner) distribution $\pi_1(s,a)$, but not π_0. This will be more explicit if we use another notation like a'.

ii. If $\hat{R}(s,a) = R(s,a)$, then

$$\begin{aligned}
&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}}\left((\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a)) + \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}(R(s,a) - \hat{R}(s,a))\right)\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_0(s,a)}}(\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a))\\
=&\mathbb{E}_{\substack{s\sim p(s)\\ a\sim\pi_1(s,a)}} R(s,a).
\end{aligned}$$

(e)

i. Since drugs are randomly assigned, the logging policy π_0 has good coverage over all possible actions. This makes importance sampling reliable, even if the relationship between drug, patient, and lifespan is very complex. Regression would struggle here because it would require fitting a complicated model. Therefore, the importance sampling estimator works better.

ii. In this case, drug assignment follows a very complicated policy π_0, but the mapping from (s,a) to $R(s,a)$ is simple. That means regression can accurately learn the reward function with limited bias, while importance sampling would suffer from large variance due to poor coverage of π_0. Therefore, the regression estimator works better.

3 PCA

$$\begin{aligned}
f_u(x) &= \arg\min_{v\in\mathcal{V}} ||x-v||_2^2 \\
&= \arg\min_{\alpha\in\mathbb{R}} ||x-\alpha u||_2^2 \cdot u \\
&= \arg\min_{\alpha\in\mathbb{R}} (x-\alpha u)^T(x-\alpha u)\cdot u \\
&= \arg\min_{\alpha\in\mathbb{R}} (x^Tx - 2\alpha u^Tx + \alpha^2 u^Tu)\cdot u \\
&= \frac{u^Tx}{u^Tu}u.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\arg\min_{u:u^Tu=1}\sum_{i=1}^m ||x^{(i)} - f_u(x^{(i)})||_2^2 &= \arg\min_{u:u^Tu=1}\sum_{i=1}^m ||x^{(i)} - \frac{u^Tx^{(i)}}{u^Tu}u||_2^2 \\
&= \arg\min_{u:u^Tu=1}\sum_{i=1}^m (x^{(i)} - (u^Tx^{(i)})u)^T(x^{(i)} - (u^Tx^{(i)})u) \\
&= \arg\min_{u:u^Tu=1}\sum_{i=1}^m ((x^{(i)})^Tx^{(i)} - 2(x^{(i)})^T(u^Tx^{(i)})u + (u^Tx^{(i)})^2) \\
&= \arg\max_{u:u^Tu=1}\sum_{i=1}^m (u^Tx^{(i)})^2 \\
&= \arg\max_{u:u^Tu=1} u^T(\sum_{i=1}^m x^{(i)}(x^{(i)})^T)u,
\end{aligned}$$

which is what we want to maximize in the lecture note of PCA.

4 Independent components analysis

(a)

$$\begin{aligned}
\nabla_W\ell(W) =& \nabla_W\sum_{i=1}^n(\log|W| + \sum_{j=1}^d \log\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(w_j^Tx^{(i)})^2)) \\
=& \nabla_W(n\log|W| - \frac{1}{2}WX^TXW^T) \\
=& nW^{-T} - X^TXW = 0.
\end{aligned}$$

We solve that $W^TW = n(X^TX)^{-1}$. For any orthogonal matrix Q, we have that $(QW)^T(QW) = W^TW = n(X^TX)^{-1}$, so the solution W is not unique.

(b) In this case, for a single example $x^{(i)}$,

$$\begin{aligned}\nabla_W \ell(W) =& \nabla_W (\log |W| + \sum_{j=1}^{d} \log(\frac{1}{2}\exp(-|w_j^T x^{(i)}|)) \\ =& \nabla_W (\log |W| - ||Wx^{(i)}||_1) \\ =& W^{-T} - \text{sign}(Wx^{(i)})(x^{(i)})^T.\end{aligned}$$

So the update rule should be

$$W := W + \alpha(W^{-T} - \text{sign}(Wx^{(i)})(x^{(i)})^T).$$

(c) It is hard to distinguish between mixed or split audio tracks, cause they are too noisy 😢.

5 Markov decision processes

(a)

$$\begin{aligned}||B(V_1) - B(V_2)||_\infty =& \max_{s\in S} |V_1'(s) - V_2'(s)| \\ =& \gamma \max_{s\in S} \max_{a\in A} |\sum_{s'\in S} P_{sa}(s')(V_1(s') - V_2(s'))| \\ \leq& \gamma \max_{s\in S} \max_{a\in A} |\sum_{s'\in S} P_{sa}(s')| \max_{s'\in S} |V_1(s') - V_2(s')| \\ =& \gamma ||V_1 - V_2||_\infty \max_{s\in S} \max_{a\in A} |\sum_{s'\in S} P_{sa}(s')| \\ =& \gamma ||V_1 - V_2||_\infty.\end{aligned}$$

(b) We assume that B has at least one fixed point. If it has more than one fixed point, then there must exist V_1 and V_2, such that $B(V_1) = V_1$ and $B(V_2) = V_2$. Therefore $||V_1 - V_2||_\infty = ||B(V_1) - B(V_2)||_\infty \leq \gamma ||V_1 - V_2||_\infty$. Since $\gamma < 1$, $||V_1 - V_2||_\infty = 0$ so that $V_1 \equiv V_2$. By way of contradiction, B has at most one fixed point.

6 Reinforcement Learning: The inverted pendulum

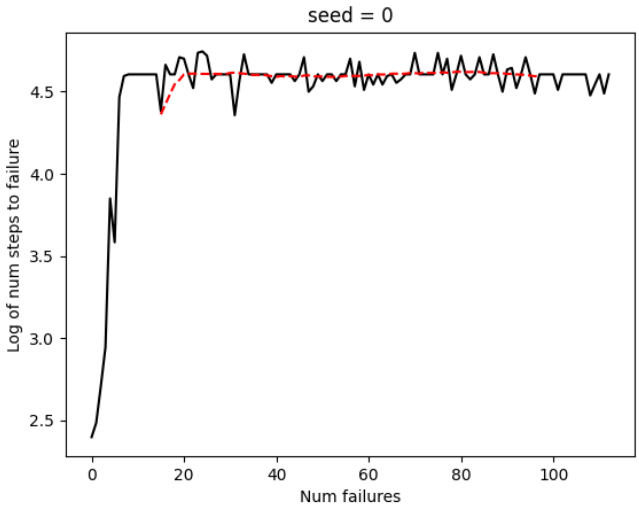

Figure 2: Problem 6 Seed 0

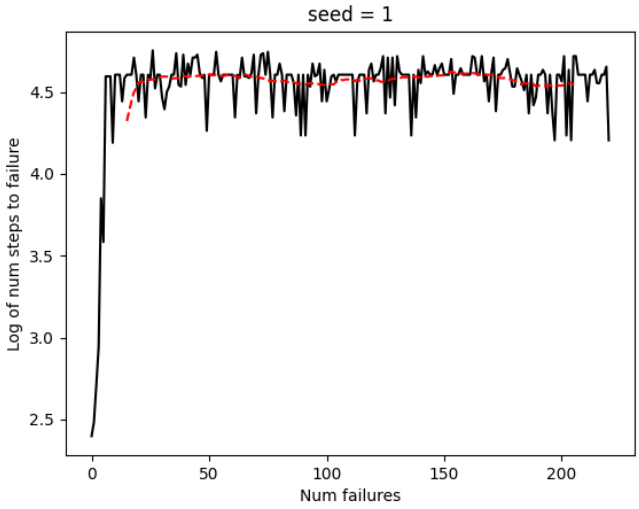

Figure 3: Problem 6 Seed 1

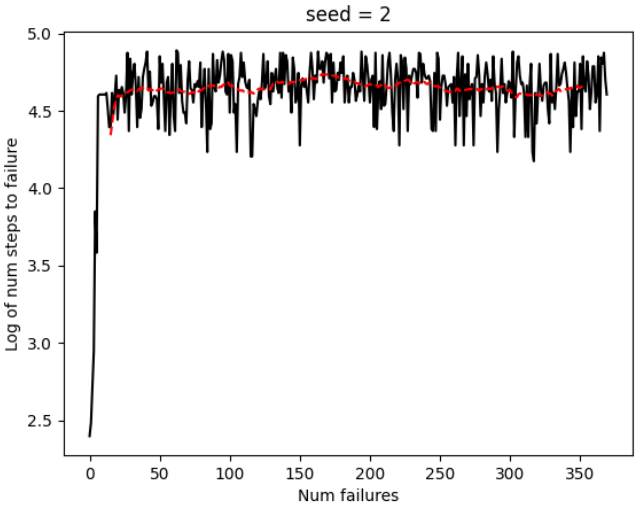

Figure 4: Problem 6 Seed 2

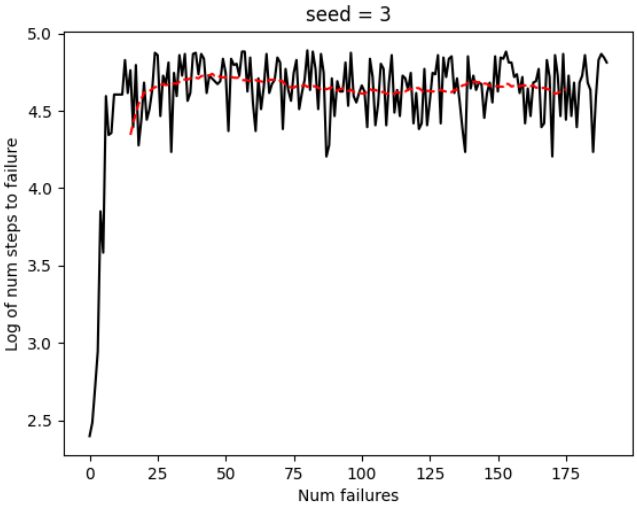

Figure 5: Problem 6 Seed 3

- It takes 113 trials before the algorithm converged.
- The black curves (raw time-steps per trial) show different levels of fluctuation across seeds. Seed 0 has the smallest fluctuations, seeds 1 and 3 are moderate, and seed 2 exhibits the largest fluctuations. The red curves

(smoothed learning curves) for all seeds eventually converge to a roughly stable value around 4.6. However, the number of trials required to reach this stable value varies depending on the seed. This implies that the algorithm is **stochastic**, and its short-term behavior can vary significantly depending on the random initialization, but the long-term performance tends to converge to a similar level.