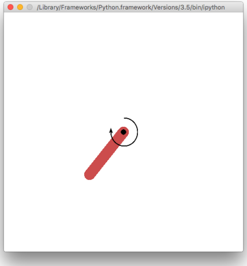
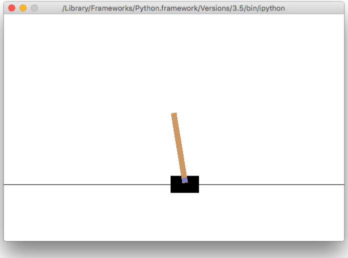
Model-Free Reinforcement Learning

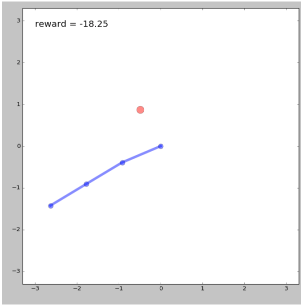
Mathias Winther Madsen

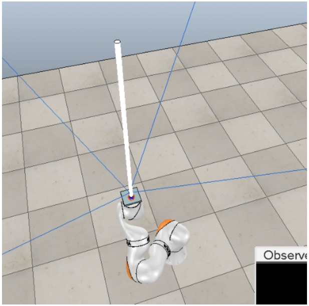
March 1, 2017

Policy Gradient Methods

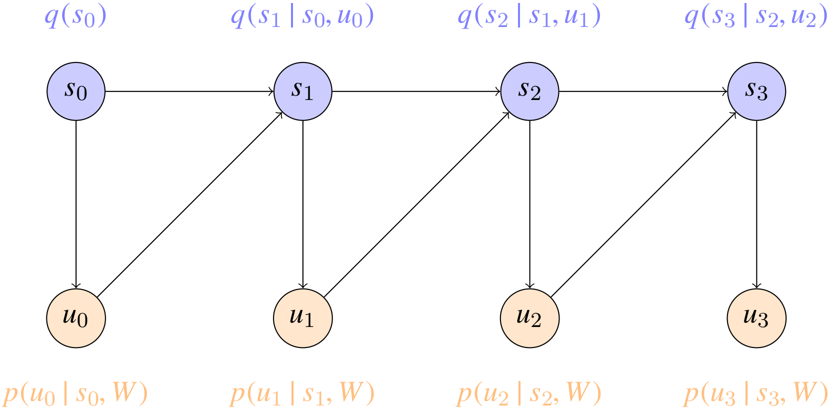

/Library/Frameworks/Python.framework/Versions/3.5/bin/ipython


/Library/Frameworks/Python.framework/Versions/3.5/bin/ipython

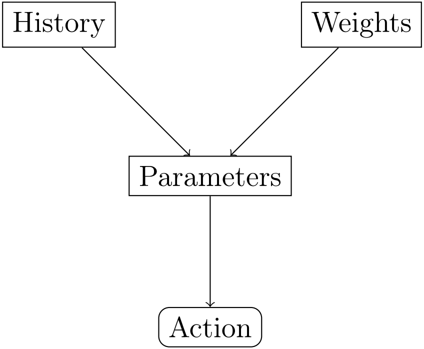




Reinforcement Learning = Game Theory

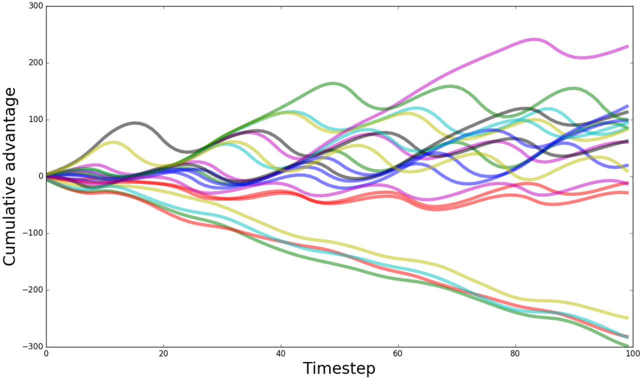


Reinforcement Learning = Game Theory



Action ~ distribution(history, weights)

The Policy Gradient Method

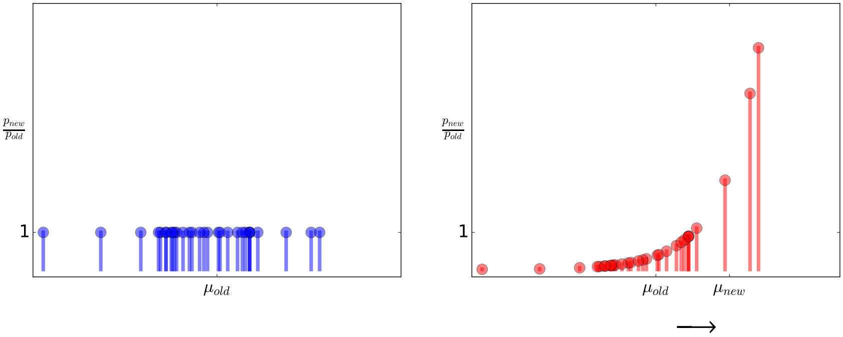


The Policy Gradient Method

```
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

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

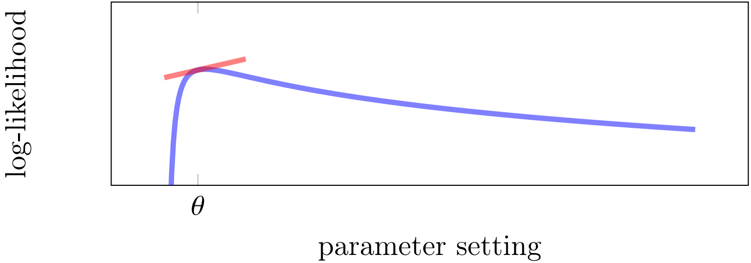


The Policy Gradient

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

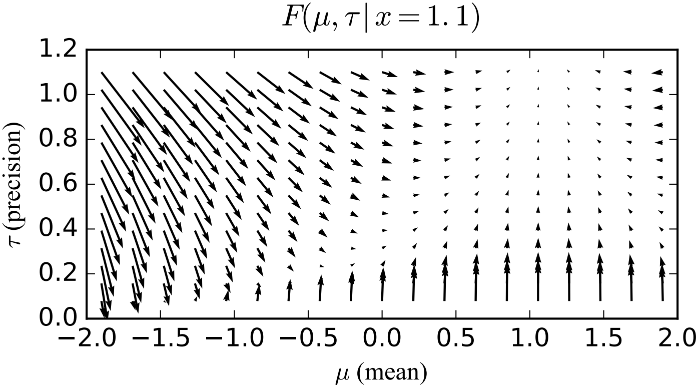
The Fisher Score

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



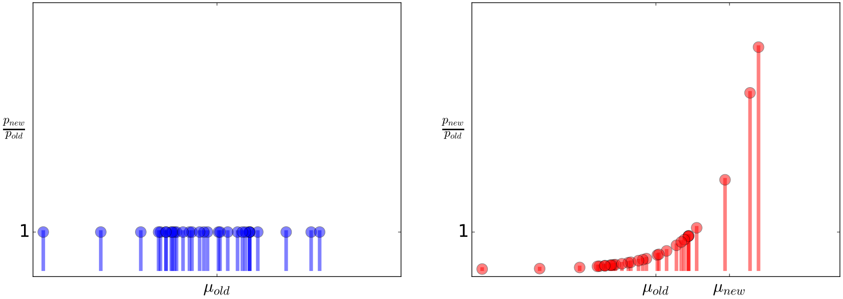
The Fisher Score

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

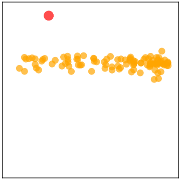


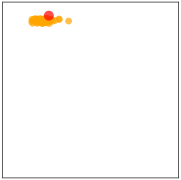
The Policy Gradient

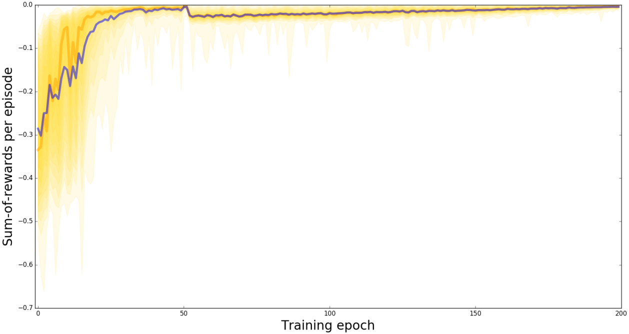
$$\nabla_W \, E_W \left[\, R \, \right] \quad = \quad E_{W_0} \left[\, R \cdot F \, \right]$$



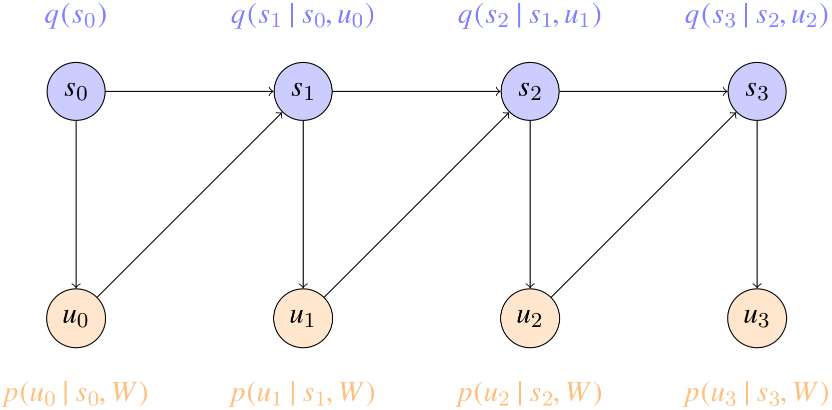
Dart-Throwing Game: $R(u) = -\|u - u^*\|^2$







Scores Given Rollouts



Scores Given Rollouts

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

Hence:

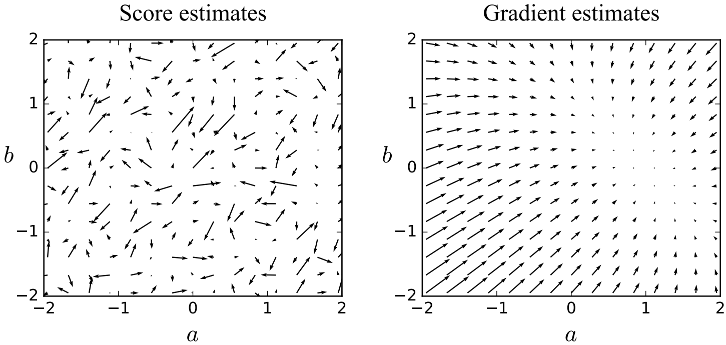
$$\boxed{F \quad = \quad \nabla \log p_0 + \nabla \log p_1 + \nabla \log p_2 + \cdots + \nabla \log p_{T-1}}$$

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

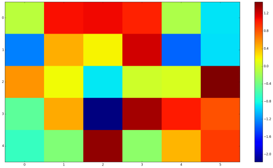
$$s \sim \mathcal{N}(1/2, 1)$$

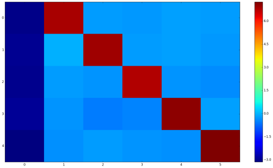
$$u \sim \mathcal{N}(as+b, 1)$$

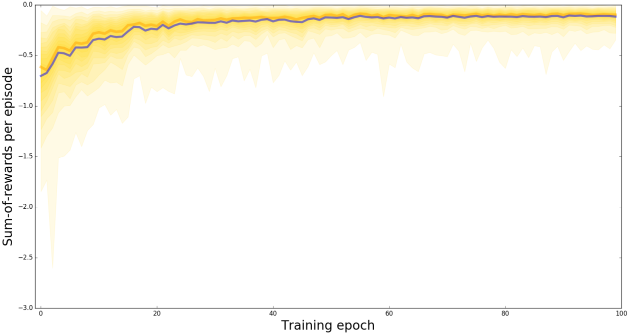
$$F\left(\begin{array}{c} a \\ b \end{array} \middle| s,u\right) = \left(\begin{array}{c} (as+b-u)s \\ (as+b-u) \end{array}\right)$$



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


0
1
2
3
4
0
1
2
3
4
5
1.2
0.8
0.4
0.0
−0.4
−0.8
−1.2
−1.6
−2.0


0
1
2
3
4
0
1
2
3
4
5
6.0
4.5
3.0
1.5
0.0
−1.5
−3.0


Sum-of-rewards per episode
Training epoch