REINFORCEMENT LEARNING: THEORY AND PRACTICE

Ch. 2: Gradient Bandits

Profs. Amy Zhang and Peter Stone

TEXAS
The University of Texas at Austin

Reading Response Questions

Sam Ziegelbein: "Why do we use the softmax distribution in gradient bandit algorithms, as opposed to some other method of converting preference values into probabilities? The exponentials sure seem nice to work with, but intuitively why is this distribution better than some other one?"

Gradient Bandits: Arm Preferences

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

Gradient Bandits: Arm Preferences

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

Differentiable

Gradient Bandits: Arm Preferences

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

Differentiable

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$

Updates can be high variance

Exact Gradient Ascent

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

Where expected reward is $$\mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x)$$

Exact Performance Gradient

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right]$$

$$= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x \big(q_*(x) - B_t \big) \frac{\partial \pi_t(x)}{\partial H_t(a)},$$

q doesn't depend on H

B can be an arbitrary scalar as long as it doesn't depend on x

Gradient Bandits: Baseline

How does expected return change w.r.t. prefs?

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right]$$

$$= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x \big(q_*(x) - B_t \big) \frac{\partial \pi_t(x)}{\partial H_t(a)},$$

Gradient Bandits: Baseline

How does expected return change w.r.t. prefs?

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right]$$

$$= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x \left(q_*(x) - B_t \right) \frac{\partial \pi_t(x)}{\partial H_t(a)},$$

Sum over Actions

How good is action?

How does policy change w.r.t. prefs?

Why are we allowed to subtract a baseline?

How does expected return change w.r.t. prefs?

$$
\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] \\
&= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \sum_x \left(q_*(x) - \boxed{B_t} \right) \frac{\partial \pi_t(x)}{\partial H_t(a)},
\end{aligned}
$$

Expected baseline contribution = 0 because…

…multiplied by term with expectation 0

Claim: a good baseline reduces variance of gradient and improves convergence

Why does the variance of the gradient matter?

Why does the variance of the gradient matter?

Theory: upper bounds on convergence rate of SGD are directly related to the variance of gradient estimates

Intuition: variance causes "overshooting" that destabilizes learning

Why does the variance of the gradient matter?

Theory: upper bounds on convergence rate of SGD are directly related to the variance of gradient estimates

Intuition: variance causes "overshooting" that destabilizes learning

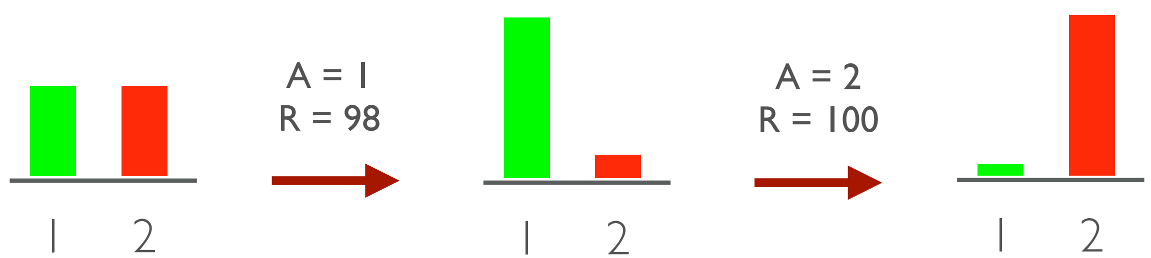
$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$

Why does the variance of the gradient matter?

Theory: upper bounds on convergence rate of SGD are directly related to the variance of gradient estimates

Intuition: variance causes "overshooting" that destabilizes learning

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$

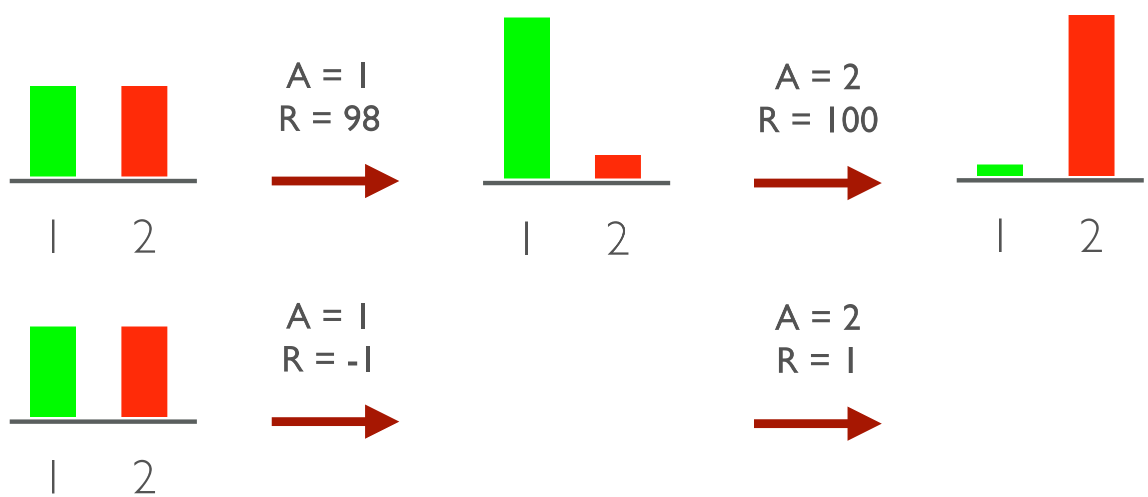


Why does the variance of the gradient matter?

Theory: upper bounds on convergence rate of SGD are directly related to the variance of gradient estimates

Intuition: variance causes "overshooting" that destabilizes learning

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$

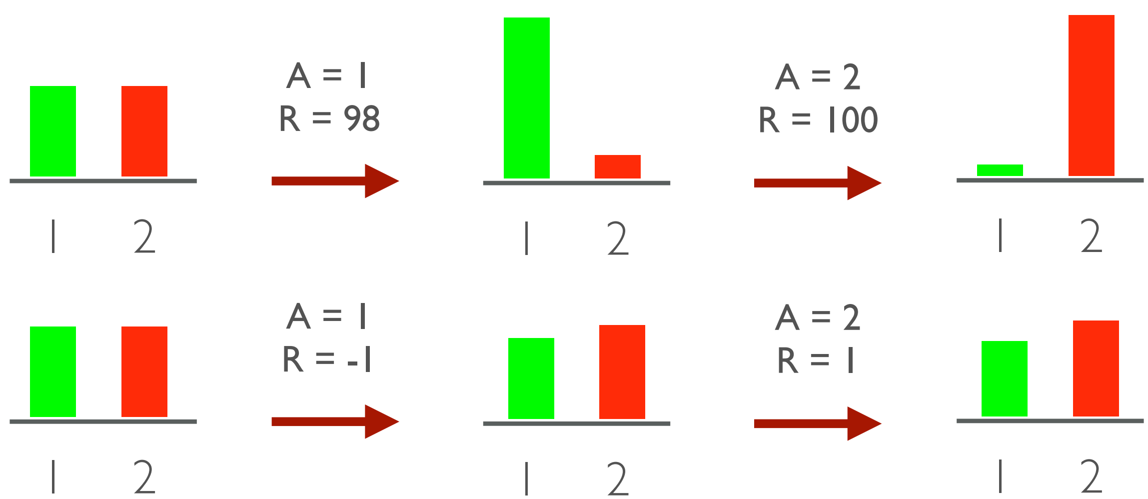


Why does the variance of the gradient matter?

Theory: upper bounds on convergence rate of SGD are directly related to the variance of gradient estimates

Intuition: variance causes "overshooting" that destabilizes learning

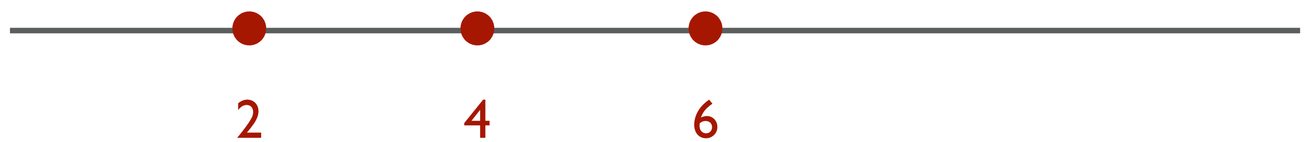
$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$



Why does baseline reduce variance?

Why does baseline reduce variance?

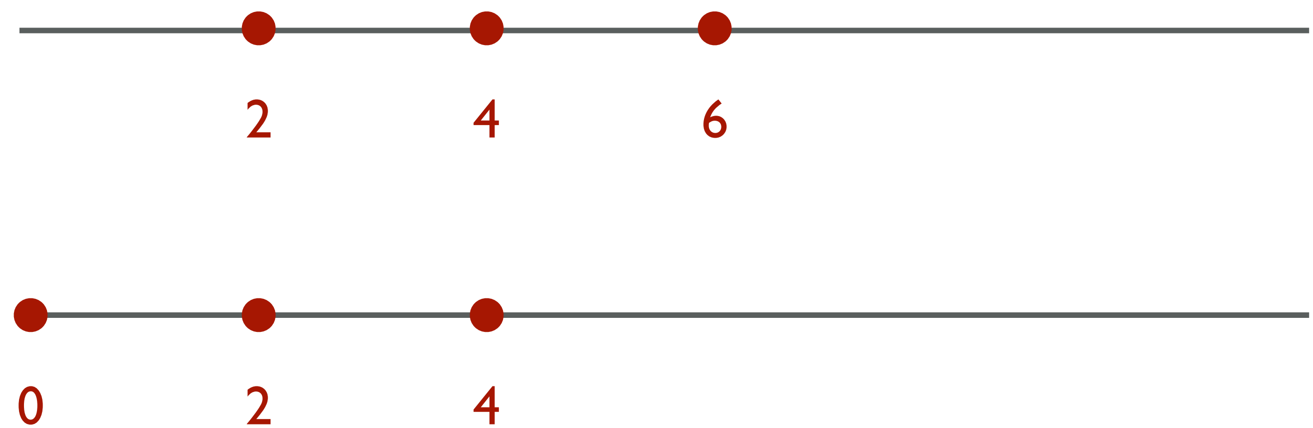
Original “gradients”



Why does baseline reduce variance?

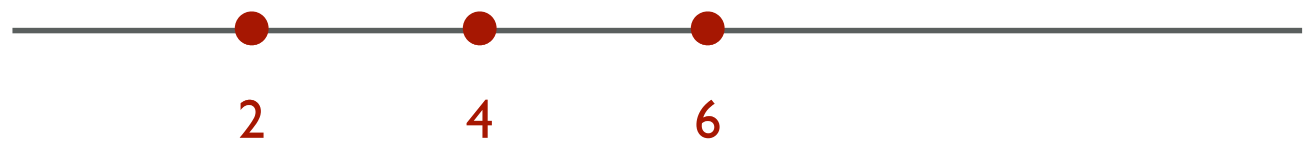
Original “gradients”

We are NOT subtracting from the gradient



Why does baseline reduce variance?

Original "gradients"



We are NOT subtracting
from the gradient

We are subtracting
from a number that
multiplies the gradient

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \left(q_*(x) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$