강화학습개론 (AIE5101)

Term Project Report

팀명: **강화성공**

120240534 김현정 (팀장)
320250053 김문일 (팀원)
120250251 민경현 (팀원)

Github link

https://github.com/ghmin-cnl/25_2_RL_project

목차

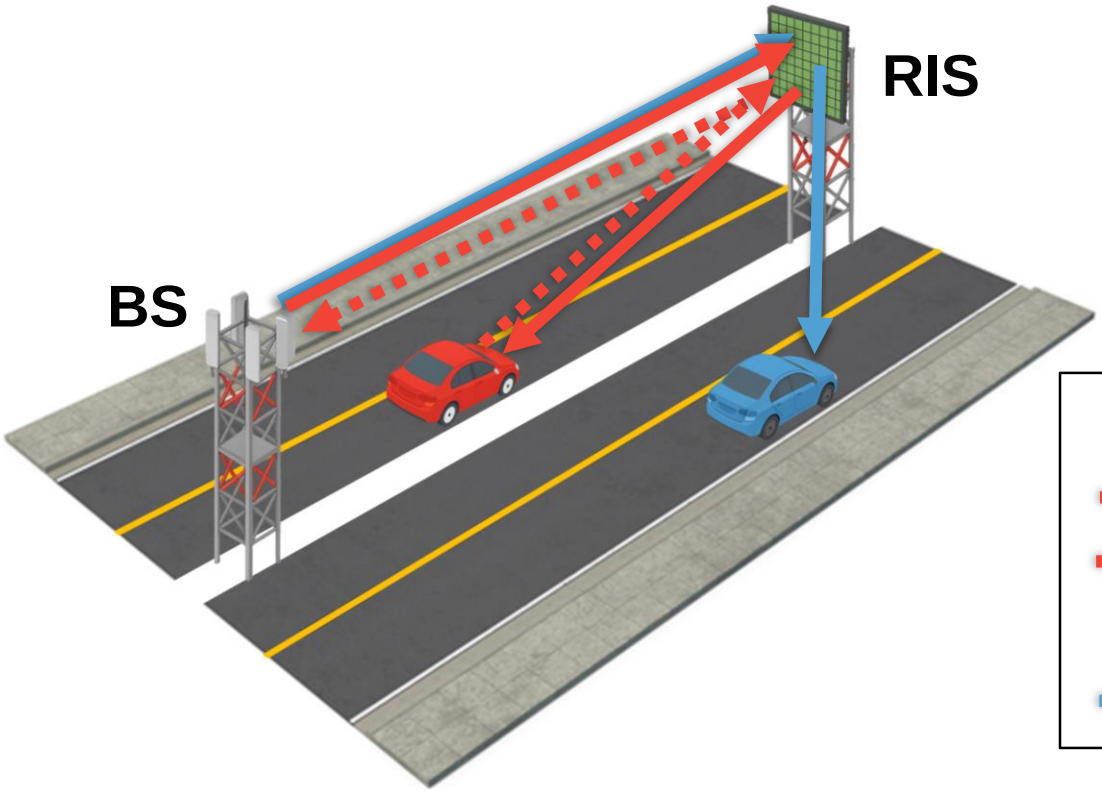
프로젝트 주제 및 목표

- 프로젝트 제목
  - Reinforcement Learning-Based Joint Beamforming Design for RIS-ISAC Vehicular Network with Delayed Channels
  - 지연된 채널을 가지는 RIS-ISAC 차량 네트워크에서 강화학습 기반 공동 빔포밍 설계

- RIS(Reconfigurable Intelligent Surface)의 지원으로 타겟 센싱과 사용자 통신을 동시에 수행하는 ISAC(Integrated Sensing and Communication) 차량 통신 네트워크를 다룸
- 각 RE(RIS Element)의 위상 변화(phase shift)로 전파 경로 조정 가능 → 에코 신호(sensing)와 사용자 수신 신호(communication) 동시에 향상 가능
- RL을 이용하여, communication SNR(Signal-to-Noise Ratio)을 보장하면서 sensing SNR을 극대화하는 RIS 위상 변화 도출 → 이후 RIS 위상 변화 기반으로 기지국의 최적 송신 빔포머 계산

시스템 모델 (1/3)

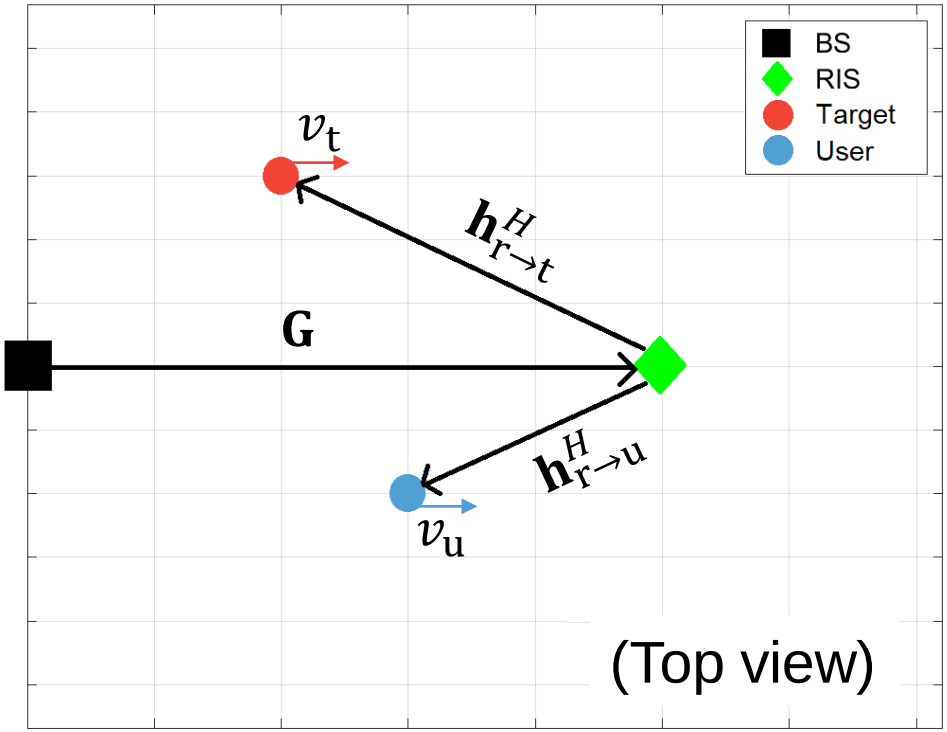
- 센싱 타겟과 통신 대상이 이동 중인 차량인 **도심 RIS-ISAC 차량 네트워크**
  - 차량 이동성과 C-RAN 제어·연산/Edge-offloading 처리 등의 요인으로 **RIS가 받는 CSI(Channel State Information) 관측이 지연**되는 환경
  - 기지국(BS; Base Station)은 RIS로 다운링크 센싱/통신 신호 송신
  - RIS는 위상을 조정하여 타겟/사용자에게 신호 전달
    - 이때 센싱 타겟의 에코 신호는 다시 RIS → BS 경로로 전달



: 센싱 타겟(차량)
: 다운링크 센싱 신호
: 에코 신호
: 통신 사용자(차량)
: 다운링크 통신 신호

시스템 모델 (2/3)

기지국 안테나 M개	RIS element N개	타겟/사용자 단일 안테나
기지국 → RIS 채널: $\mathbf{G} \in \mathbb{C}^{N \times M}$	기지국 송신 빔포머: $\mathbf{w} \in \mathbb{C}^{M \times 1}$	
RIS → 타겟 채널: $\mathbf{h}_{\mathrm{r}\to\mathrm{t}}^{H} \in \mathbb{C}^{1 \times N}$	센싱 잡음: $n_{\mathrm{s}} \sim \mathcal{CN}(0, \sigma_{\mathrm{s}}^2)$	
RIS → 사용자 채널: $\mathbf{h}_{\mathrm{r}\to\mathrm{u}}^{H} \in \mathbb{C}^{1 \times N}$	통신 잡음: $n_{\mathrm{c}} \sim \mathcal{CN}(0, \sigma_{\mathrm{c}}^2)$	
RIS 위상 변화 행렬: $\boldsymbol{\Theta} = \mathrm{diag}(\mathrm{e}^{j\hat{\theta}_1}, \mathrm{e}^{j\hat{\theta}_2}, \ldots, \mathrm{e}^{j\hat{\theta}_N}) \in \mathbb{C}^{N \times N}$ ($\hat{\theta}_n$은 phase shift)		

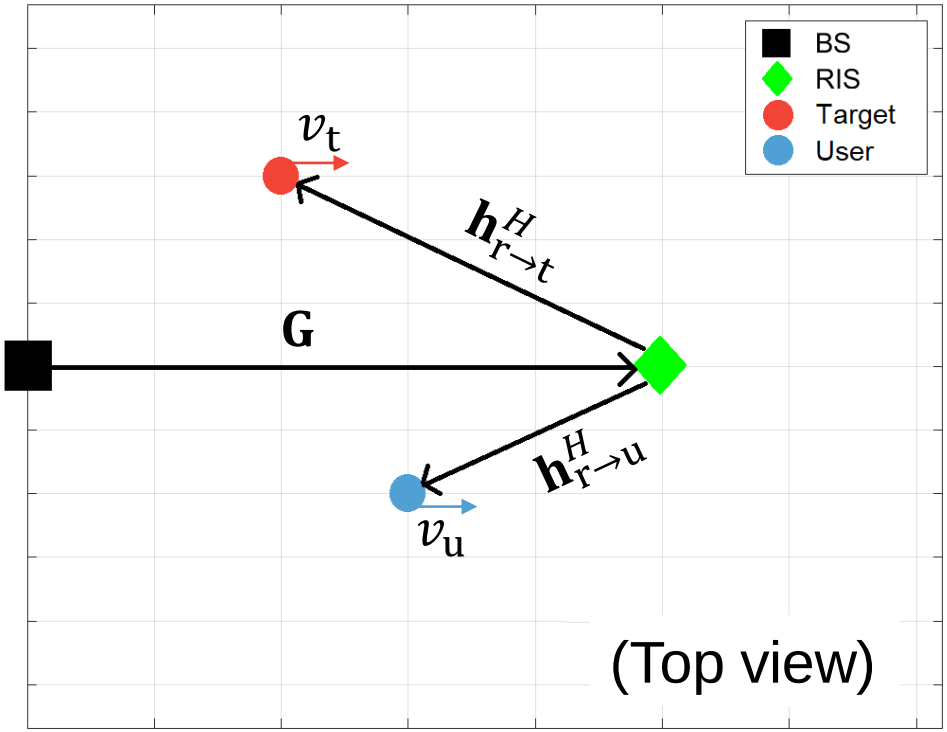


타겟 속도: v_{t}

사용자 속도: v_{u}

시스템 모델 (3/3)

- 통신 사용자의 수신 신호: $y_c = \mathbf{h}_{\mathrm{r}\to\mathrm{u}}^H \boldsymbol{\Theta}^H \mathbf{G}\mathbf{w}s + n_{\mathrm{c}}$ (s는 power-normalized signal)
- 통신 사용자의 수신 SNR(dB): $\gamma_c = 10\log_{10}\left(\frac{\left|\mathbf{h}_{\mathrm{c}}^H \mathbf{w}\right|^2}{\sigma_{\mathrm{c}}^2}\right)$, where $\mathbf{h}_{\mathrm{c}}^H = \mathbf{h}_{\mathrm{r}\to\mathrm{u}}^H \boldsymbol{\Theta}^H \mathbf{G}$
- 기지국의 센싱 에코 수신 신호: $y_s = \mathbf{G}^H \boldsymbol{\Theta}\mathbf{h}_{\mathrm{r}\to\mathrm{t}}\mathbf{h}_{\mathrm{r}\to\mathrm{t}}^H \boldsymbol{\Theta}^H \mathbf{G}\mathbf{w}s + n_{\mathrm{s}}$
- 기지국의 에코 수신 SNR(dB): $\gamma_s = 10\log_{10}\left(\frac{\|\mathbf{H}_{\mathrm{t}}\mathbf{w}\|^2}{\sigma_s^2}\right)$, where $\mathbf{H}_{\mathrm{t}} = \mathbf{h}_{\mathrm{t}}\mathbf{h}_{\mathrm{t}}^H$, $\mathbf{h}_{\mathrm{t}}^H = \mathbf{h}_{\mathrm{r}\to\mathrm{t}}^H \boldsymbol{\Theta}^H \mathbf{G}$



문제 정식화

- 최적화 목적: 기지국의 에코 수신 SNR, 즉 센싱 SNR γ_s를 최대화

$$\max_{\mathbf{w},\boldsymbol{\Theta}} \gamma_s$$

s.t. $\gamma_c \geq \tau_c$, ----------→ 통신 SNR γ_c를 threshold τ_c 이상 보장

$\|\mathbf{w}\|^2 \leq P_{\text{tx}}^{\max}$, ----------→ 기지국 최대 송신 전력 $P_{\text{tx}}^{\max}$ 제한

$\left|\boldsymbol{\Theta}_{n,n}\right| = 1, \forall n = \{1, 2, \ldots, N\}$ ----------→ RIS 반사는 크기 변화 없이 위상만 변화

- $\boldsymbol{\Theta}$가 정해지면, 상기 제약을 만족하는 조건 하에서 최적 $\mathbf{w}$를 closed-form으로 계산 가능함[1] → RL을 이용하여 $\boldsymbol{\Theta}$ 결정하고 $\mathbf{w}$ 계산 → SNR 도출

Theorem 1: The optimal transmit beamformer $\mathbf{w}$ is

$$\mathbf{w} = \begin{cases} \sqrt{P_t}\frac{\mathbf{h}_t}{\|\mathbf{h}_t\|}, & \text{if } P_t|\mathbf{h}_c^H\mathbf{h}_t|^2 \geq \tau_c\sigma_c^2\|\mathbf{h}_t\|^2, \\ \mathbf{x}_1\mathbf{u}_1 + \mathbf{x}_2\mathbf{u}_2, & \text{otherwise,} \end{cases}$$

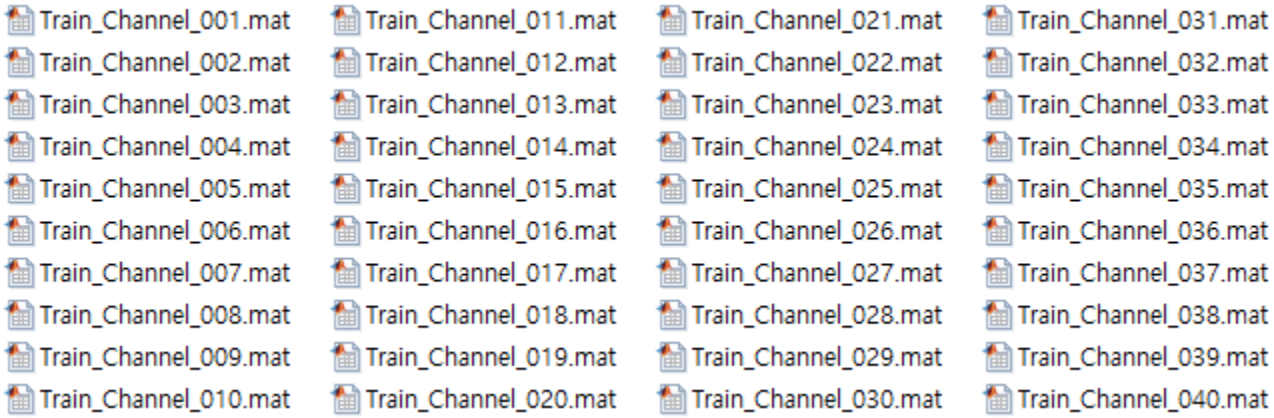
where

$$\mathbf{u}_1 = \frac{\mathbf{h}_c}{\|\mathbf{h}_c\|}, \quad \mathbf{u}_2 = \frac{\mathbf{h}_t - (\mathbf{u}_1^H\mathbf{h}_t)\mathbf{u}_1}{\|\mathbf{h}_t - (\mathbf{u}_1^H\mathbf{h}_t)\mathbf{u}_1\|}, \text{ and } \quad \mathbf{x}_1 = \sqrt{\frac{\tau_c\sigma_c^2}{\|\mathbf{h}_c\|^2}}\frac{\mathbf{u}_1^H\mathbf{h}_t}{|\mathbf{u}_1^H\mathbf{h}_t|}, \quad \mathbf{x}_2 = \sqrt{P_t - \frac{\tau_c\sigma_c^2}{\|\mathbf{h}_c\|^2}}\frac{\mathbf{u}_2^H\mathbf{h}_t}{|\mathbf{u}_2^H\mathbf{h}_t|}.$$

[1] F. Liu, Y. Liu, A. Li, C. Masouros, and Y. C. Eldar, "Cramér-Rao bound optimization for joint radar-communication beamforming," *IEEE Trans. Signal Process.*, vol. 70, pp. 240–253, Jan. 2022.

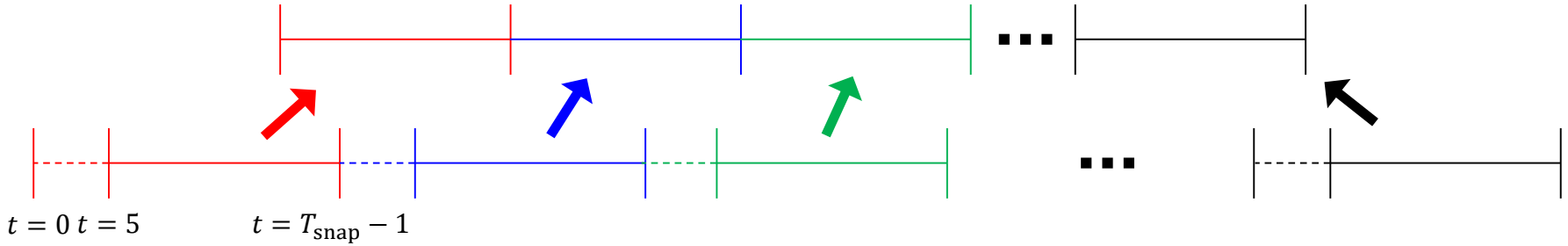
채널 데이터셋 생성 및 전처리 (1/2)

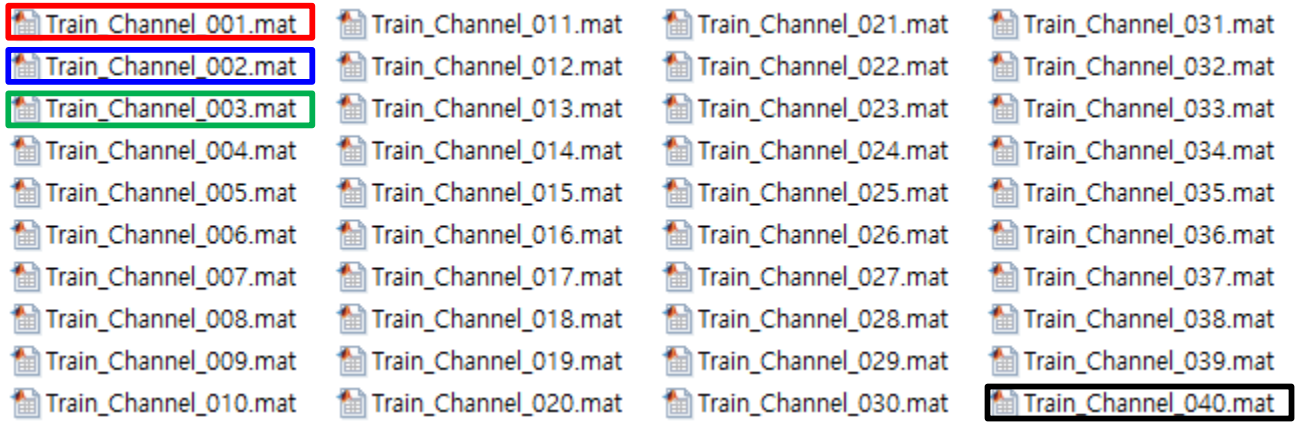
- MATLAB으로 Jakes model 시뮬레이션 및 채널 데이터셋 생성
  - Jakes model: 차량 이동으로 전파가 여러 경로로 반사되어 도달 → 시간에 따른 무선 채널 변화를 수학적으로 모델링(Doppler 효과 포함)
  - 시뮬레이션 환경 (40회 독립적으로 채널 생성)
    - Carrier frequency: $f_c = 3.5$ GHz
    - 안테나 구성: ULA(Uniform Linear Array) 기지국 $M = 8$, RIS $N = 32$
    - Sampling period: $T_s = 0.1$ s, # of snapshots $T_{\text{snap}} = 30$
    - 기지국 위치: $(0,0)$ m, RIS 위치: $(+50,0)$ m
    - 초기 타겟/사용자 위치: $(0, \pm 5)$ (m), 타겟/사용자 속도 $v_t, v_u \sim \mathcal{U}[+30, +100]$ km/h
  - BS–RIS–target/user 간 채널 40×30 = 1200개씩 생성



채널 데이터셋 생성 및 전처리 (2/2)

- 각 시뮬레이션에서 $t = 0 \sim 4$ snapshots의 채널 데이터 제외
  - RL의 state에서 지연된 채널 상태를 고려하기 위함
    - 초기 snapshot에서는 과거의 채널 상태를 사용할 수 없으므로
  - 각 시뮬레이션의 $t = 5 \sim (T_{\text{snap}} - 1)$ snapshots을 concatenation
  - RL에서 사용할 BS–RIS–target/user 간 채널 데이터: 40×(30-5) = **1000개**
    - 각 데이터는 snapshot t마다 $\mathbf{G}_t$, $\mathbf{h}_{\mathrm{r}\to\mathrm{t},t}^H$, $\mathbf{h}_{\mathrm{r}\to\mathrm{u},t}^H$를 포함





MDP 설계

- RL을 이용하여 센싱 SNR을 극대화하는 RIS 위상 변화 행렬 $\boldsymbol{\Theta}$을 도출하기 위해 다음과 같이 MDP를 정의한다: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$
  - $\mathcal{S}$는 state space, $s_t \in \mathcal{S}$는 time step t에서의 state
  - $\mathcal{A}$는 action space, $a_t \in \mathcal{S}$는 time step t에서의 action
  - $\mathcal{P}$는 state transition probability
    - 생성한 채널 데이터셋을 state의 일부분으로 사용하고, action에 따라 state의 나머지가 결정되는 deterministic environment로 모델링함
  - $\mathcal{R}$은 reward function, $r_t \in \mathcal{R}$는 time step t에서의 reward
  - γ는 discount factor, $\gamma \in (0,1]$

- 본 프로젝트에서 RIS를 에이전트로 간주
  - 즉, RIS가 기지국 및 센싱/통신 사용자들로부터 보고받는 **지연된 채널 정보**와 **이전 슬롯의 SNR** 값을 기반(state)으로 위상 변화 행렬 $\boldsymbol{\Theta}$을 결정(action)
  - 센싱 SNR 값과 통신 SNR의 최소값 제약 만족 여부를 반영 누적(reward)
  - 이동중인 센싱/통신 사용자들로 인해 시간에 따른 채널 정보 변화(next state)

MDP – State (s_t) design

- 현재 시스템 모델의 두가지 가정
  1) 차량 이동성과 C-RAN 제어·연산/Edge-offloading 처리 등의 요인으로 인한 **RIS가 받는 채널 정보(CSI) 관측 지연**은 5 slots(time step)으로 가정
  2) 센싱/통신 신호의 SNR 값은 RIS의 직전 slot $\boldsymbol{\Theta}$에 의존

- $s_t = \left(\widehat{\boldsymbol{\Xi}}_{t-5}, \boldsymbol{\Gamma}_{t-1}\right)$

  - $\boldsymbol{\Xi}_t = \begin{bmatrix} \Re(\boldsymbol{\psi}_{\mathrm{s},t}\boldsymbol{\psi}_{\mathrm{s},t}^H) \\ \Im(\boldsymbol{\psi}_{\mathrm{s},t}\boldsymbol{\psi}_{\mathrm{s},t}^H) \\ \Re(\boldsymbol{\psi}_{\mathrm{c},t}\boldsymbol{\psi}_{\mathrm{c},t}^H) \\ \Im(\boldsymbol{\psi}_{\mathrm{c},t}\boldsymbol{\psi}_{\mathrm{c},t}^H) \end{bmatrix} \rightarrow \widehat{\boldsymbol{\Xi}}_t = \frac{1}{\sigma_{\boldsymbol{\Xi}_t}}\left(\boldsymbol{\Xi}_t - \boldsymbol{\mu}_{\boldsymbol{\Xi}_t}\right) \in \mathbb{R}^{4\times(N\times N)}$

    - 채널 정보 (CSI) 텐서
      Cascaded channel의 autocorrelation에서 실수/허수부 분리
      훈련 안정성을 위해 정규화(normalization)하여 사용

    where $\boldsymbol{\psi}_{\mathrm{s},t} = \mathrm{diag}(\mathbf{h}_{\mathrm{r}\rightarrow\mathrm{t}}^H)\mathbf{G} \in \mathbb{C}^{N\times M}$, $\boldsymbol{\psi}_{\mathrm{c},t} = \mathrm{diag}(\mathbf{h}_{\mathrm{r}\rightarrow\mathrm{c}}^H)\mathbf{G} \in \mathbb{C}^{N\times M}$

  - $\boldsymbol{\Gamma}_t = \begin{bmatrix} \gamma_{\mathrm{s},t} \\ \gamma_{\mathrm{c},t} \end{bmatrix} \in \mathbb{R}^{2\times 1}$

    - 센싱/통신 SNR 정보 벡터
      실수 값 그대로 사용

MDP – Action (a_t) & Reward (r_t) design

- $a_t = (\boldsymbol{\Theta}_t, \boldsymbol{\Theta}_t^H)$
  - 현재 step t에서 RIS 위상 변화 행렬 $\boldsymbol{\Theta}_t$과 해당 행렬의 Hermitian 행렬 $\boldsymbol{\Theta}_t^H$를 action으로 정의

- $r_t = \underbrace{\alpha_{\mathrm{corr}}\,\rho_t}_{(1)} + \underbrace{\beta_{\mathrm{w}}\gamma_{\mathrm{s},t}}_{(2)} + \underbrace{(1-\beta_{\mathrm{w}})\gamma_{\mathrm{c},t}}_{(3)} - \underbrace{\lambda_{\mathrm{p}}\left(\tau_{\mathrm{c}} - \gamma_{\mathrm{c},t}\right)}_{(4)}$

(1) 센싱/통신 채널이 유사할수록 센싱 SNR ↑ **(2)** 최대화 목적(센싱 SNR)
(3) 제약 조건(통신 SNR) 만족을 위한 항 **(4)** 제약 조건 위반 시 패널티로 반영

  - $\alpha_{\mathrm{corr}}, \beta_{\mathrm{w}}, \lambda_{\mathrm{p}}$는 각각 채널 상관 계수 반영, 센싱/통신 SNR 가중 반영, 제약 위반 여부 반영에 대한 파라미터

  - $\rho_t = \frac{\left|\mathbf{h}_{\mathrm{t}}^H \mathbf{h}_{\mathrm{c}}\right|^2}{\|\mathbf{h}_{\mathrm{t}}\|^2 + \|\mathbf{h}_{\mathrm{c}}\|^2}$는 센싱/통신 effective channel의 상관 계수

RL 알고리즘: SAC

- 알고리즘 선택 근거
  1) RIS 위상(각도)은 연속 변수이므로, continuous action space를 가짐
     - DQN과 같이 연속 변수를 이산화 시 dimension 급증, 오차로 인해 성능 하락
     - 설계한 state 역시 연속 공간이며, dimension 매우 큼 → deterministic policy 계열보다 효율적
  2) 탐색(Exploration)이 전체 성능에서 매우 중요한 문제임
     - RIS 위상에 따라 BS-타겟/사용자 간의 end-to-end 채널이 결정 → 각도의 미세한 변화가 센싱/통신 SNR 값에 큰 영향
     - SAC는 불확실성(Entropy)을 유지하면서 stochastic policy를 출력하므로, exploration 능력이 deterministic policy 계열보다 우수
  3) 센싱 SNR을 극대화 하면서도 통신 SNR을 일정 수준 보장해야 함
     - SAC의 stochastic policy는 다양한 trade-off를 탐색하면서 deterministic policy 계열보다 빠르게 균형점을 찾을 수 있어 multi-objective reward에 유리
  4) 데이터 효율성이 중요한 환경임
     - 시뮬레이션 관점과 실제 구현의 관점 모두에서, 채널 데이터를 위한 환경 상호작용 비용이 매우 큼(즉, 생성 또는 수집/연산 비용이 큼) → 적은 데이터로도 학습이 효율적인 알고리즘이 유리
     - SAC는 off-policy method로서, 학습 과정에서 경험한 transition을 replay buffer에 저장하여 사용하므로 데이터 효율이 높고, 환경과의 상호작용 비용을 낮출 수 있음

RL 알고리즘: SAC 모델 구조 (1/2)

- Actor network: $\pi_\phi(a|s)$
  - Input: s_t ← 이때, $s_t = [\Re(\boldsymbol{\Psi}_{\mathrm{s},t}\boldsymbol{\Psi}_{\mathrm{s},t}^H), \Im(\boldsymbol{\Psi}_{\mathrm{s},t}\boldsymbol{\Psi}_{\mathrm{s},t}^H), \Re(\boldsymbol{\Psi}_{\mathrm{c},t}\boldsymbol{\Psi}_{\mathrm{c},t}^H), \Im(\boldsymbol{\Psi}_{\mathrm{c},t}\boldsymbol{\Psi}_{\mathrm{c},t}^H), \gamma_{\mathrm{s},t}, \gamma_{\mathrm{c},t}] \in \mathbb{R}^{4N^2+2}$ 확장 및 flatten
  - Output: 각 RE의 위상(각도 $\hat{\theta}_n$)에 대한 가우시안 평균과 로그 분산
    - 이후 $\boldsymbol{\Theta} = \mathrm{diag}\left(\mathrm{e}^{j\hat{\theta}_1}, \mathrm{e}^{j\hat{\theta}_2}, \ldots, \mathrm{e}^{j\hat{\theta}_N}\right) \in \mathbb{C}^{N\times N}$ 으로 변환
- Critic – main Q network Q_{θ_j}, $j = 1,2$
  - Input: s_t, a_t ← 이때, $s_t \in \mathbb{R}^{4N^2+2}$, $a_t = \left[\Re\left(\mathrm{e}^{j\hat{\theta}_n}\right), \Im\left(\mathrm{e}^{j\hat{\theta}_n}\right)\right] \in \mathbb{R}^{2N}$ 확장 및 flatten
  - Output: $Q_{\theta_j}(s_t, a_t), \forall j = 1,2$
    - 이후 $\min_{j=1,2} Q_{\theta_j}(s_t, a_t)$을 사용하여 Q 값 overestimation을 줄임
- Critic – target Q network : $Q_{\theta'_j}$, $j = 1,2$
  - 각 main Q network에 대해 동일한 구조의 target Q network를 두고, parameter soft update ($\theta'_j \leftarrow \tau_{\mathrm{soft}}\theta_j + (1-\tau_{\mathrm{soft}})\theta'_j$)로 학습 안정성 향상

RL 알고리즘: SAC 모델 구조 (2/2)

- Target state value (soft value)
  - 다음 state s'에서의 action $a' \sim \pi_\phi(\cdot|s')$에 대해, soft value $V(s') = \min_{j=1,2} Q_{\theta'_j}(s', a') - \alpha \log \pi_\phi(a'|s')$

- 각 main Q network의 loss function 및 parameter update
  - $\mathcal{L}_Q(\theta_j) = \frac{1}{K}\sum_{i=1}^{K}(r_i + \gamma V(s'_i) - Q_{\theta_j}(s'_i, a'_i))^2,\ \theta_j \leftarrow \theta_j - \eta_\theta \nabla_{\theta_j}\mathcal{L}_Q(\theta_j),\ \forall j = 1,2$

- Actor network의 loss function 및 parameter update
  - $\mathcal{L}_\pi(\phi) = \frac{1}{K}\sum_{i=1}^{K}(\alpha \log \pi_\phi(a_i|s_i) - \min_{j=1,2} Q_{\theta_j}(s_i, a_i)),\quad \phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}_\pi(\phi)$

- Temperature α의 loss function 및 autotuning
  - $\mathcal{L}_\alpha = -\mathbb{E}_{\alpha\sim\pi_\phi}[\alpha(\log \pi_\phi(a|s) + \mathcal{H}_{\text{target}})],\quad \log\alpha \leftarrow \log\alpha - \eta_\alpha \nabla_{\log\alpha}\mathcal{L}_\alpha$
  - Policy entropy가 $\mathcal{H}_{\text{target}}$보다 낮으면 α ↑ → 탐색 ↑, 높으면 α ↑ → 탐색 ↓
    - 학습 초반에는 탐색 ↑, 학습 후반에는 안정적 수렴

실험 셋업

- 실험 환경
  - GPU: NVIDIA RTX 4060/4090
  - Episodes: 1000
    - 1000 steps/episode × 1000 episodes = **1×10^6** total environment interactions

- Simulation parameters

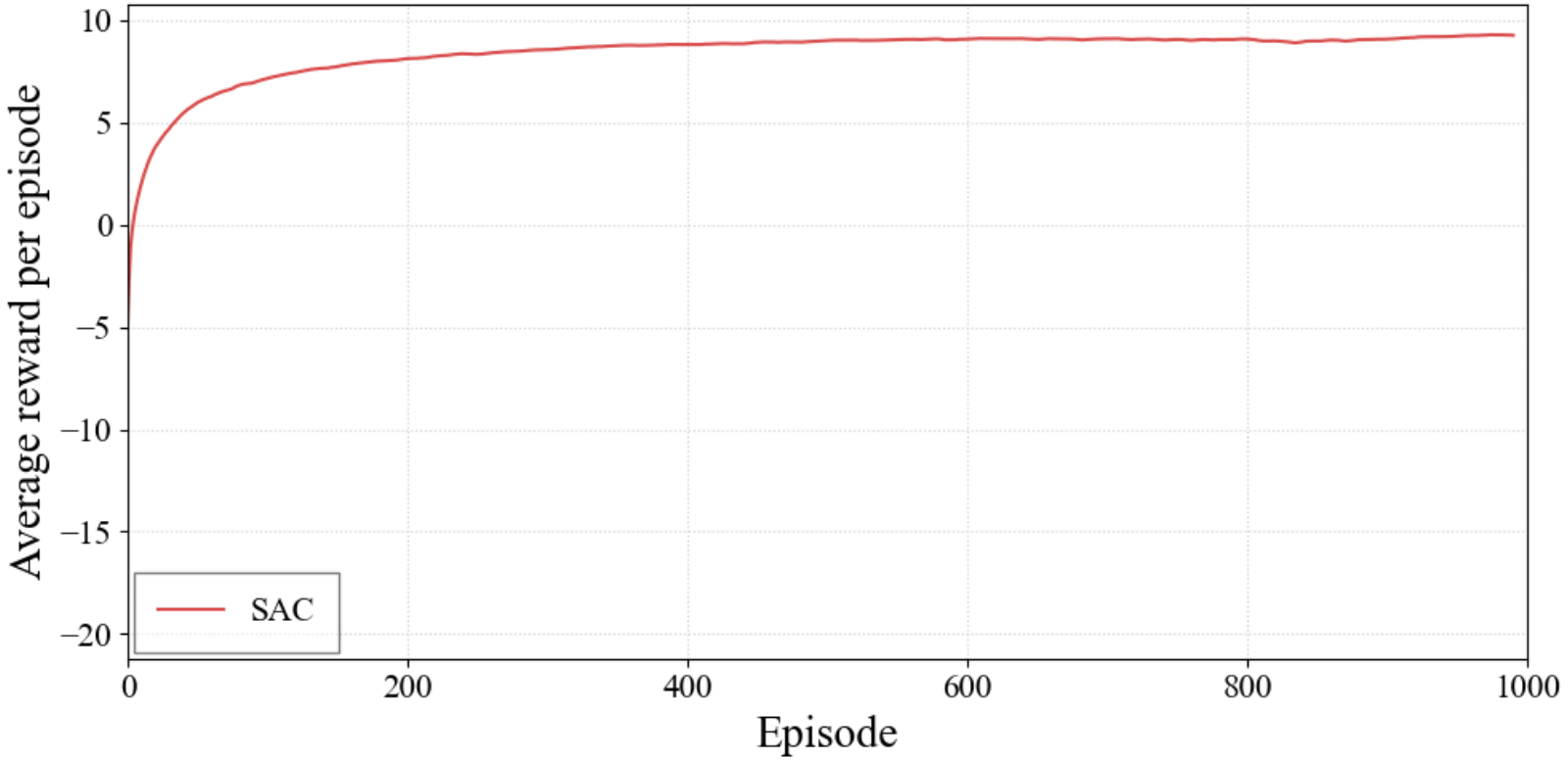
Parameter	Value
Number of BS antennas, M	8
Number of RIS elements, N	32
Sensing noise variance, σ_s^2	0.01
Communication noise variance, σ_c^2	0.01
Communication SNR threshold, τ_c	10 dB
BS maximum transmit power, $P_{\mathrm{tx,max}}$	10 W
Correlation weight (in reward), α_{corr}	1.0
SNR balancing weight (in reward), β_w	0.99
Penalty weight (in reward), λ_p	0.5

- SAC hyperparameters

Parameter	Value
Learning rate for actor, critic, temperature, η_ϕ, η_θ, η_α	1×10^{-3}
Batch size, K	256
Replay buffer size, $\lvert\mathcal{D}\rvert$	1×10^6
Discount factor, γ	0.99
Soft update rate, τ_{soft}	0.005
Actor hidden dims	[256, 64]
Critic hidden dims	[512, 256]
Target entropy, $\mathcal{H}_{\mathrm{target}}$	$-N$ (default)
Initial temperature, α	0.2
Temperature autotuning	active

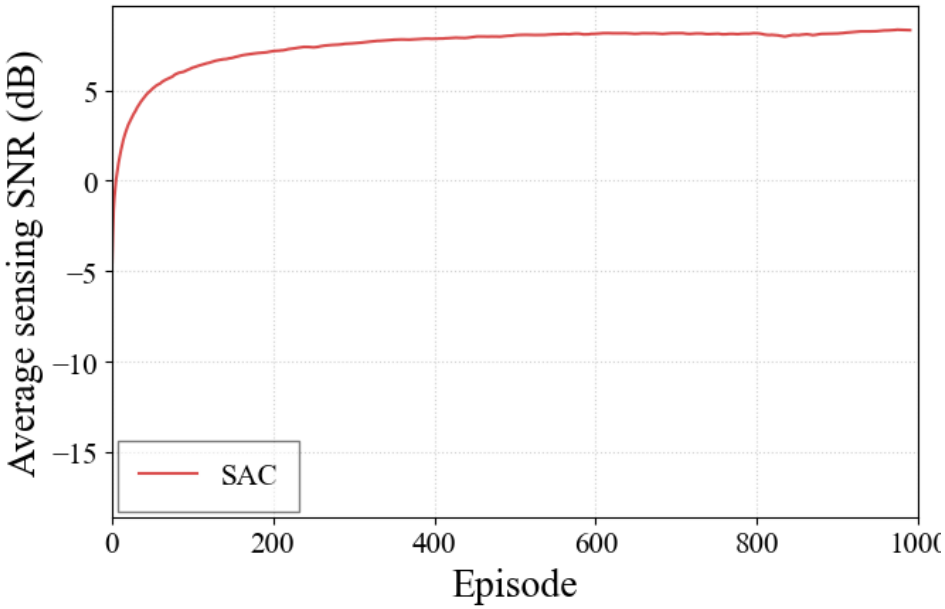
실험 결과 ①: 강화학습 수렴 (1/3)

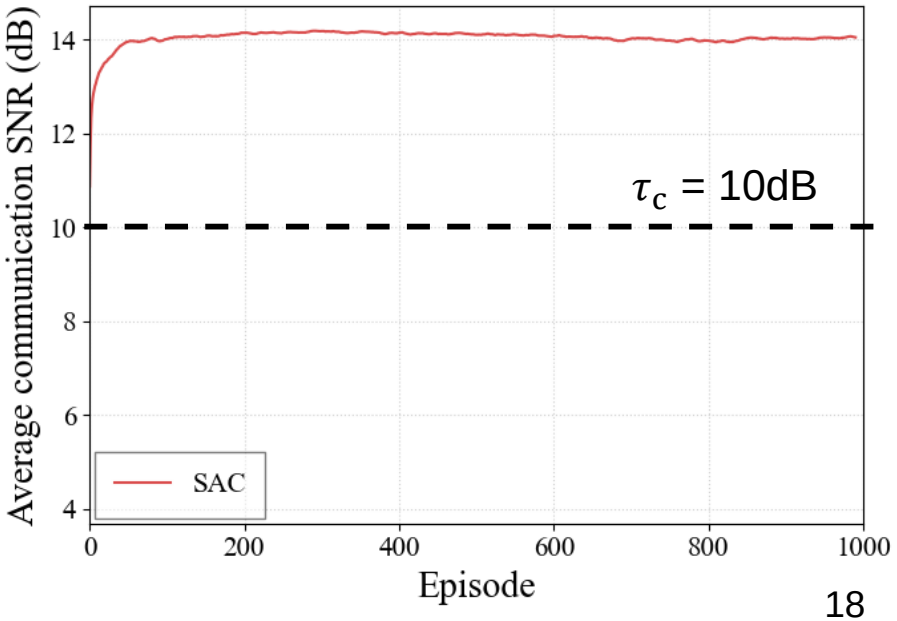
- **Learning curve:** 에피소드 당 reward 평균 값
- 단일 수행에 결과에 대해 moving average로 plot
  - Window size = 10 episodes
  - Stochastic policy와 exploration으로 인해 episode reward fluctuation → 학습 안정성의 trend 관찰을 위해 smoothing
  - 학습 초반 에피소드에서 빠르게 증가, 학습 후반 에피소드에서 수렴하는 양상



실험 결과 ①: 강화학습 수렴 (2/3)

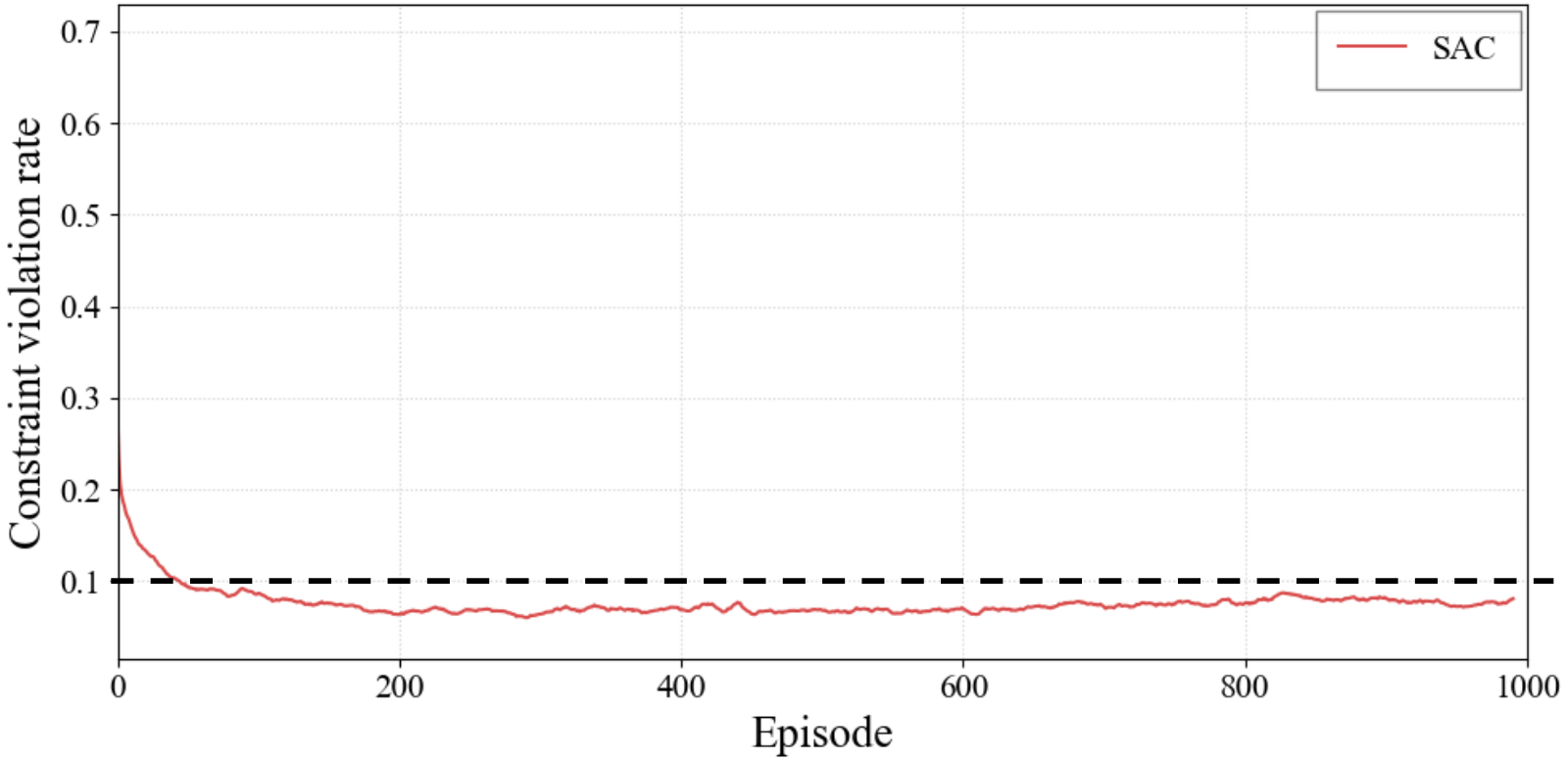
- **SNR curve:** 에피소드 당 센싱 SNR(최대화 목적)의 평균 값과 통신 SNR(제약 조건)의 평균 값
- 단일 수행에 결과에 대해 moving average로 plot (Window size = 10 episodes)
  - 센싱 SNR은 reward curve와 유사한 개형 → 강화학습 수렴과 동시에 목적함수 최대화
  - $r_t = \alpha_{\mathrm{corr}}\,\rho_t + \beta_{\mathrm{w}}\gamma_{\mathrm{s},t} + (1-\beta_{\mathrm{w}})\gamma_{\mathrm{c},t} - \lambda_{\mathrm{p}}(\tau_{\mathrm{c}} - \gamma_{\mathrm{c},t})$)에서 센싱 SNR이 가장 큰 비중을 차지함 (β_{w}=0.99)
  - 또한, 통신 SNR은 r_t에서 작은 비중을 차지함에도 채널 상관 계수와 패널티 항으로 인해 수렴과 동시에 증가하는 양상, 평균적으로 τ_{c} = 10dB 이상 제약 조건을 만족하는 결과





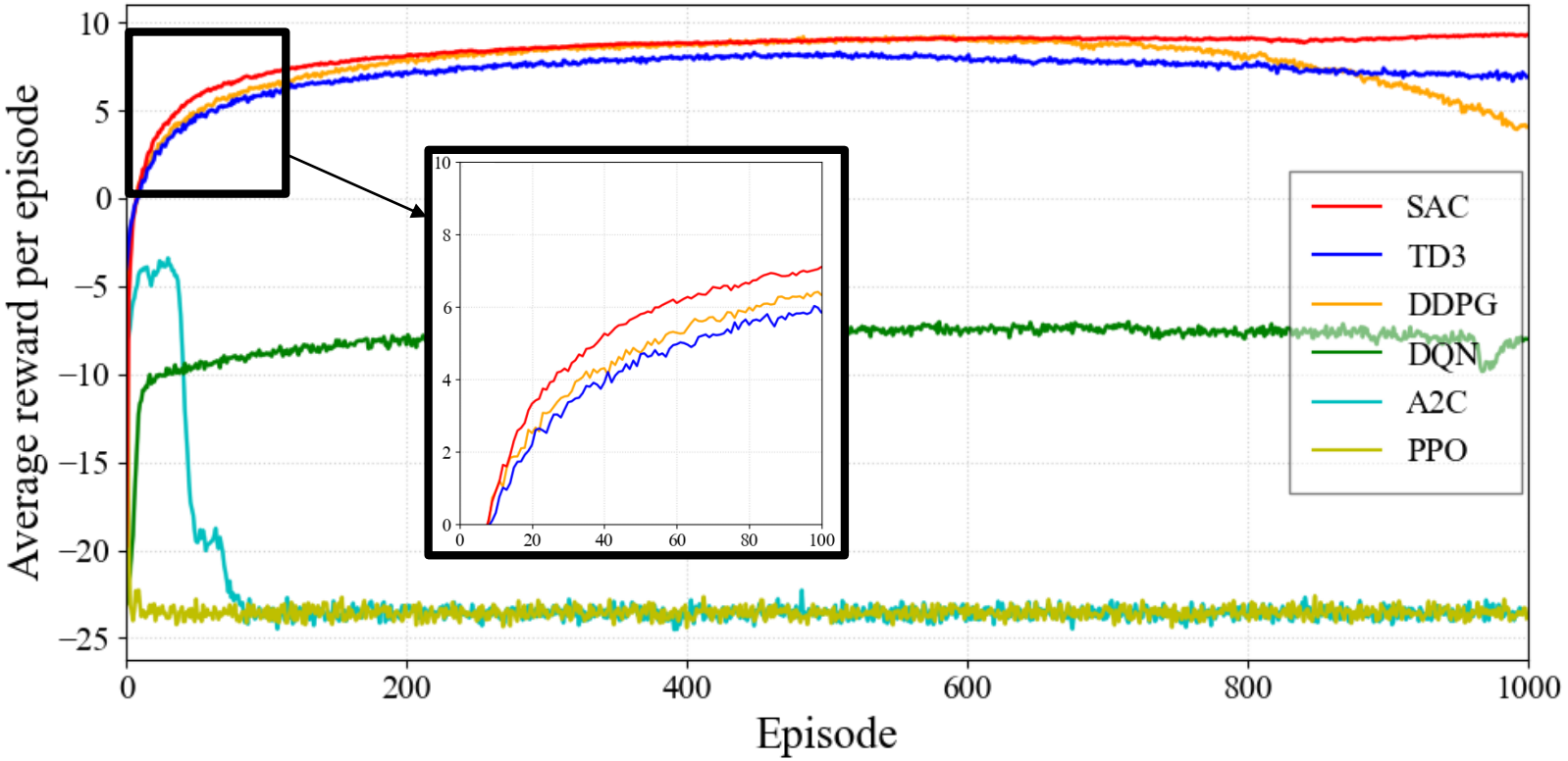
실험 결과 ①: 강화학습 수렴 (3/3)

- **Constraint violation ratio curve:** 에피소드 전체 step 중에서 제약을 위반한, 즉 통신 SNR이 임계값 τ_c를 넘지 못한 step의 평균 비율
- 단일 수행에 결과에 대해 moving average로 plot (Window size = 10 episodes)
  - 제약 위반율은 학습 초반 에피소드에서는 높은 값을 가지다가, 학습 후반 에피소드에서 수렴하면서 낮아짐
  - 에피소드 당 전체 step의 0.1 (10%)미만 비율로 제약을 위반



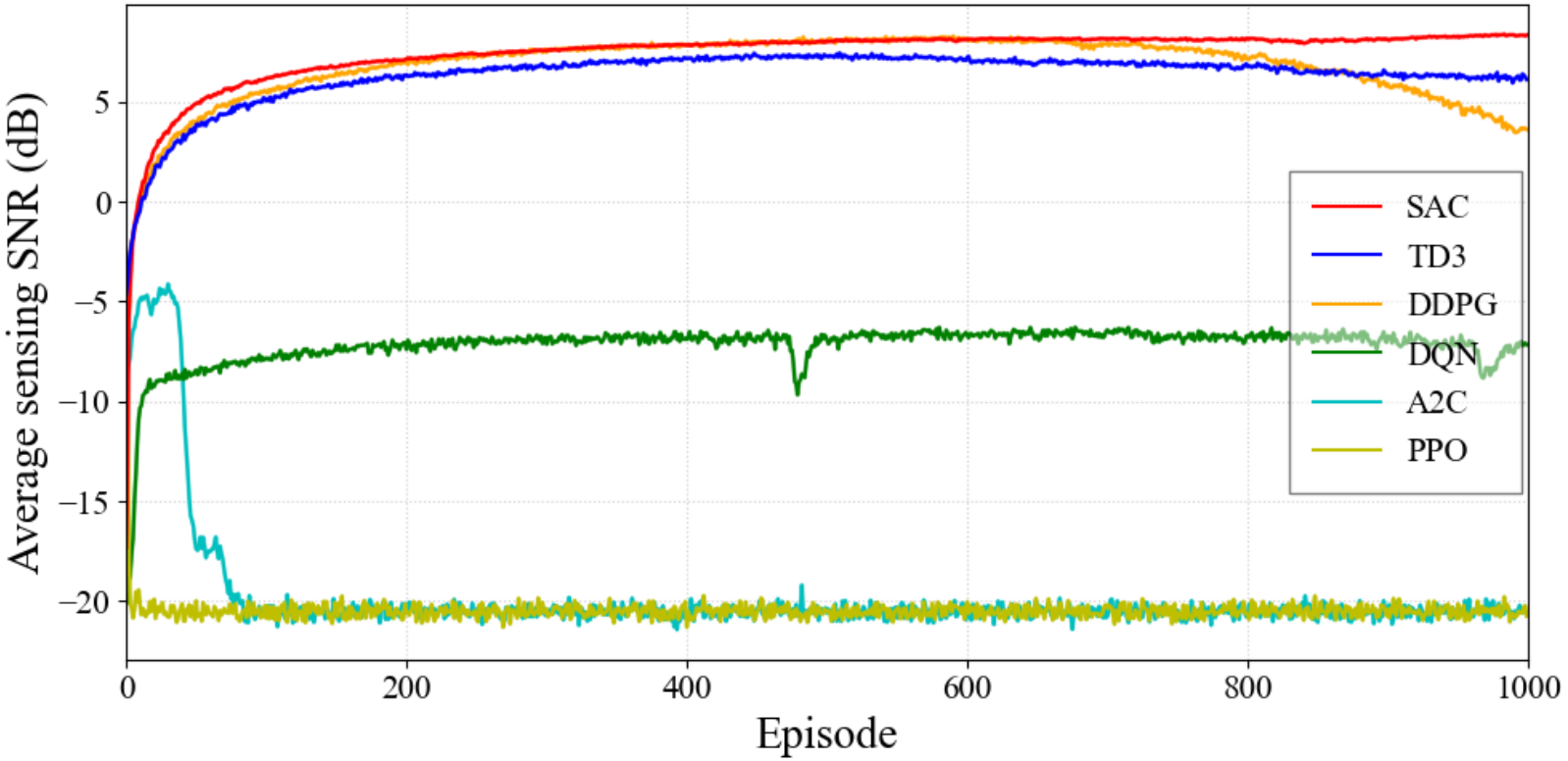
실험 결과 ②: 다른 모델과 비교 (1/4)

- **Learning curve:** 에피소드 당 reward 평균 값
  - 각 모델 랜덤 시드 5회 씩 실행 후 수렴 reward가 가장 큰 값을 가지는 단일 수행 결과에 대해 plot하여 비교 (moving average 미적용) (cf. 여기서 SAC의 $\alpha = 0.2$ 결과임)
- 비교 모델: off-policy 계열의 TD3, DDPG, DQN / on-policy 계열의 A2C, PPO
  - SAC, TD3, DDPG 결과 유사–SAC가 근소하게 높은 성능 보임(빠른 수렴,훈련 안정성)
  - DQN은 action space를 64개로 discretization → 수렴성은 어느정도 보이나, 이산화 오류 및 모델 구조 차이로 인해 낮은 성능을 보임



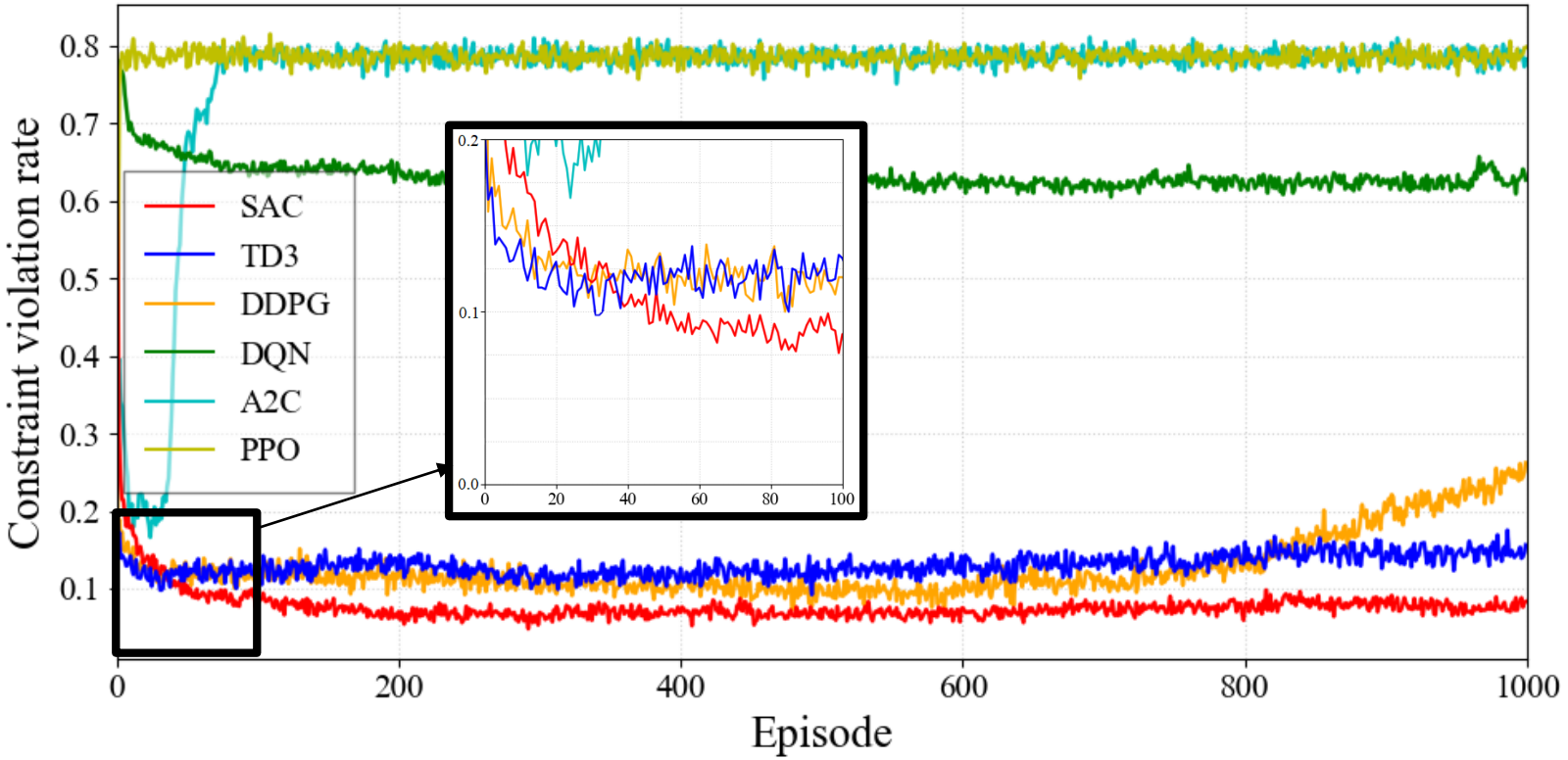
실험 결과 ②: 다른 모델과 비교 (2/4)

- **Sensing SNR curve:** 에피소드 당 센싱 SNR 평균 값
  - 모든 모델에서 센싱 SNR과 reward curve와 유사한 개형 확인
  - A2C, PPO는 환경 동일, 데이터를 활용 방식에만 차이를 두었으나, 두 모델 모두 좋은 성능을 보이지 못함 → on-policy와 off-policy의 차이인 파라미터 업데이트 방식과 사용하는 샘플 특성(효율성)차이에서 기인하는 것으로 추측
    (이에 대한 자세한 분석은 결과 토의에서 작성함)



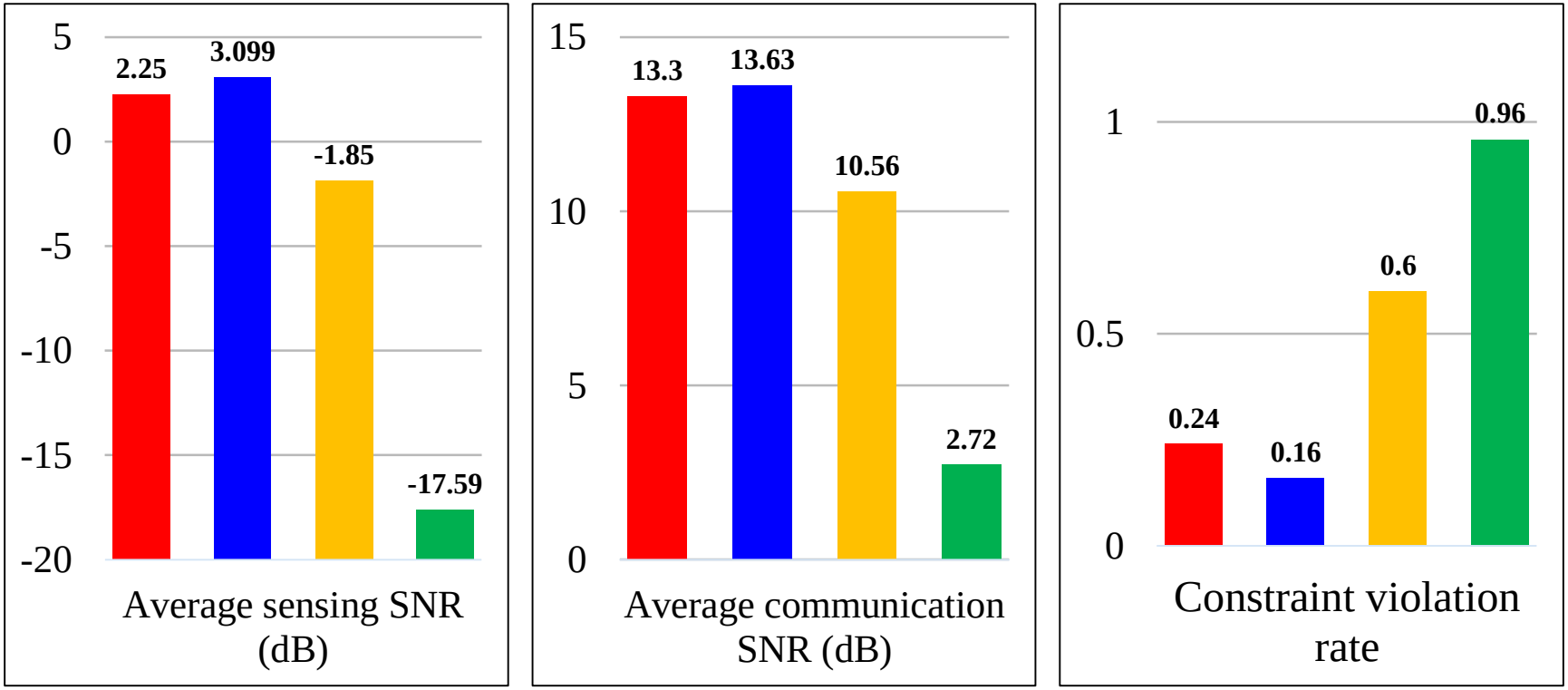
실험 결과 ②: 다른 모델과 비교 (3/4)

- **Constraint violation ratio curve:** 에피소드 전체 step 중에서 제약을 위반한 step의 평균 비율
  - SAC > TD3 > DDPG 순으로 위반율 낮음
    - DQN, A2C, PPO는 제약 위반이 매우 심각 → 현재 모델링으로는 실환경 적용 거의 불가 수준
  - 학습 초반 에피소드에서는 TD3의 위반율 감소 속도가 가장 빨랐으나, 수렴이 진행되면서 SAC의 위반율이 더 낮아짐
  - 학습 후반 에피소드에서 TD3가 DDPG보다 안정적인 훈련이 진행된 것으로 보임



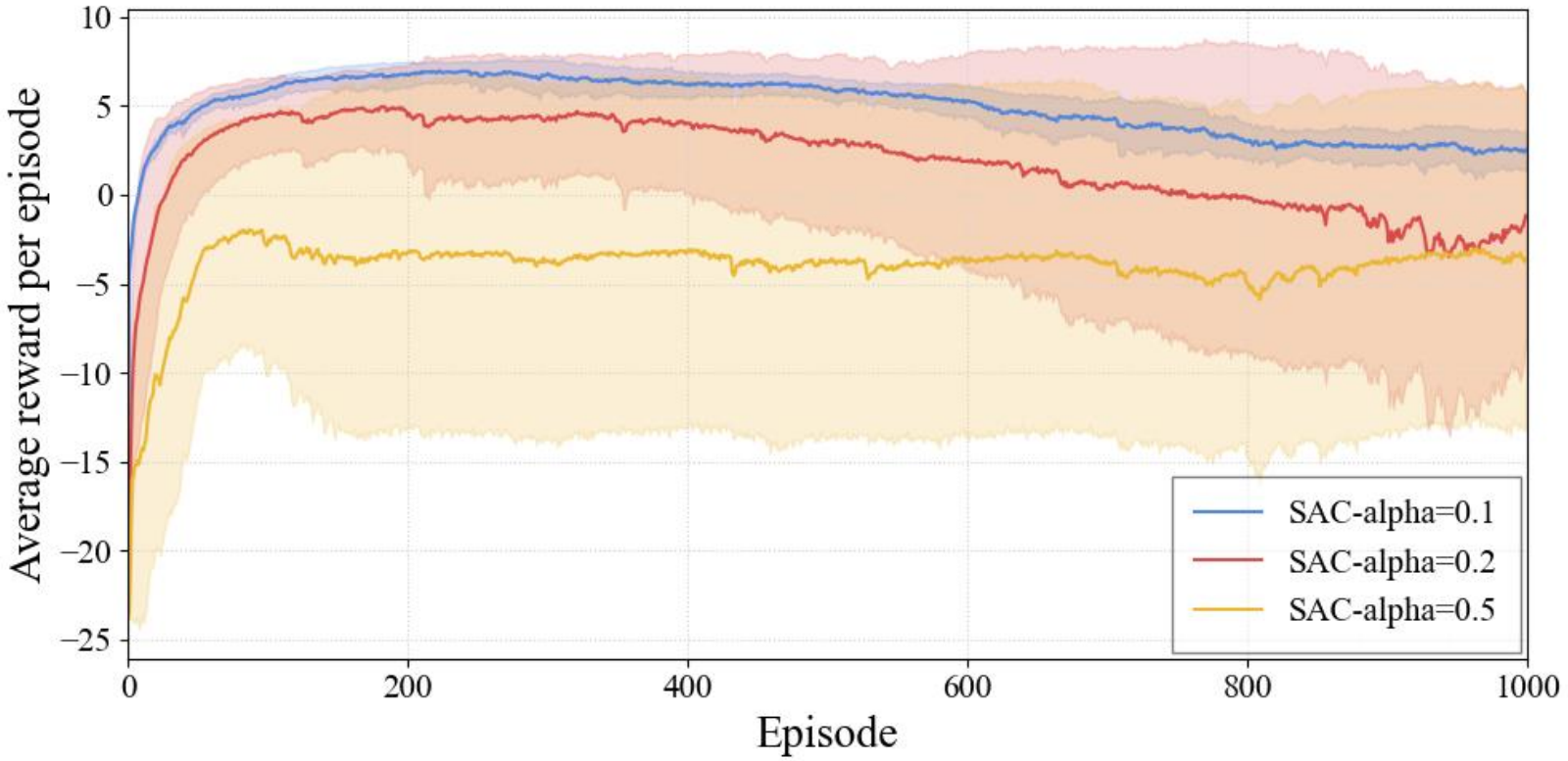
실험 결과 ②: 다른 모델과 비교 (4/4)

- **Generalization performance:** 훈련이 완료된 모델을 테스트 데이터로 평가한 결과 – 훈련 데이터와 동일한 구조의 채널 데이터를 재생성 (단, 사용자의 속도는 랜덤)\
  - Off-policy 계열 4개 모델만 비교: 일반화 성능은 TD3 > SAC > DDPG > DQN 순서를 보임
  - 특히, SAC와 TD3는 새로운 데이터에서도 제약 위반율을 낮게 유지하여 robustness와 generalization performance를 확인할 수 있음
  - 훈련 성능과 반대로 일반화 성능은 TD3 > SAC → SAC의 훈련데이터 overfitting 추측



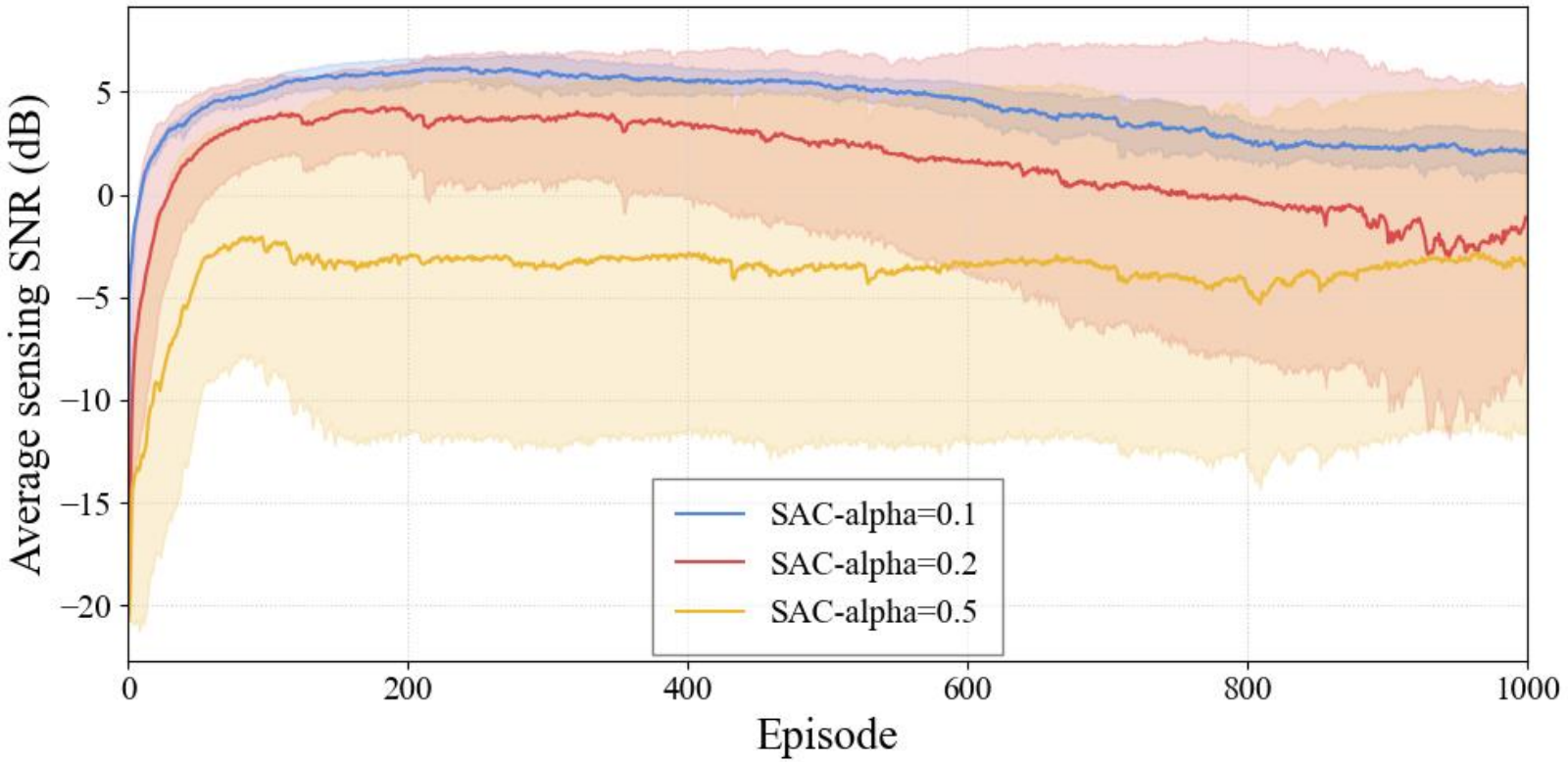
실험 결과 ③: Hyperparameter sensitivity (1/3)

- **Learning curve:** 에피소드 당 reward 평균 값
  - Initial temperature α값에 따른 모델 학습 성능 비교 ($\alpha \in \{0.1, 0.2, 0.5\}$)
  - 각 경우 랜덤 시드 5회 씩 실행 후 평균값을 bold line, 95% 신뢰구간(CI)을 shaded region으로 plot
  - α값이 클수록 평균 reward의 신뢰구간의 영역이 넓은 것을 확인
    - 초기 α가 크면 정책의 stochasticity가 비교적 오랜 학습동안 유지되는 경향



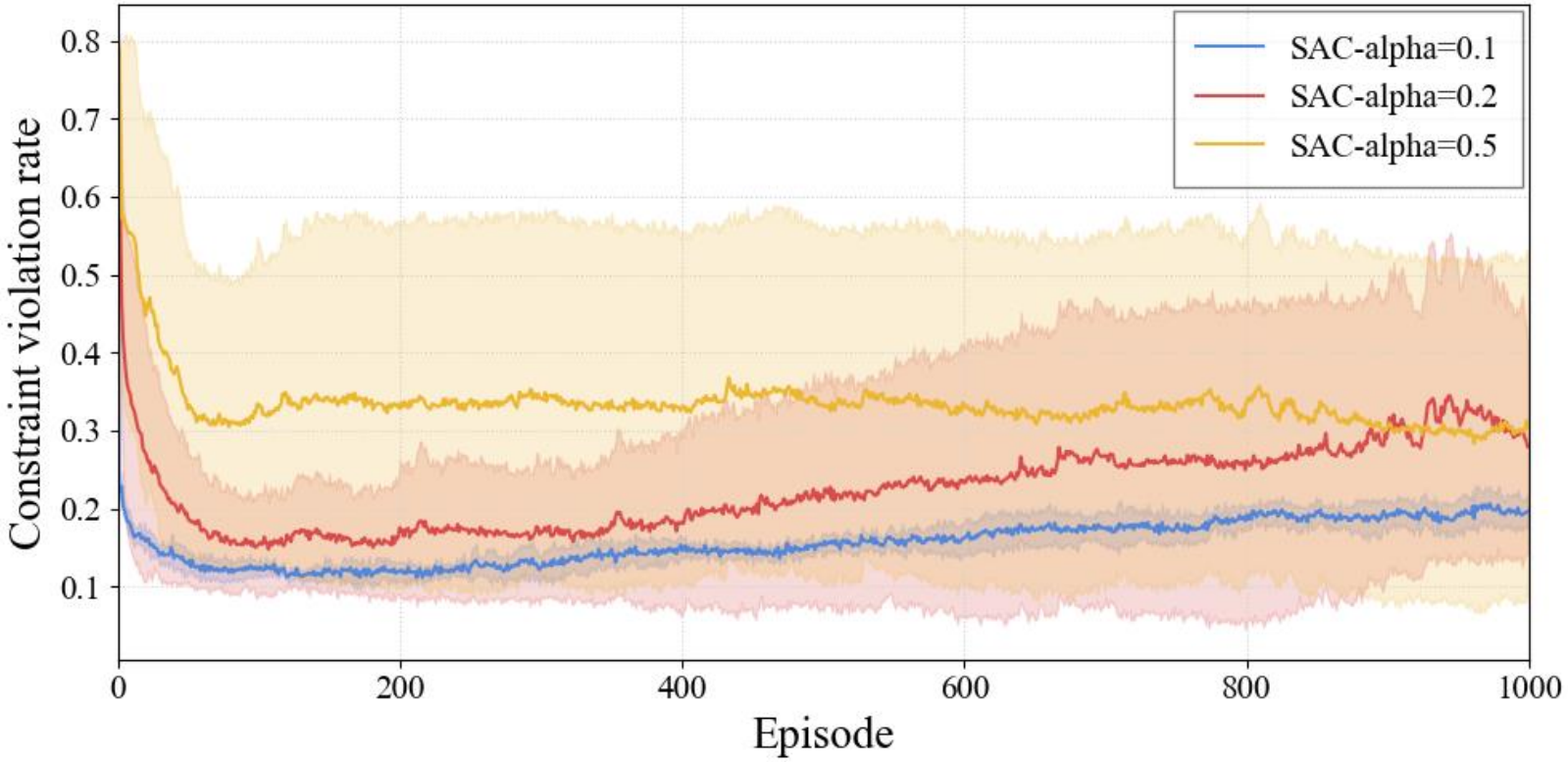
실험 결과 ③: Hyperparameter sensitivity (2/3)

- **Sensing SNR curve:** 에피소드 당 센싱 SNR 평균 값
  - Initial temperature α값에 따른 모델 학습 성능 비교 ($\alpha \in \{0.1, 0.2, 0.5\}$)
  - 센싱 SNR은 reward curve와 유사한 개형
  - α값이 작으면 학습 초반 entropy 유지 약화 → policy 빠르게 deterministic하게 수렴, exploration 비교적 부족하므로 local optimum에 도달하거나 변동성이 커질 위험 있음
  - α값이 크면 학습 entropy를 매우 강하게 유지 → policy는 매우 느리게 수렴, exploration 비교적 풍부하므로 초반에 낮은 값에 머무르거나 variance가 매우 큰 단점 있음



실험 결과 ③: Hyperparameter sensitivity (3/3)

- **Constraint violation ratio curve:** 에피소드 전체 step 중에서 제약을 위반한 step의 평균 비율
  - Initial temperature α값에 따른 모델 학습 성능 비교 ($\alpha \in \{0.1, 0.2, 0.5\}$)
  - 제약 조건 위반율 역시 α값 대소에 따른 추이는 Learning curve/ Sensing SNR curve와 동일 해석 가능
    - α값이 클수록 평균 위반율의 신뢰구간 영역 넓음
    - α=0.5 > 0.2 > 0.1 순으로 학습 초반 위반율 감소 속도 느림



토의 및 결론 (1/4)

- **설계한 MDP 구조가 off-policy에 유리하고 on-policy에 매우 불리함**
  - 상태 s_t의 일부는 시뮬레이션으로 이미 정해진 채널 변화 데이터로 주어지고, 정책 $\pi(a_t|s_t)$와 행동 a_t는 state transition에 결정적인 영향을 미치지 못함
    - RIS 위상 변화는 cascaded 채널 $\boldsymbol{\Xi}_t$ 자체에는 영향을 주지 않고, SNR $\boldsymbol{\Gamma}_t$에만 영향을 줌 (그러나, $\boldsymbol{\Gamma}_t$ 의 dimension $\gg$ $\boldsymbol{\Xi}_t$의 dimension)
    - 즉, 설계한 MDP 구조 상 $P(s_{t+1}|s_t, a_t) \approx P(s_{t+1}|s_t)$이고, 이런 환경에서는 policy의 변화가 state의 방문 분포를 급격히 바꾸지 못함 → off-policy 데이터 재사용이 학습에 매우 안정적임
  - Action과 state의 coupling이 낮은 실험 환경이 오히려 off-policy에서 발생 가능한 distribution shift 문제를 완화시킴
  - 채널 데이터가 외부에서 생성되면서, 상태 s_t의 분포가 매우 다양함
    - Replay buffer에 쌓인 데이터들이 다양한 채널 snapshot으로 구성
    - 이런 경우, Replay buffer에서 랜덤으로 mini-batch sampling하는 것이 unbiased training distribution을 제공함 → actor-critic 학습을 매우 안정화 시킬 수 있음

토의 및 결론 (2/4)

- **State 차원을 너무 크게 설계**
  - 데이터 생성 비용이 높고(채널 시뮬레이션), 복잡도가 매우 큼
  - 따라서 실환경 구현을 고려하면, 복잡도 및 샘플 수집 측면에서 보완 필요
    - CNN과 같이 channel feature를 효과적으로 추출 가능한 신경망을 배치하거나 채널 수집 시에 전처리하여 state/input 차원을 줄일 수 있을 것으로 생각
  - 프로젝트 시간 관계상 많은 수의 반복 실험을 하지 못하여 유의미한 통계적 특징 관찰 부족
  - 다른 모델과의 정확한 성능 비교 또한 반복 실험 후 오차범위 기반의 비교 필요

- **학습이 진행될 수록 신뢰구간(CI)가 넓어지는 원인**
  - 랜덤 시드마다 서로 다른 방향으로 학습하여 정책 간 분산 증가 가능성
    - 시드마다 critic error가 누적되면서 gradient noise 증가 및 stochastic policy → 에피소드 진행될 수록 reward curve 다른 양상으로 진행
  - Reward variance 증가 → episode 진행하면서 critic의 불안정성 증가

토의 및 결론 (3/4)

- **실험 결과 요약 및 프로젝트 결론**
  - 학습 과정 분석
    - Learning curve와 sensing SNR curve 모두에서 SAC는 학습 초반 빠르게 증가하고 학습 후반 안정적으로 수렴하는 경향
    - Stochastic policy와 exploration 특성으로 인해 개별 episode reward는 변동성을 보였으나, moving average 기준으로는 분명한 증가 추세를 확인
  - 제약 위반율 분석
    - 학습 초반에는 높았으나, 학습이 진행될수록 감소하여 에피소드당 전체 step의 약 10% 미만 수준으로 수렴 → SAC 에이전트가 제약 조건을 반영한 reward를 바탕으로 통신 SNR을 일정 수준 이상 유지하도록 학습
  - 다른 모델과의 비교 및 일반화 성능
    - SAC는 TD3/DDPG 대비 더 낮은 제약 위반율과 더 빠른 수렴 안정성
    - DQN/A2C/PPO는 연속 행동공간 제어 및 on-policy 특성(MDP 설계와 맞지 않음)으로 인해 성능 ↓
    - Test dataset 평가에서 어느정도 robustness와 일반화 성능 확인
  - Initial temperature α sensitivity
    - α 클수록 정책 stochasticity 유지↑ → 학습 초반 variance 증가, 수렴 속도 ↓
    - α 작을수록 탐색 부족 → 빠른 deterministic 수렴 → local optimum 위험 ↑

토의 및 결론 (4/4)

- **개선 방안**
  - 1) MDP 설계 수정
    - 예를 들어 현재 reward는 SNR 값과 제약 위반 시 패널티 항을 더하는 방식
    - → 이를 제약 위반에 따른 adaptive penalty로 반영하여, 제약 위반 빈도에 따라 동적으로 parameter 값을 조정하도록
  - 2) 문제를 POMDP(Partially Observable Markov Decision Process)로 설계
    - 에이전트가 관측한 과거 채널 정보를 바탕으로 현재 채널을 직접 예측하도록
    - 지연된 채널 시퀀스 관측 o_t → 현재 채널 상태 s_t을 예측하는 시계열 네트워크 모델(e.g., LSTM, Transformer)을 이용하여 채널 예측
    - 이 경우, reward에 채널 예측 또는 센싱 추정 성능을 직접 반영하는 것도 가능

Appendix: 팀원 기여 사항

- 구성원 별 기여 항목

이름 / 항목	김현정 (팀장)	김문일	민경현
문제 정의 및 환경 구현	최적화 문제 설계 및 데이터셋 생성	시스템 모델 설계	강화학습 MDP 설계 및 학습 환경 구현
알고리즘 구현 및 실험 설계	DDPG, TD3 구조 설계 및 구현, 센싱 및 통신 수신 SNR 실험 설계	Actor/critic 신경망 구조 설계 및 구현, 하이퍼 파라미터 설정 실험 설계	SAC, DQN 구조 설계 및 구현, 에피소드에 대한 누적 보상 실험 설계, 제약 조건 위반 관련 실험 설계
성능 평가 및 해석	센싱 및 통신 SNR 성능 평가 및 해석	하이퍼 파라미터에 따른 성능 평가 및 해석	에피소드에 대한 누적 보상 및 제약 조건 위반 성능 평가 및 해석
실험 결과 정리 및 보고서 작성	전체 실험 결과 통합 및 검토 보고서(PPT) 작성	코드 정리 및 재현 가능성 검증 보고서(PPT) 작성	실험 결과 그래프 시각화, 실험 결과 신뢰 구간 작성, 보고서(PPT) 작성
기여도	100 %	100 %	100 %