Deadlines

March 14, 2021

1 3/9

1.1 Goals

1. Work on thesis: finish 2 sections
2. Work on complex: finish 2 problems
3. Work on drl: finish 1 problem

1.2 DRL

Remark 1. Recall: Goal is to learn $v_\pi(s)$ from episodes of experiece under π.(In MC or TD learning)

For Monte Carlo: Update $V(S_t) := V(S_t) + \alpha(G_t - V(S_t))$ over random trajectories

Remark 2. Monte-Carlo: G_t is unbiased estimator of $V_\pi(S_t)$. But potentially high variance

Temportal Difference: $R_{t+1} + \gamma V(S_{t+1})$ is biased estimator but lower variance. True target $R_{t+1} + \gamma v_\pi(S_{t+1})$ is unbiased estimate of $v_\pi(S_t)$

Remark 3. Note this is idea of bootstrapping: using data to generate model which we then use in estimator: estimator uses another estimator.

Remark 4. SARAS and q-learning method of updating q values

2 3/10

2.1 Goals

1. Finish complex/study
2. Study DRL
3. Read evolution

2.2 Complex Analysis

Question 1. If f entire can we expand in powerseries converging everywhere?

2.3 DRL Review

`https://cmudeeprl.github.io/403_website/assets/lectures/s21/s21_rec2_gaussian_process.pdf`

Remark 5. Gaussian Process OPtimization:

`C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 20`

Remark 6. Kernel Cookbook:

`https://www.cs.toronto.edu/~duvenaud/cookbook/`

Remark 7. Example of learning continuous problem: ON some manifold: transition function is $T(s,a) = cos(sa)$ and reward function is $r(s,a) = -s^2$.

Question 2. Difference between $GP - CEM$ and regular CEM?

Remark 8. Limitations of GP:

1. Hard to approximate kernel in DRL
2. COmputation complexity of inference hard $O(n^3)$ (matrix inversion)
3. Hard to design differentiable policy/action optimization techniques
4. Designing multi-variante GPs is hard

Remark 9. GP: Can fully represent epistemic uncertainty, but not allows practical.

Remark 10. Limitations of learning by interaction:

1. needs chance to try and fail many times
2. Hard when safety a concern
3. hard inr eal life which takes time

Remark 11. Challenges in imitation learning:

1. Compounding errors
2. Non-markovian observation
3. Lack of generalization

Remark 12. Compunding errors happen when we make an error which causes us to deviate farther from expert which makes us more likely to make error at next time step.

Fix is to augment training with error cases so we can self correct when necessary

Remark 13. Can concatenate states to make markovian issues nonissues. Just redfine "state". Or use RNNs, which are inherently nonmarkovian, since they feed input as well as transformed input

Remark 14. There is always one optimal policy: $v_*(s) = max_\pi(\pi(s))$

Remark 15. Solving the MDP is finding the state and action value functions given a policy

Remark 16. Optimal value functions measure the best possible goodness of states or state/action pairs under all policies. So actually this is THE optimal policy vs. all others.

Question 3. If the optimal policy is simply the one which maximizes return at each state, what's the problem?

Question 4. I guess the definition is recursive.

Remark 17.

$$\mathbb{E}[G_t|S_t = s] = \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s]$$

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})] = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

Remark 18. Bellman expectation equations give us a S equations linear where S is number of states. Can be solved with linear system solver. q^* unique solution to system of nonlinear equtions

Remark 19. MDP under fixed policy is MRP:

$v_\pi(s) = r_s^\pi + \gamma \sum_{s' \in S} T_{s's}^\pi v_\pi(s')$

where $r_s^\pi = \sum_{a \in A} \pi(a|s) r(s, a)$ and $T_{s's}^\pi = \sum_{a \in A} \pi(a|s) T(s'|s, a)$

Question 5. What does it mean under fixed policy? I thought policy already given? Do we mean deterministic rewards? THis is mathematically plausible

Remark 20. $v_\pi = (I - \gamma T^\pi)^{-1} r^\pi$ where we have a matrix over states T^π which are transitions from one to the other. But matrix inversion costly

The advantage in fixing a policy is that we have a transition matrix(since we know what actions we'll take).

Remark 21. We know there is a unique optimal policy π^* w.r.t total dominance partial ordering $\pi \geq \pi'$

3 3/11

3.1 Goals

3.2 Modeling Evolution

Remark 22. Stochastic Switching: phenotypic hetergeneity despite genotypic uniformity. A bet heding strategy when mutation isn't enough.

Question 6. How is phenotypic configuration preserved if genotype uniform(from one generation to next). What else is passed on (methylation patterns?).

Answer 1. Epigenetic factors are the mediators. Internal fluctations in mRNA transcription and protein translation. Higher number of feedback loops allows for higher number of steady states leading to diff. expressions

Remark 23. Assumptions:

1. Model assumes infinitely large popluation

Question 7. Major vs. modifier locus?

Remark 24. It seems optimal switching rate exactly inversely proportional to n-stability of environment.

Question 8. What about asymmetric environment conditions? Seem more relevant(stable conditions and then shock, followed by more stable conditions)

Question 9. When can a mutation invasion be successful?

Remark 25. Mutation selection balance equation:

$$\mu_M w_A x^2 + (1-\mu_M)(w_A - w_a)x - \mu_M w_a = 0$$

Remark 26. Equilibria x^* stable if $\mu_m > \mu_M$ and unstable if $\mu_m < \mu_M$. Because of matrix eigenvalue stuff. If selection is too high then invader does not invade. No invasion if $\mu_m > \mu_M$. Independent of fitness of values. I $\mu_M > \mu_m$ then unstbale and invasion

0(with 0 mutation rate) cannot be invaded. Optimal mutation rate under this model

Remark 27. Environemtnal sensing: switching phenotypes but in response, not stochastically

Remark 28. Epigenetic transimssion: How are non-genetic factors inherited? Lots of controversial papers about epigenetic inheritance.

Somehow epigentic variance is less risky than genetic variance. So more workable in practice.

Remark 29. Fitness matrix:

$$\begin{bmatrix} 1 & 1-s_0 \\ 1-s_1 & 1 \end{bmatrix}$$

where col corresponds to allele, row corresponds to environment

Question 10. When are reductions between models possible???

3.3 DRL

Remark 30. In TD can update q values after each action instead of after trajectory b/c of recursive update rule

Remark 31. Dealing with large state spaces: Find parameterized function $\hat{v}(S, w)$, parameterized by w. Instead of having a table for all states.

Remark 32. To solve want to minimize least squares problem over w parameters. But no supervisor so need to subsitute target for examples. For example TD Target $R + \gamma\hat{v}(S', \theta)$ is biased example of truth

$\theta \to \theta + \alpha(R + \gamma\hat{v}(S', \theta) - \hat{v}(S, \theta)]\nabla\hat{v}(S, \theta)$

Remark 33. When you don't know the dynamics we need to use q values instead of state values to estimate.

Remark 34. In a similar case when you don't know dynamics in continuous case we parameterize q with $\hat{q}$ and learn

3.4 DRL Review

3.4.1 Path Perspective on Value Learning

`https://distill.pub/2019/paths-perspective-on-value-learning/`

Remark 35. Unlike monte carlo, td updates merged intersections so that return flows backwards to all preceding states.

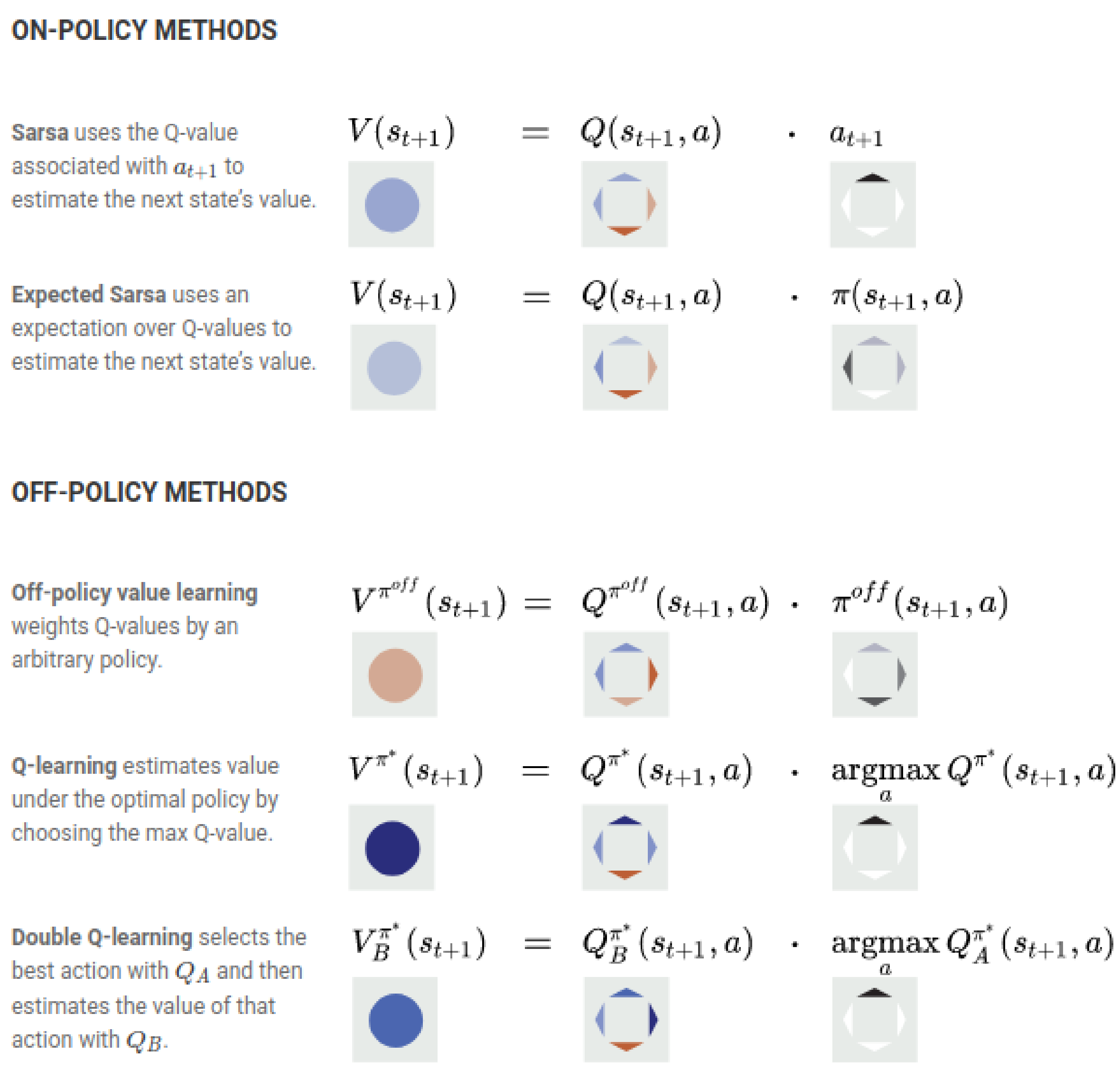
Remark 36. MC averaging over real trajectories whereas TD averaging over all possible paths

Remark 37. TD may tend to outperform MC in tabular environments becuase it averages over at least as many trajectories

Remark 38. SARSA uses $r_t + \gamma Q(s_{t+1}, a_{t+1})$ update rule but not ideal, really want to be using $V(s_{t+1})$. Q learning prunes away all but the highest valued paths

Remark 39. Q learning is biased(cause self-referential) so try to use double q learning to correct

Remark 40. Sarsa, Expected sarsa, q, and double q diff. ways of estimating $V(s_{t+1})$ in a td update



3.4.2 Learning by Cheating

https://arxiv.org/abs/1912.12294

Remark 41. Decompose imitation learning into two stages. First train cheating model copying expert and accessing ground state and then train sensorimotor model copying cheater

Remark 42. Advantages:

1. Privileged agent operates on compact space representation
2. The priveleged agent provides stronger supervision

3. Internal state of privileged agent "white box" ie. can be examined at will

3.4.3 A tutorial on bayesian optimization

`https://arxiv.org/pdf/1012.2599.pdf`

Remark 43. Value iteration(and q iteration) independent of policy.

Remark 44. Policy iteration vs. value iteration. Policy iteration faster under certain conditions. Simply becuase actions change less often. But hard to tell when we've converged.

Value iteration gives us more info.

Note we still compute value function with policy iteration.

Value iteration converges when we have no change. Policy iteration converges when at every we take the maximal action.

These only useful when we have full knowlege of the dynamics

Remark 45. TD/MC useful when we don't know the dynamics.

4 3/14

4.1 Goals

1. Finish complex
2. Thesis
3. Review complex
4. Transcription

4.2 Complex Review

4.2.1 Fourier Stuff

Theorem 1. *Phragmen-Lindelof Lemma: bounds F in a sector if bounded on boundary and sub-exponential*

Proof. $F_\epsilon(z) = F(z)e^{-\epsilon z^{3/2}}$. By construction $cos(3\theta/2)$ positive so we get good decay for F_ϵ. Then if $|F_\epsilon| \leq 1$ then $|F| \leq 1$ via continuity.

Let $M = sup|F_\epsilon|$ then $\exists w_j \to w$ toward M. It must be $w \in \partial S$ which is bounded by 1. So done. Key is that w_j are bounded since $F_\epsilon \to 0$ as $|z| \to \infty$. □

Theorem 2. *If* $f \in \mathcal{F}_a$ *then* $|\hat{f}(\xi)| \leq B_f e^{-2|\xi|}$ *for* $0 \leq b < a$

Proof. If $b = 0$

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x)e^{-2\pi i x\xi}dx \implies |\hat{f}(\xi)| \leq \int_{\mathbb{R}} |f(x)|dx \leq \int_{\mathbb{R}} \frac{A_f}{1+x^2}dx = \pi A_f$$

If $b > 0$ the idea is to shift contour of integration down imaginary line. Note vertical sides go to 0 as $R \to \infty$ since norm is large. So can shift down with a negation. □

Theorem 3. *Fourier Inversion:*

$f(x) = \int_{\mathbb{R}} \hat{f}(\xi)e^{2\pi i x\xi}dx$

Proof. First note when $A > 0$ and B real

$$\int_0^\infty e^{-(A+iB)\xi}d\xi = \frac{1}{A+iB}$$

Via checking the finite case and sending to ∞.

Then we argue by splitting across im line. In the positive case we can simply use definition and interchange integration and resolve with cauchy's integral formula. For the other case we consider a reverse contour and apply the current result. □

Theorem 4. *If* $f \in \mathcal{F}$ *then*

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n)$$

Proof. Key Idea 1: Idea is to pick out points being summed as residues. First note $\frac{1}{e^{2\pi i z}-1}$ has simple poles with residue $1/2\pi i$ at integers. Then Apply residue formula to $\frac{f(z)}{e^{2\pi i z}-1}$ which generates residues with $\frac{f(n)}{2\pi i}$. Integrating over rectangle contour(off integers).

Key idea 2: We then CLEVERLY rewrite $\frac{1}{e^{2\pi i z}-1} = -\sum e^{2\pi i n z}$ if $|z| < 1$ and similarly for complement case. Allows us to rewrite as fourier transform □

Remark 46. Residue Formula Computation Tools:

Idea is to find contour s.t.

$$\lim_{R\to\infty} \int_{\gamma_R} f(z)dz = \int_{-\infty}^{\infty} f(x)dx$$

Which is easier to evaluate because we simply compute residues

EX 1:

Consider

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi$$

Using the half circle γ_R we have

$$\int_{\mathbb{R}} \frac{1}{1+x^2}dx = -\lim \int_{\gamma_R} \frac{1}{1+z^2}dz = -\lim \int_{\gamma_R} \frac{1}{(z-i)(z+i)}dz =$$

Partial fraction decomposion yields $\frac{1}{(z-i)(z+i)} = \frac{1}{2i}\frac{1}{z-i} - \frac{1}{2i}\frac{1}{z+i}$ so integrating over half ciricle gives $2\pi i/2i = \pi$.

Further note we have equality since the integral over the polar section goes to 0(b/c of large norm).

EX 2:

5 3/16

5.1 Goals

1. Go to open house
2. Revamp site/apply to summer stuff
3. Finish DRL