

CS 760: Machine Learning **Reinforcement Learning III**

University of Wisconsin-Madison

Josiah Hanna

December 7, 2023

Announcements

- Homework 7 due December 12 at 9:30 am.
- Final exam: December 18 from 2:45 4:45 pm in the Social Sciences building.
- Course evaluations available until 12/13.
 - Currently at 52% participation. > 75% to receive 2 points extra credit on final.

Lecture Goals

At the end of today's lecture, you will be able to:

- 1. Explain the key techniques necessary for using neural networks in qlearning.
- 2. Explain the policy gradient objective and optimization approach.
- 3. Explain how to combine value-based and policy-based reinforcement learning to obtain actor-critic methods.

Beyond Tables

So far:

- Represent everything with a table
 - Value function V: table size
 - •Q function: table size $|S| \times |A|$
- Too big to store in memory for many tasks
 - Backgammon: 10²⁰ states. Go: 3³⁶¹ states
 - Need some other approach

$|S| \times 1$

Concerts of	County of Hislory for and received, and returned to the AUDIALS. Post Childs. Post Law and Repleton	Rate of S	a partment partment 10 Eans 10 . 4.	deni deni	nd in ng the	face face in qu	hat Of	For end	sing .	Ale some 3	den	. 18	state ofB_la	120	Property Street	And the second	and a state	ia	i	Parties.	1
NEW DATA DATA DATA DATA DATA DATA DATA DAT	ANDELES, Possil Golds	EDMARS OR COTTINO,	ти Елер и . 4/,	A man		tra qu ja	- pro	The Qu	oung .	June 2	39, 1.			1.00		12			Company of the local division of the local d		
Name Normal	ANTICLES. Postel Godds	UNITED OF	04 Ease 10 - - - 4/,	101	11	ten Qu	-	Bre Qu	APRIL 1			19%		13	Ascent.	41 ***	second.	ill	[11	115	2
Normal Goudon m <	Postal Golds	007111308.	и. 4.	1	11				Granten, Dr. Granten, H. H. H			Rear in last loss for smill,		28.15		-		8 (%	5 44	1	
Name 4	Postal Galda		4.	1		1	1	1	ŧ1	4	-1	Rath ar Res	One cost collecty policy marging	E						240	3
Normal Dama III. Deployment 44 4 <td< td=""><td>Post Levs and Republican</td><td></td><td>4.</td><td></td><td>12</td><td>11</td><td>£1</td><td>1</td><td>13</td><td>1</td><td>11</td><td>Free Inc.</td><td>Parrowt orderary putting damples of</td><td></td><td></td><td></td><td></td><td></td><td></td><td>5 85</td><td>4</td></td<>	Post Levs and Republican		4.		12	11	£1	1	13	1	11	Free Inc.	Parrowt orderary putting damples of							5 85	4
Nome have the large block	Point Law and Regulations			1				-			14	-	Francest collinary podage entropy							4 00	1.2
Description Image:	Patrarkig Samp. Radig Samp. Jah											6-	Tra-and onlinery policy many								1.5
Name	Paring Rampt												Thistowest ordinary postage stampt								1.4
All	Fals												Maniy cost orderry policy station			-				20	
Print	Pada												Special delivery elamps	-						314	10
Table												1	Our wat postd carls (International)			- ×.				100	11
Data air propies many												1	One-cost portugador strange							- 134	19
Table and particular distance and particular distance Table and particular distance and particular distance State Conce and particular distance and particular distance State Conce and particular distance and particular distance State Conce Conce and particular distance and particular distance State Conce Conce Conce and particular distance Conce and particular distance State Conce													ent potago-los stamps,	-	-					100	10
Parking and parking and parking and parking State of Musics Bit of Data Law and Data law and the Data lawa and the Data lawa and the Data law and the Data law and the Dat													and patter de charpe	-							10
Table	Doks												wai petge-berstanje								154
Come Image: Section of the section	Tables				-								and support printed stops								17
Balan	Ches												out a copuper & periodical stamps					-			18
Bands	Belle									1			mat sew-paper & periodical enough		-		-	-			19
Mail Key, SA, A.J.J.J., Annotation and the second seco	Boord Books									-6-		4	wat som paper & paris Lock stamps					-			20
Xall Key, No. 444 And A. Gald													and sempoper & periodical straige								21
Mail Rep: No. 44/1 2011 Garde Ga													out sempsys & periodical ensure								12
Much Age, Ma, Mi Advid, Gaple Ga	Mail Key, Na. 44 2. 1 Joursen	dan										1	One wat special request carefugut	-							94
[™] Exercised this <u>B</u> / <u>a</u> / <u>day</u> of <u>fully</u> , 1899 from <u>Rufac</u> h (fifts, <u>bits P. M.</u> the following frameword articles in the Prot Office at <u>Mark productions frameword articles in the Prot Office at</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark productions frameword articles in the Prot Office at <u>Mark productions</u> <u>Mark pr</u></u></u></u></u></u></u></u>	Vail & up No 21. 2. 4. J.	basele											rent special respect samplinger								25
* Treverine of this B1 at day of fully , 187/f from Buffet h 4 files, the files of the server of the file of the f													An out through carrieges							1.40	28
*Received this B/ at day of fuely, 1870 from Rufu bills, construct bills, and the base of											1		mut shaped cardison.							223	27
* Received this 61/22 day of			01			and		W	4.1	G	41	alda.	Letter shark				-	-			28
Inter P. M., the full opting ing manned articles in the Part Office at (1,2,2,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4	*Received this . 5/ at	day of -	July	1		7 E I	rom.	- N.	10	.18	16.	hi to	On-old hempiper weppers							-	27
India P. J. L., the presenting realized production in the constraint of the Chailed Studer, rist: Image: Constraint of the Chailed Studer, rist: <t< td=""><td>a sea a sale deserved</td><td>articles in the</td><td>Past Offic</td><td>e al .</td><td></td><td>Class</td><td>det .</td><td>, O</td><td>hose y</td><td>15.0</td><td>Xirs.</td><td>and as</td><td>and mandades, and done service</td><td>-</td><td></td><td></td><td></td><td>1</td><td></td><td></td><td>20</td></t<>	a sea a sale deserved	articles in the	Past Offic	e al .		Class	det .	, O	hose y	15.0	Xirs.	and as	and mandades, and done service	-				1			20
Specific of Distance	late P. M., the following numer	1.1	and the floor	Parite.	d Ste	des. ri	2.7						DVDsi#						_	30 28	1.7
Since Postal Oxides	State of Blie	, beconge	27 10 ture		-								Different from how to provide the state of t	A 1		_	-	1	-	a	-
Arrist Dark of Paul Laws and Equilibrium, Dark Arrist Dark of Paul Laws and Equilibrium, Dark Arrist Later Solo, Solo, Care, Dark of Paul Laws and Equilibrium, Solo, Theras Harrist Roompoon Solo, Care, Dark of Paul Laws and Equilibrium, Solo, Theras Harrist Roompoon Solo, Care, Dark, Dark, Dark, Dark, <t< td=""><td colspan="9">View Part Galles</td><td></td><td>Plangt tald during the mouth as shown</td><td>dara .</td><td></td><td></td><td></td><td>Dr. Person</td><td>Special D</td><td>CO 35 36</td><td>20.2</td></t<>	View Part Galles										Plangt tald during the mouth as shown	dara .				Dr. Person	Special D	CO 35 36	20.2		
Letter bains, Salo, Data Pattark Big Bangs, Jack J, Bang Bang, Data Pattark Big Bangs, Jack J, Bang, Bang, Data Pattark Big Bangs, Jack J, Bang, Bang, Data Pattark Big Bangs, Jack J, Bang, Bang, Data Pattark Bang, Bang, Jack J, Bang, Bang, Data Pattark Bang, Jack J, Bang, Bang, Data Pattark Bang, Bang, Bang, Jack J, Bang,	Ares Task of Postal Law and Equision, Own.										Call firsts material, i.e.	-				Annes	The later	hip and stee	mbes		
Image: Statistic Statistic Statistic Jpc. Ch., Encode Reads Image: Statistic Statistic Statistic Jpc. Ch., Encode Reads Image: Statistic Statis Statis Statistic Statistic Statistic Statistic Statistin Statis	ALCA Letter Sales					Side,							full Born maked, Sec	A PER SPA	carbon, B			Amereo	I paid the	when the Re-	1.
Indiag Plantpin Mair Key, Xu, Leff M, Jun MMAA Mair Key, Xu, Leff M, Jun M, Jun Key, Xu, Leff M, Jun MMAA Mair Key, Xu, Leff M, Jun M, Jun Key, Xu, Leff M, Jun M, Ju	Ware Patrarking Sumpty			yo	all.	. Boord	Dayles.					-	Lord Bourn maked, So.	i per qu				Aller and	1 pail		-
Concert Tak, Main Key, Ke. (2000), 100 Fill Fill Transla test (100) Fill Fill Transla test (100) Fill Fill Fill Fill Fill Fill Fill Fill	Todag Namps								11.9			91828	Lord Daniele manak, Natural and	a part age	1475-0, B			- Louis	i paid		
Image: Data Imag	any Care Isk.					. Mil 3	67.34	-40	See.	17.00			Dealt by band at and of had severily was	straint.				Annes	paid		-
**Prime a distance of physical and the set of physical and the field in a set of physical and the	Mary Pals,												Dad Ju.			- F. M.		Transfe	of the Manage	Other Assa	
¹ When a strategy of Protessanse score the new P. N. and Cl is the above party with the lifes, cancer d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S. S. S. Obse, Crowy, Subs, at parts of relations of the score d law P. S.					1	here .	A-C	416	16			Tutterality.	Transition Street Manage Contract of			in Harris		Transf	it for Manag	Other Asso	- 444
*When a distance words the are P. W. and C. In the dates word for P. W. Other, Every Annual Association of product and the second of and write. *When a distance of the tensor words of the tensor words for P. W. Other, Every Annual Association of product and the second of and write. The and are reacted of and write the date of the second is the T. W. Other, Every Annual Association of product and the second of and write the second of and write the second of and the first of the second o					-								Transfer from Manay Oaks Arrange			H 11		Trande	e to Motor	Online Assos	
*When a statego of Petasanan source the tee P, St. and the data base ranged with the first, source like P, St. One, Devidy, Rate, and particular and the state and particular and the state and the st														-	There are a	· · · ·		Trand	e in Manag	Other Asses	
**When a sitting of Photometer source (all now P, X, and C) is the dates source in P, S, Ohe, Tony, Inde, all source of columnet Prod.																		Peak B			
*When a stateport Publication sources that one P. N. and C. It is the above month of the Line of the C. Tard.							L OBA	Course of	Rafe, I	and passive	11-6 20	COM DESIGN						Anone Steeres	Copyright I		-
Vision a datago of Parlament source for a serie of the series The series of the series of the series The series of the series of the series The series of the series of the series of the series The series of the series of the series of the series The series of the series		to M. search 210 in the above	on people with	the Rules	a barrent	and the second												Balator	m hand,		
Press his produces of a first band in which the deal is which it is a first band in the deal is which it is a first band in the deal in the deal is a first band in the deal in th	Although a plange of Puppman of South and South and	uplicate of this pursiple.											Tural.			-					
Dis same	from his producement, giving him or maticle should be a	and the second second																		The Party of the P	
	The sension manual												and the second						Pier .	A	





Semi-gradient Q-Learning

- Instead of using a table, represent action-values as a function with learnable parameters, $q(s, a, \theta)$.
- Semi-Gradient Q-learning:

•
$$\theta_{t+1} \leftarrow \theta_t + \alpha(r_{t+1} + \gamma \max_{a'} q(s_{t+1}, a', \theta_t) - q(s_t, a_t, \theta_t)) \nabla q(s_t, a_t, \theta_t)$$

Temporal difference error

- Use back propagation to compute gradient of $q(s, a, \theta)$ for any (s, a).

• Example: parameter vector, θ_t , could be all weights and biases of a neural network.

Adjust each weight in proportion to gradient of output times temporal difference error.

Deep Q-learning



Mnih et al, "Human-level control through deep reinforcement learning"



Stability with Neural Networks

- Neural networks are typically trained with i.i.d. data and fixed targets.
 - $x_i, y_i \sim D$ and we are learning some underlying function such that $f(x_i) = y_i$.
- Using neural networks with Q-learning breaks both assumptions.
 - Training may be unstable and diverge; lacks theoretical guarantees.
- Deep Q-Network (DQN) uses two key methods to stabilize training:
 - Experience replay: keep around old data to update network.
 - Target networks: Use an older copy of network parameters to compute the target for updates; update this older copy at a slower rate than main parameters used.

Experience Replay

- The basic semi-gradient Q-learning algorithm processes (*s*, *a*, *s'*, *r*) transitions as they are experienced and then discards them.
- Experience replay: save the most recent transitions (in DQN, the past 1 million) and use a random subset to update the action-value function.
 - Re-uses data and reduces correlation between samples.
 - Learning becomes more like supervised neural network training where we train from a static data set.
- Other choices besides random subset can improve performance [1].

[1] Prioritized Experience Replay. Schaul et al. 2015.

- The basic Q-learning algorithm always uses the most recent action-values to form the training target $r_{t+1} + \gamma \max_{a'} q(s_{t+1}, a', \theta)$
- DQN uses a separate target network to compute $\gamma \max q(s_{t+1}, a', \theta)$.
 - The target network is infrequently updated by setting the target network parameters to be the same as the main network's parameters, i.e., $\tilde{\theta} \leftarrow \theta_{-}$
 - Makes the learning target more stable as in supervised learning.

Target Networks

DQN Architecture



Looking Forward

- DQN (arguably) launched a surge of interest in deep reinforcement learning that has led to many exciting new applications and RL developments.
- DQN is widely used in practice though many improvements have been made.

Rainbow: Combining Improvements in Deep Reinforcement Learning. Hessel et al. 2018. https://www.deepmind.com/blog/agent57-outperforming-the-human-atari-benchmark



Quiz

state sequence, B, A, G, G.



r(B) = 20; r(A) = 10; r(C) = 20; r(G) = 100

Consider the following MDP which has deterministic transitions and $\gamma = 0.8$. In each state, the agent can either stay in the current state or take an action that takes it to the state given by the red arrows. We run q-learning with all action-values initially set to 0. Write the action-value table after observing the



Model-Free Reinforcement Learning



Policy-based RL

- acts greedily w.r.t. values.
- Policy-based methods instead explicitly learn the policy.

•
$$\pi_{\theta}(a \mid s) = \Pr(A_t = a \mid S_t = s; \theta)$$

Objective is to find policy that maximizes expected future reward:

 $J(\theta) := \sum_{s} \mu_{\theta}(s) v_{\pi_{\theta}}(s)$ Frequency of visitations to s under π_{θ}

So far the policy is implicit. Q-learning learns a value function and then

Expected sum of discounted future rewards from s under π_{θ}

Why Policy-based?

- Advantages to policy-based methods?
 - More easily handle continuous actions.
 - Policy gradient theorem provides stronger convergence guarantees under function approximation.
 - Useful for partial observability.
 - Policy may be simpler to approximate.
- Disadvantages? \bullet
 - May be easier to approximate action-values.
 - Policy is a simple function of the action-values.

Policy Parameterizations

- Policy can be any parameterized and differentiable distribution.
- Need $\pi_{\theta}(A_t = a \mid s)$ and $\nabla_{\theta} \pi_{\theta}(A_t = a \mid s)$ exists.
- For discrete action RL tasks, typically use a softmax distribution with logits given by a neural network.
 - Same model that we use for multi-class classification.
- For continuous action RL tasks, typically use a Gaussian distribution with mean and variance each given by a neural network.
 - Same model used for multiple regression.

Policy Gradient Theorem

$$J(\theta) := \sum_{s} \mu_{\theta}(s) v_{\pi_{\theta}}(s)$$

$$\nabla_{\theta} J(\theta) \propto \sum_{s} \sum_{a} \mu_{\theta}(s) \pi_{\theta}(a \mid s) q_{\pi_{\theta}}(s)$$
$$= \mathbf{E}[q_{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(s)]$$

- The direction in which an infinitesimally small change to θ produces the maximum increase in $J(\theta)$.
- Nice property: $\nabla_{\theta} J(\theta)$ does not depend on any gradients of p or $\mu_{\theta}(s)$.
 - Only have to differentiate policy which is known by the learner.

- $(s, a) \nabla_{\theta} \log \pi_{\theta}(a \mid s)$
- $[a \mid s) \mid s \sim \mu_{\theta}, a \sim \pi_{\theta}(\cdot \mid s)]$ ally small change to θ produces the

REINFORCE

- directly computed.
 - But it can be estimated using data obtained by running π_{θ} .
- Approximate $q_{\pi_{a}}(s_{t}, a_{t})$ with the sum of discounted rewards following s_{t} and a_t , i.e., G_t .
- $\theta_{t+1} \leftarrow \theta_t + \alpha G_t \nabla_{\theta} \log \pi(a_t | s_t)$

• $\nabla_{\theta} J(\theta) = \mathbf{E}[q_{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a \mid s) \mid s \sim \mu_{\theta}, a \sim \pi_{\theta}(\cdot \mid s)]$ cannot be

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \theta)$ Algorithm parameter: step size $\alpha > 0$ Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to **0**)

Loop forever (for each episode): Loop for each step of the episode $t = 0, 1, \ldots, T - 1$: $\begin{aligned} G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \\ \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t \boldsymbol{G} \nabla \ln \pi (A_t | S_t, \boldsymbol{\theta}) \end{aligned}$

Usually dropped in practice

Is the policy gradient a gradient? Nota and Thomas. 2020. **Bias in Natural Actor-Critic Algorithms. Thomas. 2014.**

REINFORCE

- Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot | \cdot, \boldsymbol{\theta})$

 (G_t)

Actor-Critic Methods

- Basic REINFORCE does not use a value function.
 - Learns directly from the sum of discounted rewards following an action.
- Actor-critic methods use learned value functions to drive policy changes.
- New update:
 - $\delta_t \leftarrow R_{t+1} + \gamma \hat{v}(S_{t+1}) \hat{v}(S_t)$
 - $\theta_{t+1} \leftarrow \theta_t + \alpha \delta_t \nabla_{\theta} \ln \pi(A_t | S_t)$

Actor-Critic Methods

One-step Actor–Critic (episodic), for estimating $\pi_{\theta} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \theta)$ Input: a differentiable state-value function parameterization $\hat{v}(s, \mathbf{w})$ Parameters: step sizes $\alpha^{\theta} > 0, \ \alpha^{\mathbf{w}} > 0$ Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to **0**) Loop forever (for each episode): Initialize S (first state of episode) $I \leftarrow 1$ Loop while S is not terminal (for each time step): $A \sim \pi(\cdot | S, \theta)$ Take action A, observe S', R(if S' is terminal, then $\hat{v}(S', \mathbf{w}) \doteq 0$) $\delta \leftarrow R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$ $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S, \mathbf{w})$ $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$ 1 1 $S \leftarrow S'$

Comparing REINFORCE and Basic Actor-Critic

- REINFORCE update: $\theta \leftarrow \theta + \alpha \sum \tilde{G}_t$ t=0• Actor-Critic update: t=0
- Can generalize learning signal:

$$\delta_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n v_{\pi_\theta}(S_{t+n}) - v_{\pi_\theta}(S_t)$$

If t + n is greater than termination step T then R_{t+n} is taken to be zero.

• For comparison, assume that actor-critic only updates at end of episodes.

$$\nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t) \qquad \tilde{G}_t = \left(\sum_{t'=t}^T \gamma^{t'} R_{t'+1}\right) - v_{\pi_{\theta}}(A_t)$$

 $\theta \leftarrow \theta + \alpha \sum \delta_t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \qquad \delta_t = R_{t+1} + \gamma v_{\pi_{\theta}}(S_{t+1}) - v_{\pi_{\theta}}(s_t)$

Advantage Actor-Critic (A2C)

- Basic multi-step actor-critic method that works well with deep networks.
 - Policy and value function are represented as neural networks with parameters θ and ϕ respectively.
- A2C alternates collecting n steps of experience in task environment and then updating a state-value function and a policy with the learning signal $\delta_t^{(n)}$. $\delta_{\star}^{(n)} = R_{t+1} + \gamma R_{t+2} + \dots$
- Training losses:

*Must stop gradient through value estimate at times greater than t; be careful to not update critic when optimizing actor.

 $L(\phi) = \sum_{t} (\delta_t^{(n)})^2$

t=0

$$+\gamma^{n-1}R_{t+n}+\gamma^n v_{\phi}(S_{t+n})-v_{\phi}(s_t)$$

$$L(\theta) = \sum_{t=0}^{n} \delta_t^{(n)} \log \pi_{\theta}(a_t | s_t)$$

Summary

- Deep Q-learning: approximates Q-learning with deep neural networks.
- Policy-based methods: directly learn policy with gradient ascent.
- Actor-critic methods: learn value functions (critic) that provide a learning signal for improving the policy (actor).

Slides adapted from Advanced Topics in RL and based on Chapter 13 of Reinforcement Learning: An Introduction.

Thanks Everyone!