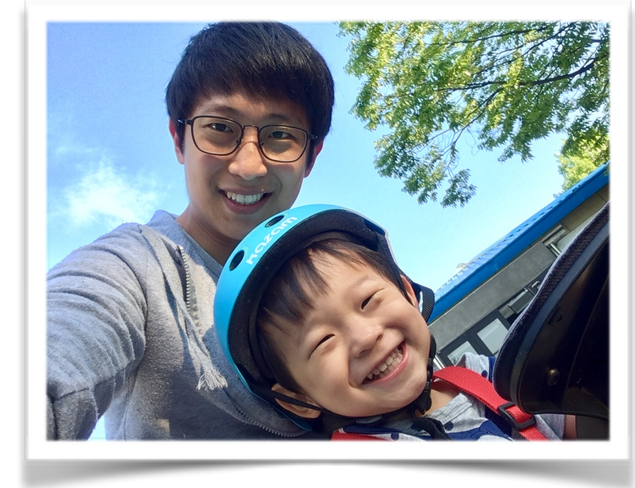
Policy Gradient
to
Actor-Critic

박진우 (Curt Park)

RL Korea Bootcamp
2019. 10. 27

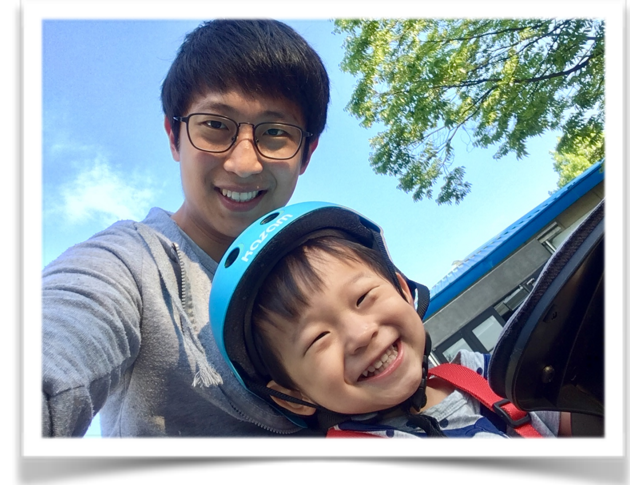
발표자 소개

걸어온 길



- 컴퓨터공학 학사 (2006 - 2014)
- SW developer at Smilegate (2013.11 - 2014.5)
- SW developer at Ericsson (2014.10 - 2017.01)
- Research Engineer at Medipixel (2018.11 - 2019.08)
- Research Engineer at J.Marple (2019.09 -)

최근 활동



- 연구주제
  - 모델 기반 동적 제어 시스템
  - 강화학습을 이용한 심혈관 시술용 가이드와이어 제어
- 기타활동
  - PG is all you need with 김경환, 김민철
  - Rainbow is all you need with 김경환
  - RL algorithms with 김경환, 김민철
  - 모두를 위한 컨벡스 최적화 @풀잎스쿨, 모두의 연구소

이 발표는?

- **범위**
  - Policy Gradient에서 A2C에 이르는 이론적 배경
  - 실습을 통한 A2C 구현의 이해 (Pytorch)
- **목표**
  - Pendulum 환경에서의 에이전트 제어

목차

1. Policy Gradient

2. Reinforce

3. Actor Critic (AC)

4. Hands-on Practice: A2C

Policy Gradient

Q. 강화학습?

문제를 풀기 위한
최적의 정책을 찾아내는 것

$$\pi^\star(a \mid s) \approx \pi(a \mid s; \theta)$$

DQN에서의 최적 정책

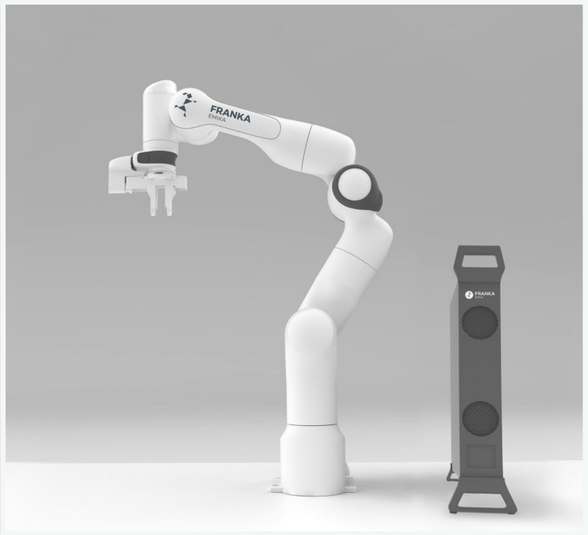
$$\pi^\star(a \mid s) = \arg\max_a Q^\star(s, a) \approx arg\max_a Q(s, a \mid \theta)$$

하지만…

상상해봅시다

- Partially observable state만을 획득할 수 있다면?
  - e.g. 포커게임

- Infinite action space를 다뤄야한다면?
  - e.g. 로봇팔 제어

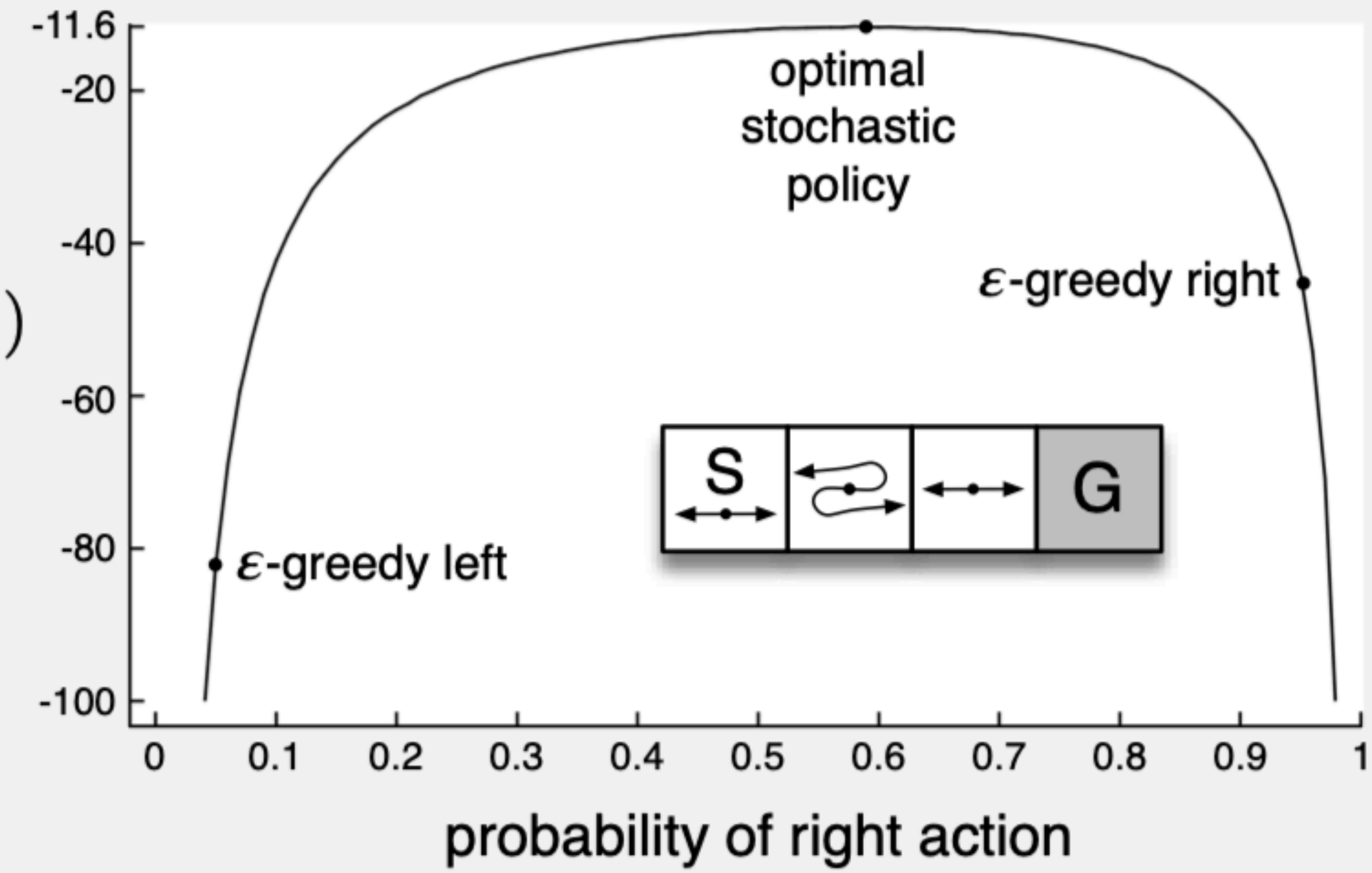


https://kr.pokernews.com/news/2013/09/hand-motions-to-catch-bluffs-kr-5709.htm
https://etinow.me/143

Short corridor with switched actions

$$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(\mathsf{S})$$

$$\epsilon = 0.1$$



- 입력되는 모든 state가 동일
- 가운데 state에서는 action의 효과가 뒤바뀜
- 최적의 policy는 좌 : 우를 0.41 : 0.59 비율로 선택하는 것

Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

Q. 강화학습?

A. 문제를 풀기 위한
최적의 정책을 찾아내는 것

적합한 목적함수의 정의!

목적함수 정의

$$\underbrace{p_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)}_{p_\theta(\tau)} = p(\mathbf{s}_1) \prod_{t=1}^{T} \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\theta^\star = \arg\max_\theta E_{\tau \sim p_\theta(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$\theta^\star = \arg\max_\theta E_{(\mathbf{s},\mathbf{a}) \sim p_\theta(\mathbf{s},\mathbf{a})} [r(\mathbf{s}, \mathbf{a})]$$

infinite horizon case

$$\theta^\star = \arg\max_\theta \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t) \sim p_\theta(\mathbf{s}_t,\mathbf{a}_t)} [r(\mathbf{s}_t, \mathbf{a}_t)]$$

finite horizon case

CS294-112 at UC Berkeley (2018). Lecture5: Policy Gradient Introduction

목적함수 정의

$$\theta^\star = \arg\max_\theta \underbrace{E_{\tau \sim p_\theta(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right]}_{J(\theta)}$$

$$J(\theta) = E_{\tau \sim p_\theta(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right] \approx \frac{1}{N} \sum_i \sum_t r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$$

N samples

CS294-112 at UC Berkeley (2018). Lecture5: Policy Gradient Introduction

Policy Gradient

$$\theta^\star = \arg\max_\theta \underbrace{E_{\tau\sim p_\theta(\tau)}\left[\sum_t r(\mathbf{s}_t,\mathbf{a}_t)\right]}_{J(\theta)}$$

a convenient identity

$$\underline{\pi_\theta(\tau)\nabla_\theta \log \pi_\theta(\tau)} = \pi_\theta(\tau)\frac{\nabla_\theta \pi_\theta(\tau)}{\pi_\theta(\tau)} = \underline{\nabla_\theta \pi_\theta(\tau)}$$

$$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[\underbrace{r(\tau)}_{\sum_{t=1}^{T} r(\mathbf{s}_t,\mathbf{a}_t)}] = \int \pi_\theta(\tau) r(\tau) d\tau$$

$$\nabla_\theta J(\theta) = \int \underline{\nabla_\theta \pi_\theta(\tau)} r(\tau) d\tau = \int \underline{\pi_\theta(\tau)\nabla_\theta \log \pi_\theta(\tau)} r(\tau) d\tau = E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]$$

Policy Gradient

- Finite Horizon Case

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}\Big[\Big(\sum_{t=1}^{T} \nabla_\theta \log \pi(a_t\,|\,s_t)\Big)\Big(\sum_{t=1}^{T} r(s_t, a_t)\Big)\Big]$$

$$\approx \frac{1}{N}\sum_{i}^{N}\Big(\sum_{t=1}^{T} \nabla_\theta \log \pi(a_t\,|\,s_t)\Big)\Big(\sum_{t=1}^{T} r(s_t, a_t)\Big)$$

N samples

Policy Gradient

- Markov Property

t < t' 일때, t'에서의 policy가 t 시점에 영향을 끼치지 않는다고 가정

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i}^{N}\Big(\sum_{t=1}^{T}\nabla_\theta \log \pi(a_t \mid s_t)\Big)\Big(\sum_{t=1}^{T} r(s_t, a_t)\Big)$$

$$= \frac{1}{N}\sum_{i}^{N}\Big(\sum_{t=1}^{T}\nabla_\theta \log \pi(a_t \mid s_t)\Big)\Big(\sum_{t=t'}^{T} r(s_{t'}, a_{t'})\Big)$$

$$= \frac{1}{N}\sum_{i}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi(a_t \mid s_t)\cdot G_t$$

Policy Gradient

- Policy Gradient

t < t' 일때, t'에서의 policy가 t 시점에 영향을 끼치지 않는다고 가정

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi(a_t \mid s_t) \cdot G_t$$

$$\underline{\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)}$$

Gradient Ascent!

직관적 해석:

Return이 곱해진Maximum log likelihood의 형태

- Action probability가 높을수록 업데이트에 덜 반영
- Return이 높을수록 업데이트에 더 반영

Policy Gradient

- Reinforce

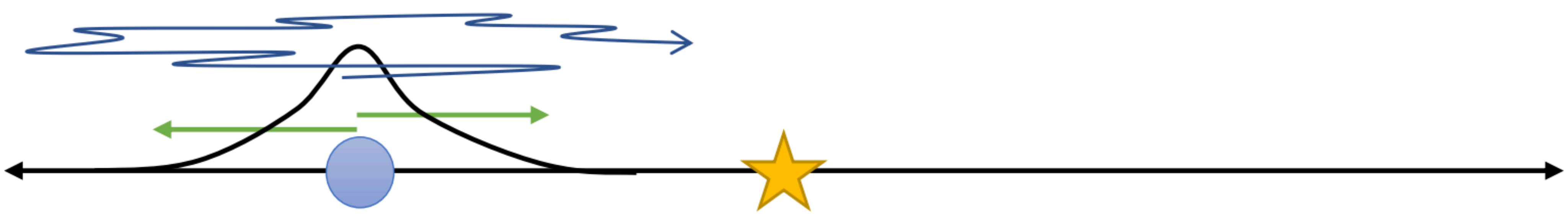
REINFORCE, A Monte-Carlo Policy-Gradient Method (episodic)

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$
Repeat forever:
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  For each step of the episode $t = 0, \ldots, T-1$:
    $G \leftarrow$ return from step t
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla_{\boldsymbol{\theta}} \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

Policy Gradient

- Reinforce: High-variance issue



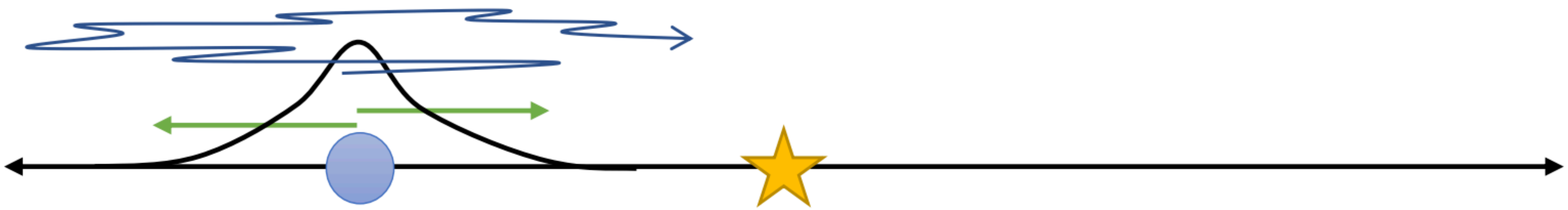
Sample trajectory에 대한 과적합으로 학습을 원활하게 하기 어려움

Variance-Bias trade-off가 필요

CS294-112 at UC Berkeley (2018). Lecture5: Policy Gradient Introduction

Policy Gradient

- Reinforce: High-variance issue



Sample trajectory에 대한 과적합으로 학습을 원활하게 하기 어려움

Variance-Bias trade-off가 필요

적절한 Baseline 함수를 이용!

CS294-112 at UC Berkeley (2018). Lecture5: Policy Gradient Introduction

Policy Gradient

- Reinforce with Baseline

$$J(\theta) \approx \frac{1}{N} \sum_{i}^{N} \sum_{t=1}^{T} \log \pi(a_t \mid s_t; \theta) \cdot G_t$$

$$\rightarrow \frac{1}{N} \sum_{i}^{N} \sum_{t=1}^{T} \log \pi(a_t \mid s_t; \theta) \cdot (G_t - \underline{v(s; w)})$$

theta에 대한 미분과 함께 소거가능

Policy Gradient

- Reinforce with Baseline

REINFORCE with Baseline (episodic)

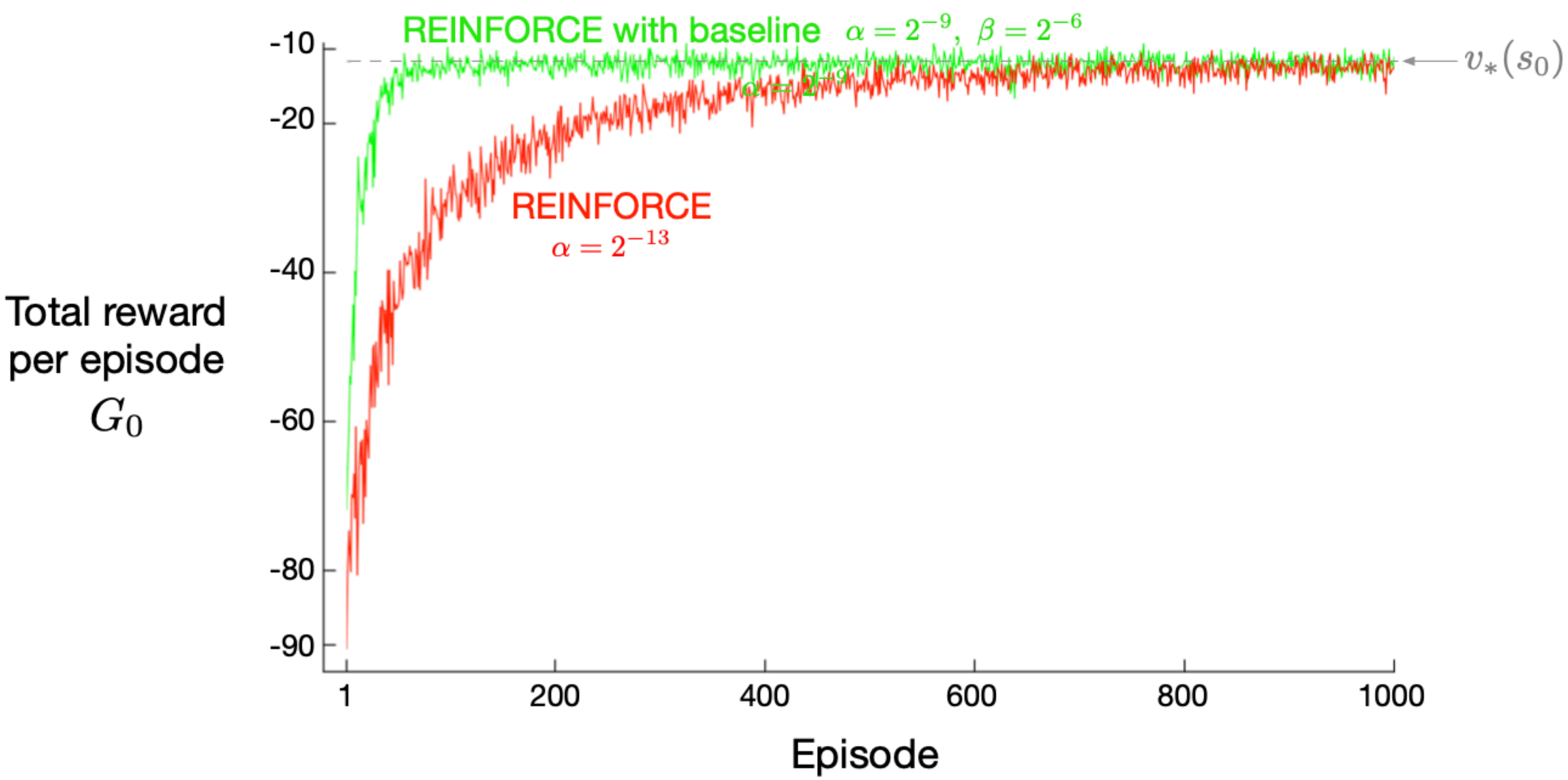
Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$

Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$
Repeat forever:
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  For each step of the episode $t = 0, \ldots, T-1$:
    $G_t \leftarrow$ return from step t
    $\delta \leftarrow G_t - \hat{v}(S_t,\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \gamma^t \delta \nabla_{\mathbf{w}} \hat{v}(S_t,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

Policy Gradient

- Short corridor with switched actions



Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

Policy Gradient

- Reinforce with Baseline

REINFORCE with Baseline (episodic)

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$

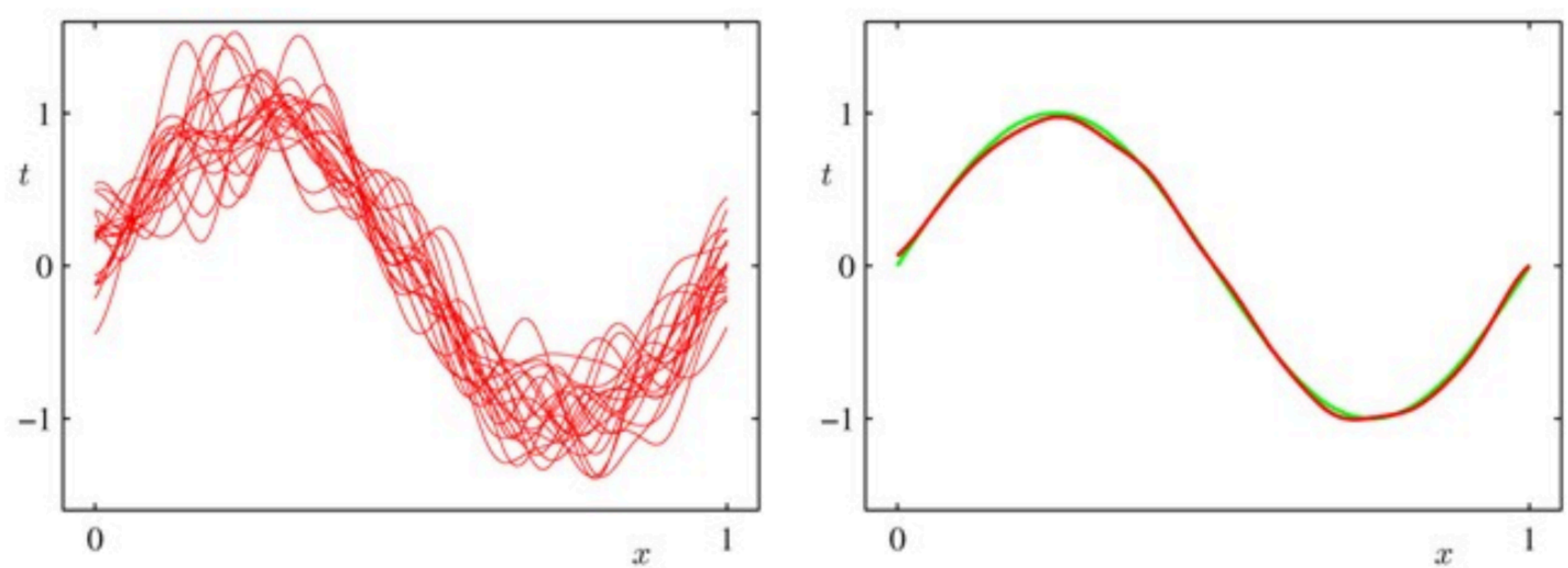
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$
Repeat forever:
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  For each step of the episode $t = 0, \ldots, T-1$:
    $G_t \leftarrow$ return from step t
    $\delta \leftarrow G_t - \hat{v}(S_t,\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \gamma^t \delta \nabla_{\mathbf{w}} \hat{v}(S_t,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla_{\boldsymbol{\theta}} \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Return 대신 Q(s,a)를 사용한다면?

Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

Policy Gradient

- Variance reduction by expectation



Q(S, A)를 G에 대한 기댓값으로 볼 수 있음

Policy Gradient

- Actor-Critic

One-step Actor–Critic (episodic)

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$

Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$
Repeat forever:
  Initialize S (first state of episode)
  $I \leftarrow 1$
  While S is not terminal:
    $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} I \delta \nabla_{\mathbf{w}} \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla_{\boldsymbol{\theta}} \ln \pi(A|S, \boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

Note: Value function을 bootstrapping!

Sutton, R. and Barto, A. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press

소화하는 시간

실습시간

https://github.com/MrSyee/pg-is-all-you-need

소화하는 시간

감사합니다!