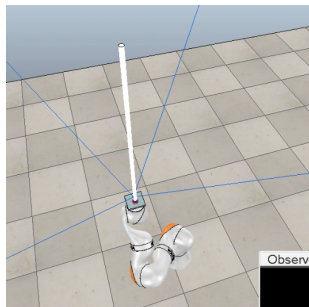
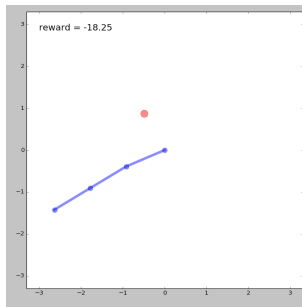
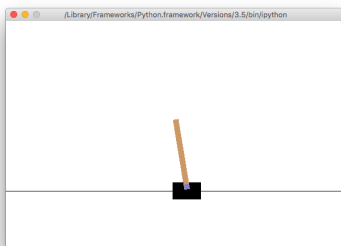
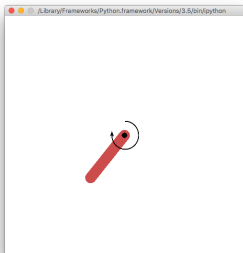


# Model-Free Reinforcement Learning

Mathias Winther Madsen

March 1, 2017

# Policy Gradient Methods



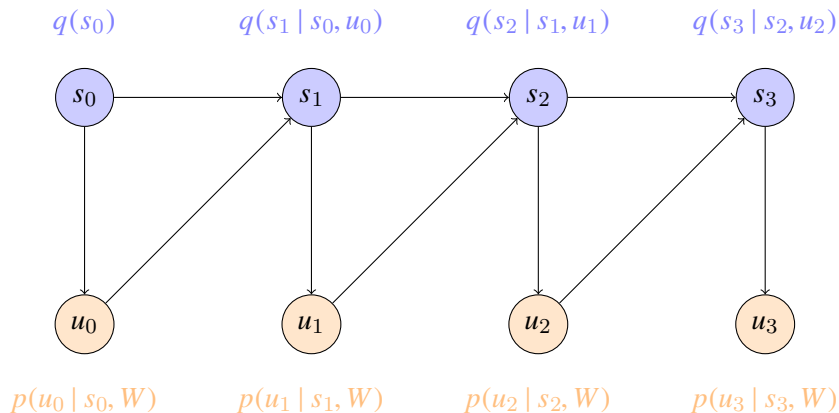
# Policy Gradient Methods

The REINFORCE algorithm: Williams, “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning” (*Machine Learning*, 1992)

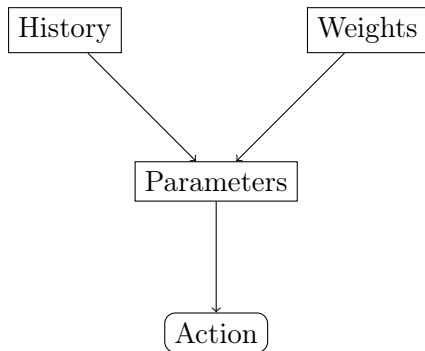
Task	Random	REINFORCE	TNPG	RWR	REPS	TRPO	CEM	CMA-ES	DDPG
Cart-Pole Balancing	77.1 ± 0.0	4693.7 ± 14.0	<b>3986.4 ± 748.9</b>	<b>4861.5 ± 12.3</b>	565.6 ± 137.6	<b>4869.8 ± 37.6</b>	4815.4 ± 4.8	2440.4 ± 568.3	4634.4 ± 87.8
Inverted Pendulum*	-153.4 ± 0.2	13.4 ± 18.0	<b>209.7 ± 55.5</b>	84.7 ± 13.8	-113.3 ± 4.6	<b>247.2 ± 76.1</b>	38.2 ± 25.7	-40.1 ± 5.7	40.0 ± 244.6
Mountain Car	-415.4 ± 0.0	-67.1 ± 1.0	<b>-66.5 ± 4.5</b>	-79.4 ± 1.1	-275.6 ± 166.3	<b>-61.7 ± 0.9</b>	-66.0 ± 2.4	-85.0 ± 7.7	-288.4 ± 170.3
Acrobot	-1904.5 ± 1.0	-508.1 ± 91.0	-395.8 ± 121.2	-352.7 ± 35.9	-1001.5 ± 10.8	-326.0 ± 24.4	-436.8 ± 14.7	-785.6 ± 13.1	<b>-233.6 ± 5.8</b>
Double Inverted Pendulum*	149.7 ± 0.1	4116.5 ± 65.2	<b>4455.4 ± 37.6</b>	3614.8 ± 368.1	446.7 ± 114.8	<b>4412.4 ± 50.4</b>	2566.2 ± 178.9	1576.1 ± 51.3	2863.4 ± 154.0
Swimmer*	-1.7 ± 0.1	92.3 ± 0.1	<b>96.0 ± 0.2</b>	60.7 ± 5.5	3.8 ± 3.3	<b>96.0 ± 0.2</b>	68.8 ± 2.4	64.9 ± 1.4	85.8 ± 1.8
Hopper	8.4 ± 0.0	714.0 ± 29.3	<b>1155.1 ± 57.9</b>	553.2 ± 71.0	86.7 ± 17.6	<b>1183.3 ± 150.0</b>	63.1 ± 7.8	20.3 ± 14.3	267.1 ± 43.5
2D Walker	-1.7 ± 0.0	506.5 ± 78.8	<b>1382.6 ± 108.2</b>	136.0 ± 15.9	-37.0 ± 38.1	<b>1353.8 ± 85.0</b>	84.5 ± 19.2	77.1 ± 24.3	318.4 ± 181.6
Half-Cheetah	-90.8 ± 0.3	1183.1 ± 69.2	<b>1729.5 ± 184.6</b>	376.1 ± 28.2	34.5 ± 38.0	<b>1914.0 ± 120.1</b>	330.4 ± 274.8	441.3 ± 107.6	<b>2148.6 ± 702.7</b>
Ant*	13.4 ± 0.7	548.3 ± 55.5	<b>706.0 ± 127.7</b>	37.6 ± 3.1	39.0 ± 9.8	<b>730.2 ± 61.3</b>	42.2 ± 5.9	17.8 ± 15.5	326.2 ± 20.8
Simple Humanoid	41.5 ± 0.2	128.1 ± 34.0	<b>255.0 ± 24.5</b>	93.3 ± 17.4	28.3 ± 4.7	<b>269.7 ± 40.3</b>	60.6 ± 12.9	28.7 ± 3.9	99.4 ± 28.1
Full Humanoid	13.2 ± 0.1	262.2 ± 10.5	<b>288.4 ± 25.2</b>	46.7 ± 5.6	41.7 ± 6.1	<b>287.0 ± 23.4</b>	36.9 ± 2.9	N/A ± N/A	119.0 ± 31.2
Cart-Pole Balancing (LS)*	77.1 ± 0.0	420.9 ± 265.5	<b>945.1 ± 27.8</b>	68.9 ± 1.5	898.1 ± 22.1	<b>960.2 ± 46.0</b>	227.0 ± 223.0	68.0 ± 1.6	
Inverted Pendulum (LS)	-122.1 ± 0.1	-13.4 ± 3.2	<b>0.7 ± 6.1</b>	-107.4 ± 0.2	-87.2 ± 8.0	<b>4.5 ± 4.1</b>	-81.2 ± 33.2	-62.4 ± 3.4	
Mountain Car (LS)	-83.0 ± 0.0	-81.2 ± 0.6	<b>-65.7 ± 9.0</b>	-81.7 ± 0.1	-82.6 ± 0.4	<b>-64.2 ± 9.5</b>	<b>-68.9 ± 1.3</b>	<b>-73.2 ± 0.6</b>	
Acrobot (LS)*	-393.2 ± 0.0	-128.9 ± 11.6	<b>-84.6 ± 2.9</b>	-235.9 ± 5.3	-379.5 ± 1.4	<b>-83.3 ± 9.9</b>	-149.5 ± 15.3	-159.9 ± 7.5	
Cart-Pole Balancing (NO)*	101.4 ± 0.1	616.0 ± 210.8	<b>916.3 ± 23.0</b>	93.8 ± 1.2	99.6 ± 7.2	606.2 ± 122.2	181.4 ± 32.1	104.4 ± 16.0	
Inverted Pendulum (NO)	-122.2 ± 0.1	6.5 ± 1.1	<b>11.5 ± 0.5</b>	-110.0 ± 1.4	-119.3 ± 4.2	<b>10.4 ± 2.2</b>	-55.6 ± 16.7	-80.3 ± 2.8	
Mountain Car (NO)	-83.0 ± 0.0	-74.7 ± 7.8	<b>-64.5 ± 8.6</b>	-81.7 ± 0.1	-82.9 ± 0.1	<b>-60.2 ± 2.0</b>	-67.4 ± 1.4	-73.5 ± 0.5	
Acrobot (NO)*	-393.5 ± 0.0	<b>-186.7 ± 31.3</b>	<b>-164.5 ± 13.4</b>	-233.1 ± 0.4	-258.5 ± 14.0	<b>-149.6 ± 8.6</b>	-213.4 ± 6.3	-236.6 ± 6.2	
Cart-Pole Balancing (SI)*	76.3 ± 0.1	431.7 ± 274.1	<b>980.5 ± 7.3</b>	69.0 ± 2.8	702.4 ± 196.4	<b>980.3 ± 5.1</b>	746.6 ± 93.2	71.6 ± 2.9	
Inverted Pendulum (SI)	-121.8 ± 0.2	-5.3 ± 5.6	<b>14.8 ± 1.7</b>	-108.7 ± 4.7	-92.8 ± 23.9	<b>14.1 ± 0.9</b>	-51.8 ± 10.6	-63.1 ± 4.8	
Mountain Car (SI)	-82.7 ± 0.0	-63.9 ± 0.2	<b>-61.8 ± 0.4</b>	-81.4 ± 0.1	-80.7 ± 2.3	<b>-61.6 ± 0.4</b>	-63.9 ± 1.0	-66.9 ± 0.6	
Acrobot (SI)*	-387.8 ± 1.0	<b>-169.1 ± 32.3</b>	<b>-156.6 ± 38.9</b>	-233.2 ± 2.6	-216.1 ± 7.7	<b>-170.9 ± 40.3</b>	-250.2 ± 13.7	-245.0 ± 5.5	
Swimmer + Gathering	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Ant + Gathering	-5.8 ± 5.0	-0.1 ± 0.1	-0.4 ± 0.1	-5.5 ± 0.5	-6.7 ± 0.7	-0.4 ± 0.0	-4.7 ± 0.7	N/A ± N/A	-0.3 ± 0.3
Swimmer + Maze	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Ant + Maze	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	N/A ± N/A	0.0 ± 0.0

Table: Duan, Chen, Houthoofd, Schulman and Abbeel, “Benchmarking Deep Reinforcement Learning for Continuous Control,” *Proceedings of the ICML*, 2016.

# Reinforcement Learning = Game Theory

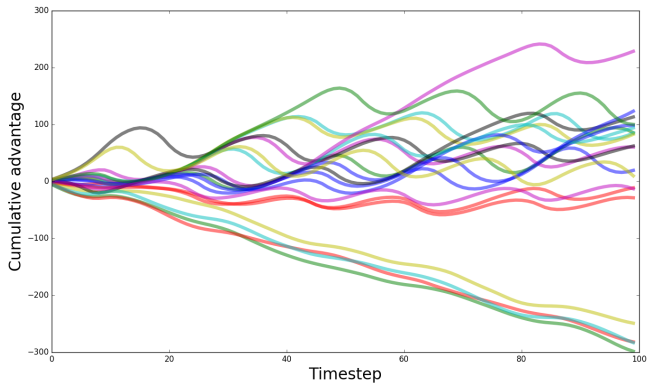


# Reinforcement Learning = Game Theory



$\text{Action} \sim \text{distribution}(\text{history}, \text{weights})$

# The Policy Gradient Method



# The Policy Gradient Method

For  $I$  epochs:

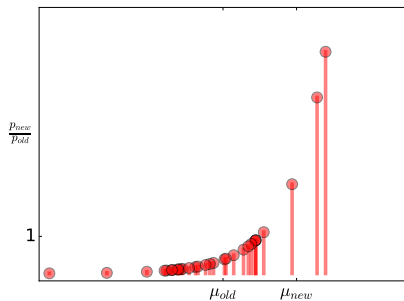
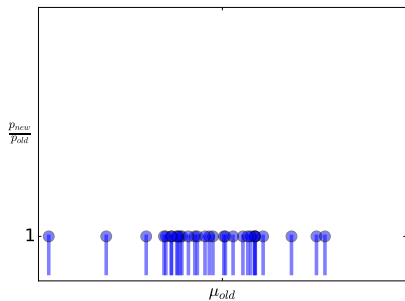
Collect  $N$  episodes, using  
your stochastic policy;

Rate your actions according to  
their (empirical) consequences.

Change  $W$  so that the good  
actions become more probable.

# Importance Weighting

$$E_{new}[X] = E_{old}\left[X \cdot \frac{p_{new}}{p_{old}}\right]$$



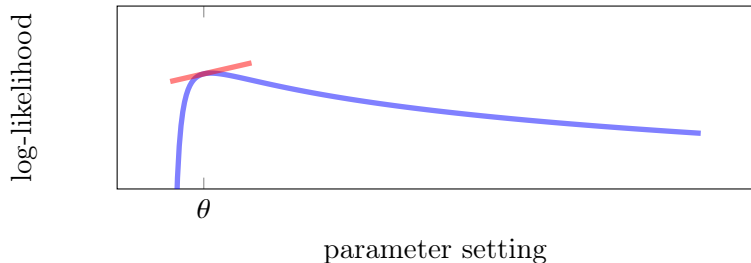


# The Policy Gradient

$$\nabla_W E_W [\textit{Reward}] = E_{W_0} \left[ \textit{Reward} \cdot \frac{\nabla_W p_W}{p_{W_0}} \right]$$

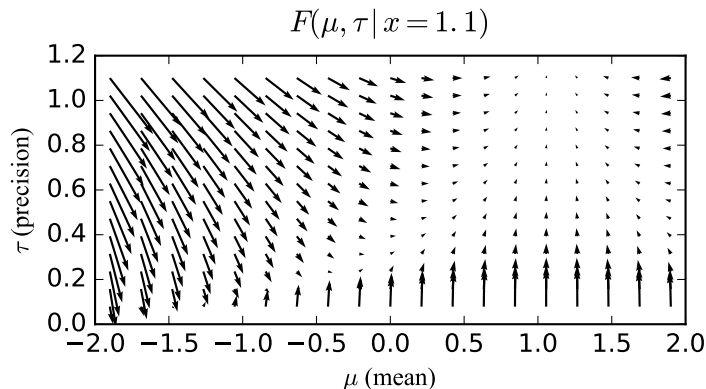
# The Fisher Score

$$F(\theta | x) = \frac{\nabla_{\theta} p(x | \theta)}{p(x | \theta)} = \nabla_{\theta} \log p(x | \theta)$$



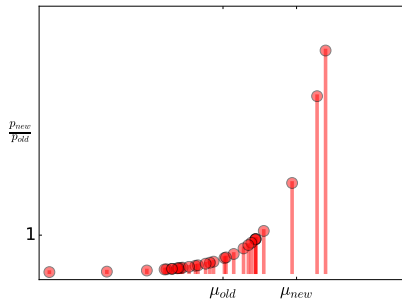
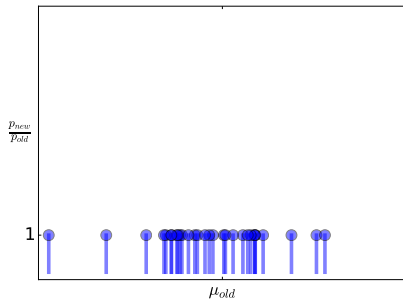
## The Fisher Score

$$\nabla_{(\mu, \tau)} \log \left( \sqrt{\frac{\tau}{\pi}} \exp \left\{ -\tau(x - \mu)^2 \right\} \right) = \begin{pmatrix} 2\tau(x - \mu) \\ (2\tau)^{-1} - (x - \mu)^2 \end{pmatrix}$$

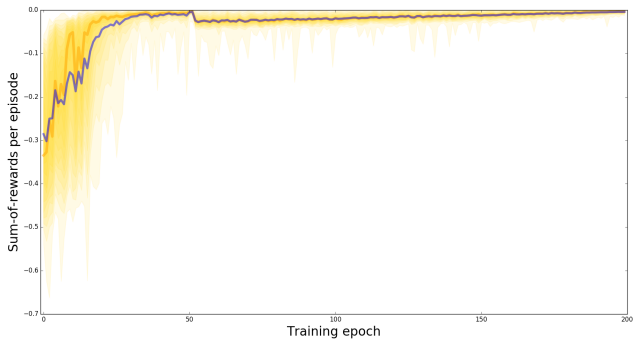


# The Policy Gradient

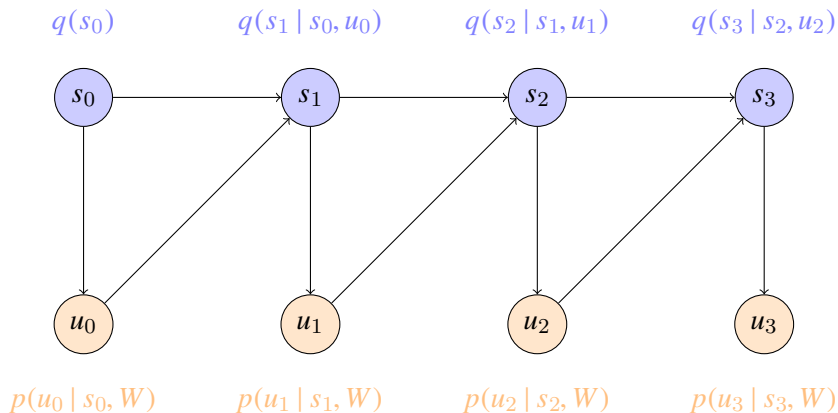
$$\nabla_W E_W [R] = E_{W_0} [R \cdot F]$$



Dart-Throwing Game:  $R(u) = -\|u - u^*\|^2$



## Scores Given Rollouts



## Scores Given Rollouts

$$\frac{\nabla (q_0 p_1 q_1 p_1 q_2 p_2 \cdots q_{T-1} p_{T-1})}{(q_0 p_1 q_1 p_1 q_2 p_2 \cdots q_{T-1} p_{T-1})} = \frac{\nabla (p_1 p_1 p_2 \cdots p_{T-1})}{(p_1 p_1 p_2 \cdots p_{T-1})}$$

Hence:

$$F = \nabla \log p_0 + \nabla \log p_1 + \nabla \log p_2 + \cdots + \nabla \log p_{T-1}$$

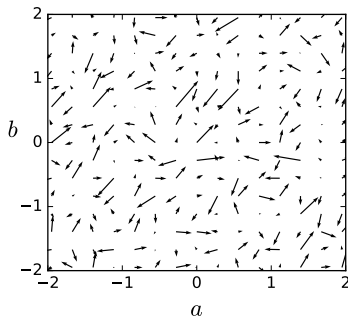
# Repeat-After-Me Game: $R(s, u) = -\|s - u\|^2$

$$s \sim \mathcal{N}(1/2, 1)$$

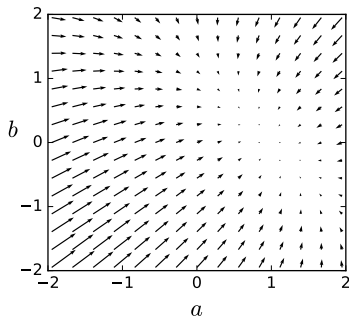
$$u \sim \mathcal{N}(as + b, 1)$$

$$F\left(\begin{matrix} a \\ b \end{matrix} \middle| s, u\right) = \begin{pmatrix} (as + b - u)s \\ (as + b - u) \end{pmatrix}$$

Score estimates



Gradient estimates





# Repeat-After-Me Game: $R(s, u) = -\|s - u\|^2$

