Journal of Institute of Control, Robotics and Systems (2025) 31(2):87-97
https://doi.org/10.5302/J.ICROS.2025.24.0264 ISSN:1976-5622 eISSN:2233-4335

안전한 모델 예측 제어와 강화 학습에 대한 연구 동향

A Survey on Safe Model Predictive Control and Reinforcement Learning

함 형 찬, 안 희 진*
(Hyeongchan Ham[1] and Heejin Ahn[1,*])
[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)

Abstract: Recent progress in model predictive control (MPC) and reinforcement learning (RL) has led to a growing interest in many complex decision-making tasks. However, MPC heavily depends on the model and RL often disregards safety concerns, which prevents them from being applied to real-world applications. This survey presents the trends in MPC and RL from the perspective of safety concerns, based on different safety levels associated with MPC and different levels of interaction with the environment in RL. Also, integrated approaches that take advantage of the strengths of MPC and RL are explored. Finally, this paper outlines future directions for safe decision making based on MPC and RL.

Keywords: model-predictive control, reinforcement learning, safety critical system

I. 서론

모델 예측 제어(model predictive control, MPC)와 강화 학습(reinforcement learning, RL)은 의사결정(decision making)을 수행하는 작업에서 많이 활용된다. 이를 통해 manipulation이나 navigation과 같은 다양한 로봇 임무에 활용되며, 정의된 비용함수를 최소화하거나, 보상함수를 최대화하는 방향으로 행동을 함으로서 로봇 제어의 작업 능력이 향상되었다. 이러한 방법을 현실에 적용하기 위해서는 안전에 대한 고려가 필수적이다[1]. 로봇이 행동을 취하며 손상이 발생하거나 동적 물체와의 충돌을 피하기 못하면 물리적 손실과 돌이킬 수 없는 피해로 이어지게 된다. 따라서 이러한 위험을 고려하며 행동하는 것은 현실의 로봇 제어에서 중요한 과제이다.

안전은 실행 단계에서의 안전과 학습 단계에서의 안전으로 구분할 수 있다. 모델 예측 제어는 정책을 학습하는 과정 없이 실행하는 매 시점마다 특정 horizon의 길이에 대하여 최적화 문제를 풀어 최적의 행동을 수행한다. 본 논문에서는 외란과 오차로 인해 불확실한 환경에서 안전하게 동작할 수 있는 방법들을 중심으로 다룬다. 최적화 문제에 주어지는 안전에 대한 제약조건의 수준을 기준으로 모델 예측 제어의 방법들을 분류할 수 있다. Robust MPC는 매 시점마다 최대의 외란에도 제약조건이 만족되도록 최적화하며, stochastic MPC는 매 시점마다 특정 확률만큼 조건을 만족하도록 최적화를 수행한다. Sample-based MPC는 조건을 만족하는 경로를 대상으로 제어를 수행한다. 이 방법들은 정책을 학습할 필요 없이 안전에 대한 제약조건을 만족하며 매 순간마다 최적 해를 찾아낼 수 있지만, 모델에 대한 의존도가 크다는 단점이 있다. 또한 특정 horizon 내에서 최적의 해를 찾아내기 때문에 전역적인 최적 해가 아닌, 지역적 최적 해를 찾는 단점이 있으며, 매 시점마다 최적화를 수행하기 위한 비용이 필요하다.

실행 시점에 최적화하는 모델 예측 제어와 달리 강화 학습은 학습하는 동안 정책 최적화를 진행하며, 학습이 끝나면 정책은 고정이 되어 변하지 않게 된다. 정책을 학습하기 위해 강화 학습은 환경과 상호작용을 하게 되는데, 이때 최적 정책으로 수렴하지 않은 정책은 위험한 행동을 취하게 되어 물리적 손상과 같은 피해가 발생할 수 있다. 강화 학습은 환경과 상호작용하는 정도를 기준으로 여러 방법들을 분류할 수 있다. 모델 프리 강화 학습(model-free reinforcement learning)은 환경과의 상호작용을 통해 시행착오를 거치며 정책을 학습하게 된다. 모델 기반 강화 학습(model-based reinforcement learning)은 환경의 변화를 예측할 수 있는 모델을 통해 미래를 예측하며 학습을 함으로서 환경과의 상호작용을 줄일 수 있다. 오프라인 강화 학습(offline reinforcement learning)은 사전에 획득한 데이터만을 통해 환경과의 상호작용 없이 정책을 학습할 수 있다. 강화 학습은 실행 시 추가적인 최적화 비용 없이 매 시점마다

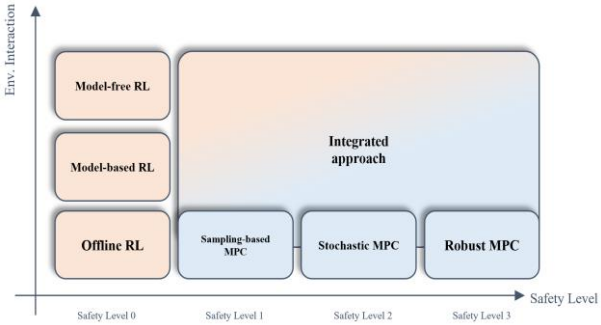

그림 1. 안전을 고려한 모델 예측 제어와 강화 학습 비교.
Fig. 1. Comparison of MPC and RL regarding safety concerns.

* Corresponding Author
Manuscript received November 8, 2024; revised December 1, 2024; accepted December 1, 2024
함형찬: 한국과학기술원 전기및전자공학부 석사과정 입학 예정(hyeongchan.ham@kaist.ac.kr, ORCID 0009-0001-3594-3709)
안희진: 한국과학기술원 전기및전자공학부 조교수(heejin.ahn@kaist.ac.kr, ORCID 0000-0001-9153-3491)
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2025-RS-2023-00259991).

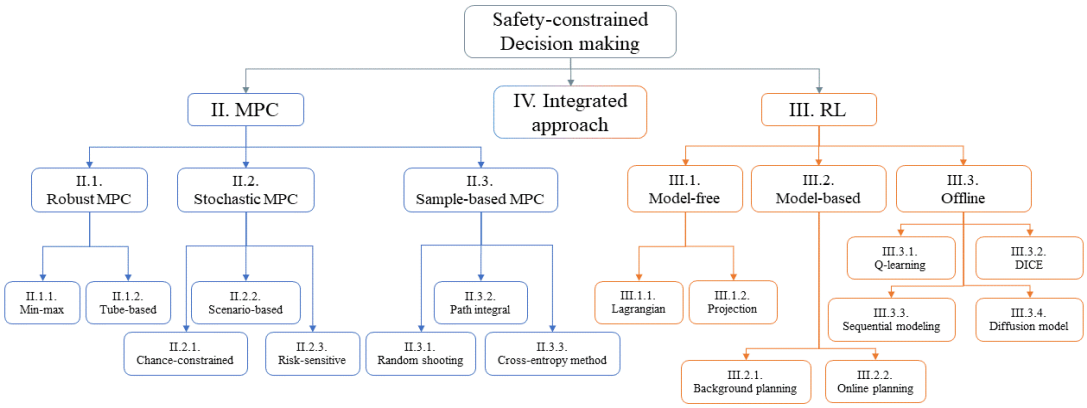

그림 2. 안전을 고려한 모델 예측 제어와 강화 학습을 이용한 방법 분류.
Fig. 2. Diagram of the methods with MPC and RL regarding safety concerns.

전역적 최적 해를 찾아낼 수 있지만, 정책을 학습 시키기 위한 비용이 드는 단점이 있다. 또한 학습을 위한 과정에서 위험한 환경에 노출이 되는 문제가 발생할 수 있다.

모델 예측 제어와 강화 학습을 안전한 환경에 적용하기 위해 본 논문에서는 그림 1과 같이 두 방식들을 안전을 보장할 수 있는 안전 단계(safety level)에 따라 분류하였다[1]. 표 1과 같이, 가장 높은 안전 단계인 safety level 3에서는 매 시점마다 제약조건을 반드시 만족하는 hard constraint를 가진다. 다음으로 높은 안전 단계인 safety level 2에서는 매 시점마다 특정 확률 이으로 제약조건을 만족하는 chance constraint를 가진다. 그리고 safety level 1에서는 제약조건 비용의 합에 대한 기댓값이 특정 값 이하로 만족하는 soft constraint를 가진다. 마지막으로 safety level 0은 가장 낮은 안전 단계이며, 실행하는 시점뿐만 아니라 학습하는 탐색 과정에서 위험에 노출될 가능성이 있는 강화 학습 정책들이 이 안전 단계에 포함된다. Safety level 0은 안전을 고려한 CMDP (Constrained Markov Decision Process)로 정의된 문제를 다루며, safety level 1과 동일하게 soft constraint를 주로 사용한다[2].

본 논문에서는 모델 예측 제어와 강화 학습을 그림 2와 같이 분류하였으며, 논문의 구성은 다음과 같다. 2장에서는 안전을 고려한 모델 예측 제어를 살펴보고, 3장에서는 안전을 고려한 강화 학습에 대해 살펴본다. 4장에서는 강화 학습에 제어 이론을 결합한 방법에 대해 살펴보고, 5장에서는 결론을 요약하며 안전한 제어를 위한 앞으로의 발전 방향을 제시한다.

표 1. Safety level 에 따른 제약조건의 형태.
Table 1. Constraint formulations with respect to safety levels.

Safety level	Constraint	Safety Guarantee Stage
Safety Level 3 (Hard constraint)	$\mathrm{c}_t(x,u,w) \le 0$	Execution
Safety Level 2 (Chance constraint)	$\Pr\left(\mathrm{c}_t(x,u,w) \le 0\right) \ge p$	Execution
Safety Level 1 (Soft constraint)	$\mathrm{c}_t(x,u,w) \le \epsilon$ or $\mathbb{E}\left[\sum_{t=0}^{T-1} \mathrm{c}_t(x,u,w)\right] \le d$	Execution
Safety Level 0	$\mathrm{c}_t(x,u,w) \le \epsilon$ or $\mathbb{E}\left[\sum_{t=0}^{T-1} \mathrm{c}_t(x,u,w)\right] \le d$	Training

II. 안전을 고려한 모델 예측 제어

모델 예측 제어는 시스템의 제약사항을 고려하며 제어함으로서 다양한 제어 분야에서 활용되었다[3,4]. 모델 예측 제어는 특정 길이의 horizon에 대해 정의된 문제를 receding horizon 방식으로 제약조건이 있는 최적화 문제를 푸는 접근법이다. 제약조건은 입력에 대한 제약을 통해 actuator의 물리적 제한을 표현하거나 시스템의 물리적 제한을 표현할 수 있고, 안전과 관련된 부분 또한 제약조건으로 표현될 수 있다. 이를 풀기 위해 목적함수를 정의하여 최적화를 진행한다. 이러한 문제는 아래와 같이 정의된다[1,5].

$$\begin{aligned} \min_{\pi_{0:H-1}} \quad & \sum_{t=0}^{H-1} l_t(x_t,u_t) + l_H(x_H) \\ & x_{t+1} = f_t(x_t,u_t,w_t), \\ \text{s.t.} \quad & u_t = \pi_t(x_t), \\ & \text{and safety constraints} \end{aligned} \tag{1}$$

비용함수는 유한한 horizon 길이 H 에 대해 매 시점의 비용의 합을 계산한다. 초기 상태 x_0가 주어졌을 때 상태 시퀀스 $x_{0:H-1} = (x_0, x_1, \dots, x_{H-1})$ 와 입력 시퀀스 $u_{0:H-1} = (u_0, u_1, \dots, u_{H-1})$, 그리고 매 시점의 비용 함수인 l_t를 사용해 누적 비용을 계산하며, 종단 시점에서의 비용을 고려하기 위해 $l_H(x_H)$를 비용함수에 더한다. 시스템 모델을 설명하는 함수인 f_t는 t 시점에서의 상태 x_t, 입력 u_t, 프로세스 잡음을 나타내는 w_t를 입력으로 받아 다음 시점의 상태인 x_{t+1}을 출력한다. 정책 π_t는 상태 x_t를 입력으로 받아 행동 u_t를 출력하는 함수이다. 비용함수를 최소화할 수 있는 최적의 행동을 출력하는 정책을 구하는 것이 목적이며, 안전한 의사결정을 수행하기 위해서는 표 1의 제약조건을 함께 고려한다. 미래의 시점을 고려하여 최적화를 하기 때문에 장기적으로 최적의 결정을 할 수 있다는 장점이 있다. 하지만 모델 예측 제어는 미래의 상태를 예측하기 위해 모델에 의존하는데, 모델이 가지고 있는 오차나 외부의 교란에 의한 불확실성으로 인해

성능이 크게 좌우되는 단점이 있다.

안전한 모델 예측 제어의 목적은 불확실성을 가지고 있는 시스템 모델에서 제약조건을 고려한 최적 정책을 찾는 것이다. 이러한 문제를 해결하기 위한 방법으로는 robust MPC, stochastic MPC, 그리고 sample-based MPC 방법이 있다. 이 방법들은 안전을 나타내는 제약조건을 어떻게 표현하는지에 따라 구분될 수 있다. 그림 1과 같이, robust MPC는 제약조건을 반드시 달성하는 hard constraint를 가지며 가장 높은 안전 단계(safety level 3)를 만족한다. Stochastic MPC는 제약조건을 특정 확률 이상으로 만족하는 chance constraint를 가지며 hard constraint 다음으로 높은 안전 단계(safety level 2)를 만족한다. 그리고 sample-based MPC는 제약조건 비용의 합에 대한 기댓값이 특정 값 이하를 만족하는 soft constraint를 가지며 비교적 낮은 안전 단계(safety level 1)를 만족한다[1].

1. Robust MPC

안전에 관한 요소를 제약조건으로 고려하기 위해 robust MPC는 제약조건을 hard constraint로 표현한다. Robust MPC는 모든 불확실한 요소에 대해 제약조건을 만족하고, 불확실성을 띄고 있는 시스템 모델에도 강건한 정책을 설계한다. 이 방식은 recursive feasibility와 제약조건 충족을 이론적으로 보장한다. Robust MPC는 가능한 불확실한 구현 중 최악의 경우에 대해 비용을 최소화하는 min-max 방식과, nominal state 주위의 bounded set 내에 존재하도록 유지하는 tube-based 방식이 있다.

1.1 Min-max 방식

Min-max 방식의 robust MPC는 외부의 교란, 혹은 불확실성이 있는 시스템에서 고려할 수 있는 최악의 경우에 대해 비용을 최소화하는 제어 입력을 찾는다[6]. 대표적으로 [8] 연구에서는 LMIs (Linear Matrix Inequalities) [7]를 min-max MPC에 적용하여 tractable state-feedback control law를 찾는 방법을 제시하였다[9]. 연구는 지속적인 교란을 가지고 있는 continuous-time 시스템에서 discontinuous feedback strategies를 사용하여 min-max 문제를 풀었다. 또한 제한된 범위의 교란 내에서 어느 정도의 steering을 보장하는 robust stability 조건을 확립하였다. [10]은 기존의 min-max MPC의 robust stability에 대한 결과를 상태와 입력에 종속된 불확실성이 있는 시스템으로 확장하여 덜 보수적인 방법으로 시스템 불확실성과 외부의 교란을 고려할 수 있게 설계하였다. 최근에는 불확실한 선형 시불변(linear time-invariant) 시스템에서 잡음이 있는 데이터에 대해 강건한 데이터 기반의 min-max MPC 방법이 제안되었다[11]. 여기서 min-max 문제는 데이터에 대해 일관되는 시스템 행렬 집합에서 최악의 경우를 최소화하는 문제이다. 이 문제를 SDP (semidefinite program) 로 재구성하여 풀었으며 이를 통해 폐루프 시스템(clased-loop system)을 안정화하고 입력과 상태 제약조건의 만족을 보장한다. 하지만 보통 개루프(open-loop)에서의 성능은 bounded set에서 최악의 교란을 가정하고 최적화되기 때문에 극단적으로 보수적인 제어로 이어지게 되고 [12], closed-loop에서는 feasibility가 향상되긴 하지만 최적화에서 infinite-dimensional 문제를 겪기 때문에 구현하기 어려운 문제가 있다[13].

1.2 Tube-based 방식

실용적인 구현을 위한 효과적인 또 다른 robust MPC의 방법으로 tube-based MPC가 있다[3]. Tube-based MPC는 MPC 최적화를 위해 nominal model을 사용하고, 모델링되지 않은 동역학을 설명하기 위해 제약조건을 좁힌다. Tube라고 불리는 nominal state 주위의 bounded set 내에 실제 상태가 존재하도록 유지하며, 이를 통해 실제 상태에 대한 제약조건 만족이 보장된다[1]. [14]는 두 예측 모델을 사용하여 강건한 예측 제어를 달성하였다. 불확실성을 포함한 nominal model과, 학습된 performance model을 이용하여 모든 제약조건을 강건하게 지키며 성능을 향상시킬 수 있다. 불확실한 시스템은 신경망을 통해 학습되며, nominal model과 performance model을 분리해서 사용함으로써 동작하는 동안 획득하는 데이터를 통해 성능 개선이 가능하다. 또한 feasibility와 제약조건 충족을 이론적으로 보장할 수 있다. [15]는 그림 3과 같이, 경로의 분포가 어떻게 움직이는지에 대한 모델을 신경망으로 설계하였고 이를 비선형 MPC에 활용하였다. 이를 위해 분포에서의 경계를 학습할 수 있는 deep quantile regression 방법을 tube-based MPC와 결합해 recursive feasibility와 제약조건을 만족함을 보장하였다. 하지만 이러한 robust MPC 방식들의 강건함은 목적함수를 최소화하는 부분에서 보수적인 모습을 보이게 된다.

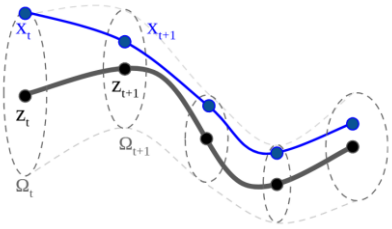

그림 3. 안전을 고려한 robust MPC 방법[15].
Fig. 3. Safety constrained robust MPC [15].

2. Stochastic MPC

Robust MPC는 안전에 대해 강건하지만 너무 보수적인 단점이 있다. 이를 개선하기 위해 stochastic MPC에서는 불확실성의 확률적 분포를 활용한다. Robust MPC는 제약조건을 반드시 만족하는 것과 달리 stochastic MPC는 제약조건을 특정 확률 이상으로 만족하는 방식이다. 최적화를 하는 매시간마다 제약조건을 충족할 확률이 특정 확률 이상이 되도록 하는 것이다. Stochastic MPC는 시스템 불확실성과 제약조건에 대한 분포를 사용한다. 이를 위해 제약조건이 특정 확률 이상으로 만족되는 chance constraint를 정의하거나 기댓값이 충족되도록 한다[16]. Stochastic MPC는 시스템의 상태와 출력을 확률 분포로 형성할 수 있는 장점이 있으며 안전과 관련된 작동 환경에서는 이 분포를 통해 동작을 안전하게 제어할 수 있다. 이러한 Stochastic MPC는 위반하는 정도를 특정 확률로 나타낸 chance-constrained 방식과, 유한 개의 샘플을 사용하여 접근하는 scenario-based 방식, 그리고 위험에 대한 지표를 이용하는 risk-sensitive 방식이 있다.

2.1 Chance-constrained 방식

Stochastic MPC에서의 제약조건은 Robust MPC에서의 hard constraint과 달리 확률적인 형태로 바뀌어 위반하는 정도를 특정 확률만큼 제한하며 제어할 수 있게 된다[17]. 하지만

robust MPC와 달리 무경계(unbounded) 불확실성을 다루게 되면 recursive feasibility를 보장하는 것이 어렵다. 이를 해결하기 위해 [18]의 연구에서는 현 시점에 충족가능한 해가 주어졌을 때, 특정 확률로 미래에도 충족가능한 해를 찾는 것을 보장할 수 있는 개루프 chance constrained MPC를 제안하였다. 또한 [19]의 연구에서는 [18]의 연구를 폐루프로 확장하여 확률적 recursive feasibility를 보장하고, 개루프의 경우에 비해 보수성을 완화하는 방법을 제시하였다. [20] 연구에서는 차량의 안전한 제어를 위해 장애물의 multi-modal 분포에 대하여 chance-constrained 방식으로 MPC를 구성하였다. 이러한 planner에서 생성된 폐루프 경로는 안전하다는 것을 증명하였으며, 자율주행 시뮬레이터에서 경로 예측 알고리즘을 사용하여 planner를 검증하였다. 또한 [21]과 [22]에서는 환경에 대한 동역학을 chance constraint로 고려하여 최적화 문제를 풀었다. 두 연구에서는 정확한 예측 모델이 있더라도 다른 차량이나 장애물의 움직임으로 인해 불확실성이 존재하기 때문에 인식을 통한 확률적 추정 결과를 활용하여 제약조건의 만족을 보장하는 PAC-MPC (Perception-Aware Chance-Constrained MPC)를 제안하였다.

Chance-constrained 방식은 주로 연속적인 도메인에서 외부의 교란에 대한 분포를 다루기 위해 확률변수가 정규분포를 따른다는 가정을 사용한다. 이는 가우시안 분포의 확률적 특성을 활용해 연산을 용이하게 하는 장점이 있지만, 이러한 가정에 맞지 않는 현실적인 경우 적용하기 어려운 한계가 존재한다.

2.2 Scenario-based 방식

Scenario-based 방식은 확률 분포에 대한 가정 없이 샘플을 이용하기 때문에 상대적으로 유연한 특징이 있다. Scenario-based 접근법은 불확실성을 가지는 시스템에서 유한 개의 샘플을 사용하여 convex한 문제를 풀며, 적은 수의 샘플 수만으로도 가능하다는 것을 [23]에서 보였다. [24]에서는 scenario-basd 방식을 선형 시불변 MPC에 적용하였고, [25]는 선형 파라메터 가변 시스템에 적용하였다. [26]는 이산적인 교란에 대해 샘플을 생성하는 stochastic MPC를 제안하였으며, 이를 통해 기존의 robust MPC 방식에 비해 덜 보수적인 제어 행동을 얻었다. [27]의 연구는 불확실한 시스템에서 환경에 대한 분포 없이 샘플만으로도 scenario 기반 최적화가 가능함을 보였다. 그리고 최적 해의 품질을 확률적으로 보장하기 위해 필요한 시나리오의 수를 유도하였고, 시뮬레이션 예시를 통해 모델과 교란의 분포에 대한 지식 없이도 적용 가능함을 입증하였다.

2.3 Risk-sensitive 방식

Chance-constrained 방식에서는 주로 충돌의 유무에 대한 확률값을 고려하는 것과 달리, risk 측정값을 사용하면 충돌의 심각도에 대한 평가가 가능하게 된다[28]. 이러한 risk metric으로는 대표적으로 VaR (Value at Risk)과 CVaR (Conditional Value at Risk)이 있다. VaR은 특정 신뢰수준에서 가지게 되는 최대값을 의미하며, CVaR은 VaR을 초과하는 값들에 대한 기대값을 의미한다. 최적화하기 어려운 VaR과 달리 convex함을 비롯하여 여러 장점을 가진 CVaR이 최적화에서 주로 사용된다[29]. CVaR은 robust 최적화에 비해서는 덜 보수적이며 chance-constraint에 비해서는 더 보수적인 특성을 가진다[28]. [30]의 연구에서는 그림 4와 같이 장애물과의 충돌을 피하기 위해 안전에 대한 제약조건을 CVaR을 통해 표현하였고 이를 MPC에 사용하였다. 또한 tractable한 방법을 구현하기 위해 CVaR constraint를 재구성하고, 샘플 평균 근사와 linearly constrained mixed integer convex program 구성을 하였다.

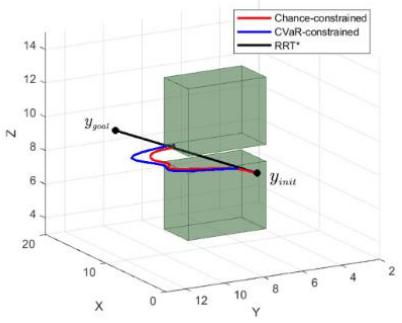

그림 4. 안전을 고려한 stochastic MPC 방법[30].
Fig. 4. Safety constrained stochastic MPC [30].

3. Sampling-based MPC

Zeroth-order 최적화라고도 불리는 sampling-based 최적화는 볼록하지 않고(non-convex) 부드럽지 않은(non-smooth) 비선형 MPC 최적화 문제를 푸는 데에 사용된다. 임의의 분포에서 샘플링을 통해 얻은 입력 시퀀스를 시스템에 적용하여 상태 시퀀스를 얻는 과정을 roll out이라고 부른다. 이를 통해 여러 입력 시퀀스들의 비용을 모두 계산한 뒤 경로 최적화를 진행한다. First-order 또는 second-order인 stochastic한 방식과 달리 sampling-based MPC는 기울기 정보를 필요로 하지 않는 gradient-free한 방식이라는 차이점이 있다. 내재된 병렬화 가능성으로 인해 GPU 활용을 가능하게 만들었고, 실시간 모션 플래닝에서 주목을 많이 받았다[31]. 예를 들어 MPPI (Model Predictive Path Integral)는 실시간 로봇 제어를 위해 온보드에서 하나의 GPU만으로 수천 개의 경로를 생성하는 데에 50Hz 이상의 속도를 낼 수 있다. 오늘날의 모델 기반 강화 학습에 관한 연구 중 sampling-based MPC를 정책으로서 활용하는 경우가 존재한다. 하지만 본 논문에서는 이러한 shooting 방식의 연구들을 sampling-based MPC에 포함하여 설명한다[32].

3.1 Random shooting 방식

[33]은 신경망으로 학습된 동역학 모델을 MPC에 활용하기 위해 random shooting [34,35] 방법을 사용하였다. Random shooting 방식은 랜덤으로 생성된 K개의 입력 시퀀스를 학습된 동역학 모델에 적용하여 해당 상태 시퀀스를 얻는다. 각 시퀀스들에 대해 비용을 계산한 뒤 누적 기대 비용이 가장 낮은 시퀀스를 선택하고 첫 번째 입력을 시스템에 적용한다.

3.2 Path integral 방식

MPPI는 제어 문제를 정보 이론 관점에서 최적 분포 매칭 문제로 재구성하였다[36]. MPPI는 특정 분포에서 샘플링된 입력 시퀀스를 통해 가중치를 계산 후 최적 제어 입력 시퀀스를 구한다. 이 최적 제어 입력 시퀀스의 첫 번째 입력을 시스템에 적용하고 위의 과정을 반복한다. 이 과정에서 여러 경로에 대한 병렬 연산을 이용하여 빠르게 최적 해를 구할 수 있다.

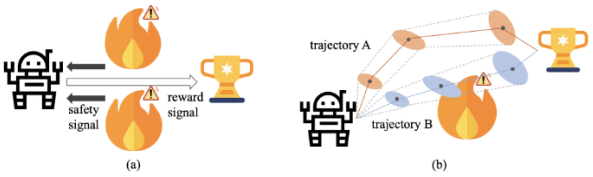

그림 5. 안전을 고려한 sample-based MPC 방법[44].
Fig. 5. Safety constrained sample-based MPC [44].

[37]은 처음으로 MPPI를 사용하여 복잡한 실제 환경인 험지 자율 주행에 성공하였다. 특히 병렬 연산의 특성을 활용하여 GPU를 통해 실시간 제어를 할 수 있게 되었다. [38]은 희소(sparse)하고 불연속적인 기울기 정보를 가진 비용 함수를 최적화할 수 있는 알고리즘을 제안하였다. 기존의 sample-based MPC에서 희소한 정보를 가진 비용 함수를 사용하게 되면 불안정하고, 예상치 못한 교란에 취약한 점을 지적하였으며, 이를 해결하기 위해 강건한 제어를 할 수 있는 tube-MPC와 결합한 Tube-MPPI를 제안하였다. [39]에서는 기계학습 알고리즘을 사용하여 동역학 모델을 학습하였다. 이 모델을 MPPI 알고리즘에 사용하여 시뮬레이션과 실제 험지 주행 태스크에서 성능을 입증하였다. [31]은 sampling-based MPC의 수렴에 대한 분석이 부족해서 하이퍼파라매터를 경험적으로 설정해야 하는 단점을 지적하였다. 이를 해결하기 위해 MPPI의 수렴성을 증명하였고 최적의 sampling covariance 설계를 제안하였다.

3.3 Cross-entropy method 방식

MPPI에서는 여러 입력 시퀀스를 고정된 분포에서 샘플링하여 경로적분을 통해 비용을 계산 후 가중치를 구해 최적해를 구한다. 이와 달리 CEM (Cross-Entropy Method) 방식은 초기 분포에서 여러 입력 시퀀스를 생성한 뒤 비용이 낮은 시퀀스를 선별 후, 이 시퀀스들을 이용하여 분포를 업데이트한다. 이 과정을 반복적으로 진행한 뒤 얻게 되는 최적 시퀀스의 첫 번째 입력을 시스템에 적용한다[40,41]. 이 방법은 MPPI와 달리 넓은 분포에서 시작하여 비용을 낮출 수 있는 최적 해로 업데이트가 진행되기 때문에 전역 해를 찾을 가능성이 높다. 하지만 반복적인 분포 업데이트 과정이 필요하기 때문에 추가적인 연산이 필요하다는 단점이 있다. CEM 방식은 비선형 모델을 활용하는 모델 기반 강화 학습에서 최적의 정책을 찾는 데에 주로 사용된다. PETS [42]는 probabilistic ensemble dynamics model을 MPC 제어에 활용하였다. iCEM [43]은 CEM이 샘플 비효율성으로 인해 실시간 제어에 활용되지 못하는 단점을 개선하였다. 시간적으로 연관된 입력과 메모리를 활용하여 더 적은 샘플을 필요로 하고 더 좋은 성능을 달성하였다. [44]의 연구는 기존의 CEM에서 안전에 대한 제약조건을 만족시키기 위해 RCE (Robust Cross-Entropy)를 제안하였다. RCE는 그림 5와 같이 각 시퀀스의 비용을 계산함과 동시에 제약조건에 대한 값도 함께 계산하여 제약조건이 최소가 되는 시퀀스를 선별하여 분포를 업데이트한다.

III. 안전을 고려한 강화 학습

강화 학습을 이용한 방법은 정책이 환경에 행동을 취한 뒤 보상을 받으며 행동을 강화하는 방법이다. 이 방법은 정책이 최종적으로 받을 수 있는 보상의 합인 리턴이 최대가 되는 방향으로 정책을 최적화하기 때문에 전역적으로 최적화가 되어 다양한 태스크에서 높은 성능을 보이고[45,46] 일부 부분에서는 사람보다 우수한 성능을 보이고 있다[47,48]. 하지만 기존의 성공들은 시뮬레이션 환경과 같이 시행착오에 대한 피해가 크지 않은 상황에서 이루어졌으며, 현실의 환경과는 차이가 있다. 동작하는 장치의 충돌, 또는 움직이는 물체와의 피해를 최소화해야 하는 현실에서 이러한 정책을 학습시키기 위해서는 시행착오를 겪으며 행동을 강화해야 하기 때문에 위험한 행동에 빠지게 되는 문제가 있다. 만약 안전하게 학습하기 위해 시행착오를 겪는 탐색을 줄이게 된다면 보수 적인 정책으로 이어지게 되어 최종 성능이 기대에 미치지 못하게 된다. 이러한 문제를 해결하기 위해 안전을 고려한 강화 학습은 문제를 CMDP (Constrained Markov Decision Process)로 정의하여 형식화하였다[2]. CMDP는 기존의 MDP (Markov Decision Process)에 추가로 제약조건에 관한 누적 비용의 합을 특정 값 이하로 유지하는 방법이 주로 사용된다.

이러한 강화 학습 방법은 환경과 상호작용을 하는 정도를 기준으로 모델 프리 강화 학습, 모델 기반 강화 학습, 그리고 오프라인 강화 학습으로 분류할 수 있다. 모델 프리 강화 학습은 환경에 대한 모델 없이 환경과의 상호작용을 통해 정책을 학습하는 방법으로, 실제 환경으로부터 획득하는 데이터가 많이 필요하기 때문에 위험에 많이 노출되는 단점이 있다. 반면 모델 기반 강화 학습은 환경에 대한 모델을 활용하여 정책을 학습하는 방법으로, 상대적으로 적은 데이터만으로도 정책을 학습할 수 있다는 장점이 있다. 오프라인 강화 학습은 환경과의 상호작용 없이 사전에 획득한 데이터만으로 정책을 학습하는 방법이며 학습 과정 동안 위험에 노출될 가능성을 완전히 배제할 수 있다.

1. 모델 프리 강화 학습

모델 예측 제어에서는 환경 모델을 사용하여 미래 상태를 예측하지만 실제 현실에서 정확한 모델을 구현하는 것은 많은 어려움이 있다. 이러한 어려움을 겪는 모델 예측 제어와 달리 환경을 나타내는 모델 없이 실제 환경과의 상호작용을 통해 시행착오를 거쳐가며 정책을 학습하는 방법이 모델 프리 강화 학습이다. 이 방법을 통해 고차원, 비선형의 공간에서 환경에 대한 모델 없이도 미래의 기대 보상의 합을 최대화하게 되는 최적 정책 학습이 가능하다. 이러한 모델 프리 강화 학습은 실제 환경과의 상호작용을 하며 상태, 행동, 다음 상태, 보상 등으로 이루어진 데이터를 수집한다. 이후 이 데이터를 통해 상태를 입력 받아 보상을 최대화하는 행동을 만들어내는 최적 정책을 학습하게 된다. 모델 프리 강화 학습을 안전에 적용한 방법은 라그랑지안(Lagrangian) 방식과 projection 방식으로 분류할 수 있다.

1.1 라그랑지안 방식

라그랑지안 방식은 제약조건이 있는 최적화 문제를 푸는 대표적인 접근법이다. 제약조건이 있는 문제를 듀얼 변수(dual variable)인 라그랑주 승수(Lagrange multiplier)를 통해 제약조건이 없는 (unconstrained) 문제로 변형한다. 라그랑주 승수는 목적함수에서 페널티 계수의 역할을 하여 제약조건을 만족하는 해를 만들어낸다[49-51]. [52] 연구에서는 라그랑지안을

그림 6. Projection 방식 제약조건이 있는 정책 최적화 과정[58].
Fig. 6. Update procedures for the projection-based constrained policy optimization [58].

이용하여 제약조건을 만족하는 정책으로 수렴하는 프라이멀-듀얼(primal-dual) 방법을 제안하였다. 제약조건을 포함한 문제를 제약조건이 없는 max-min 최적화 문제의 형태로 변형하여 정책이 제약조건을 달성하도록 하였다. 또한 [53] 논문에서는 안전한 강화 학습을 위해 safety gym 벤치마크를 제시함과 동시에 PPO (Proximal Policy Optimization) [54], TRPO (Trust Region Policy Optimization) [55]와 같은 제약조건이 없는 모델 프리 강화 학습 알고리즘을 제약조건이 있는 형태로 풀 수 있는 PPO-Lag, TRPO-Lag를 제시하였다. [56]의 연구에서는 액터-크리틱(actor-critic) 방식에 라그랑지안 기법을 적용하였다. 이 연구에서 제안한 RCPO (Reward Constrained Policy Optimization)는 제약조건을 보상함수에 페널티 시그널의 형태로 포함한다. 또한 RCPO는 multi-timescale 접근법을 사용하는데, 정책은 빠른 timescale로 업데이트 되는 반면 라그랑지안 승수는 느린 timescale로 업데이트가 된다. 하지만 기존의 연구들은 라그랑지안 승수법을 이용해 학습하는 과정에서 제약조건을 위반하는 경우가 발생할 수 있는데, PID-Lag [49]에서는 PID 제어를 통해 학습 과정에서 발생하는 제약조건 위반을 억제하는 방법을 제안하였다. 제어 관점에서 라그랑지안 승수 업데이트가 integral control과 같이 행동하기에 이 연구에서는 학습 알고리즘을 동역학 시스템으로 간주하고 proportional control과 derivative control을 도입한다. 이를 통해 제약조건을 만족하는 성능을 개선하였으며 하이퍼파라메터에 강건함을 보였다. 이러한 라그랑지안 기반의 알고리즘들은 연산량이 많이 필요한 projection 방식의 알고리즘들에 비해 비슷하거나 나은 성능을 보이기도 하였다[49]. 하지만 이러한 라그랑지안을 이용한 방식은 라그랑주 승수의 초기값과 학습률에 민감하다 [51]. 학습하는 동안 변동성이 크며 학습하는 과정에서의 정책은 제약조건을 일관적으로 만족하지 않는다. 또한 라그랑지안 승수가 느린 time-scale로 풀리는 것이 최적화하는 것에 어려움을 주게 된다.

1.2 Projection 방식

Projection 방식은 라그랑지안 방식과는 달리 학습하는 모든 과정동안 제약조건을 만족한다. 대표적으로 one-stage projection 방식인 CPO (Constrained Policy Optimization) [57]는 그림 6과 같이 강화 학습의 학습 과정에서 안전한 탐색을 할 수 있도록 제약 조건을 유지하며 보상을 최대화하는 정책을 학습하는 기법을 제안하였다. 이는 unconstrained MDP의 TRPO [55]에서 trusted region의 개념과 유사하지만, 근사 오차, 혹은 초기값이 안전하지 못한 경우 trust region과 제약조건을 만족하는 constraint set이 교차하지 않게 되어 CPO는 infeasible하게 되는 문제가 존재한다. 이를 해결하기 위해 two-stage projection을 하는 PCPO (Projection-based CPO) [58]가 제안되었다. 이 방법은 첫 번째 단계에 보상을 향상시키는 방향으로 정책을 업데이트 한다. 다음으로 정책을 constraint set으로 projection함으로서 제약조건 위반을 조정한다. 이를 통해 학습하는 매 업데이트 마다 정책의 안전성을 유지할 수 있다. 하지만 CPO와 PCPO는 고차원의 Hessian 행렬의 역함수를 구해야 하는데 quadratic approximation을 필요로 해서 연산이 어렵다는 단점이 있다. 이러한 second-order 문제를 해결하기 위해 FOCOPS (First Order Constrained Optimization in Policy Space) [59]는 non-parametric space에서 제약조건을 고려한 최적화를 수행하며 first-order gradient를 구한다. 이 접근법은 학습하는 동안 제약 조건을 위반하는 최악의 경우에 대한 상한의 근사를 알 수 있으며, first-order이기 때문에 구현하기 쉬운 장점이 있다. 하지만 모델 프리 강화 학습처럼 시행착오를 겪는 방식은 위험 한 행동을 겪은 뒤 그 행동을 penalize하기 때문에 안전하게 제어 정책을 학습하는 것이 어렵다는 단점이 존재한다[60].

2. 모델 기반 강화 학습

모델 프리 강화 학습은 명시적인 모델을 정의할 필요 없이 실제 환경과 상호작용을 통해 정책을 학습할 수 있다는 장점이 있지만, 정책을 학습하는 동안 불안전한 정책을 실제 환경에 적용하게 되어 위험에 노출되는 단점이 존재한다. 이에 반해 실제 환경을 표현하는 모델을 정의하여 샘플 효율성이 증가한 모델 기반 강화 학습은 보다 안전한 학습이 가능하다. 모델 기반 강화 학습은 MDP 동역학을 설명하는 모델을 사용하고 전역 해를 찾아내는 방법이다[61,62]. 실제 환경과 유사하게 동작하는 모델을 이용하여 정책을 학습하는 데에 도움이 되는 샘플을 생성함으로서 실제 환경과의 상호작용을 줄일 수 있다. 이러한 모델 기반 강화 학습은 먼저 임의의 초기 정책이나 사용자가 개입한 정책을 이용하여 환경의 데이터를 수집 후 모델을 학습하는 것으로 시작한다. 이후 학습된 모델을 바탕으로 정책이 누적 보상을 최대화할 수 있도록 업데이트를 진행하고, 업데이트된 정책을 통해 다시 환경에서 데이터를 모으는 과정을 반복한다[62]. 이러한 과정을 통해 강화 학습이 탐색하는 과정 동안 환경에 위험하게 노출되는 횟수를 줄이게 되어 보다 안전에 도움이 될 수 있다는 장점이 있다. CMDP로 정의된 모델 기반 강화 학습은 background planning과 online planning으로 구분할 수 있다.

2.1 Background planning

Background planning은 학습하는 과정동안 안전한 최적 정책을 학습한다. [63]은 다른 값으로 초기화된 여러 동역학 모델을 학습하고, 앙상블 기법을 통해 모델 불확실성과 데이터 불확실성을 나타낸다. SAMBA [64]는 PILCO (Probabilistic Inference for Learning Control) [65] 기반으로 학습된 가우시안

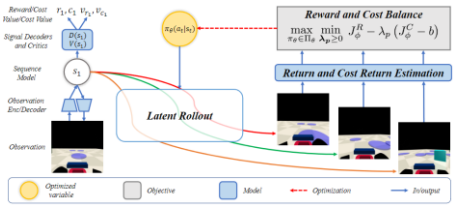

그림 7. 안전을 고려한 모델 기반 강화 학습 방법[66].
Fig. 7. Safe model-based reinforcement learning process [66].

프로세스(gaussian process) 동역학 모델을 이용하여 안전을 표현하는 제약조건 함수를 다룬다. SafeDreamer [66]는 그림 7과 같이 월드 모델(world model)을 이용하여 경로를 생성하고, 이를 통해 안전한 정책을 학습한다.

2.2 Online planning

Online planning은 실행하는 동안 경로 최적화를 통해 최적 입력을 결정한다. Sample-based MPC에서 언급된 RCE [44]와 같은 방법들도 이 분류에 포함되지만, 그들은 학습된 정책이 아닌 MPC로 제어한다는 차이가 있다. MDP 문제에서 학습된 정책으로 online planning을 수행하는 POLO (Plan Online Learn Offline) [67], TD-MPC [68]과 유사하게, 본 단락에서는 CMDP 문제에서 학습된 정책을 이용하여 경로 최적화를 수행하는 방법을 소개한다.

[69]은 학습된 모델과 종단 가치 함수(terminal value function)를 통해 H-step lookahead를 하는 LOOP (Learning Off-policy with Online Planning) 방법을 제시하였다. 이는 MPC와 유사하지만 LOOP는 parametrized된 정책을 모델 프리 off-policy 알고리즘을 이용하여 학습한다. 학습된 정책은 실행 단계에 매 시점마다 동역학 모델을 이용해 H-step만큼 rollout을 하여 보상이 최대가 되는 행동 시퀀스를 찾아낸다. 이러한 모델을 이용한 online planning과 모델 프리 off-policy 학습의 장점을 이용하여 샘플 효율성과 연산 효율성을 얻게 된다. 또한 actor-divergence 문제를 지적하며 ARC (Actor Regularized Control)라는 경로 최적화 방법을 제시하였다. 이를 통해 CMDP와 동일한 제약 조건으로 목적함수를 정의함으로써 매 rollout마다 안전을 고려한 행동이 선택될 수 있도록 하는 것이 가능하다. [70]은 제약조건이 있는 모델 기반 정책 최적화와 안전 필터(safety filter)를 결합한 방법인 CASE를 제시하였다. Parametrized된 정책과 크리틱을 학습하기 위해 액터-크리틱 방법을 사용하였으며, 이 학습된 정책의 행동을 고려한 safety filter를 설계하여 online planning을 수행한다. 특히 모델 기반 엑터-크리틱의 planning에서 편향이 낮은 특성을 이용하여 제약조건 위반을 줄이게 되었다.

3. 오프라인 강화 학습

기존의 안전을 고려한 강화 학습 방식들은 학습하는 동안 안전을 보장할 수 있는 방법을 찾기 위해 발전하였지만, 전체 학습 단계 동안 안전을 보장하지는 못하였다. 하지만 오프라인 강화 학습은 사전 수집된 데이터로부터 학습하기 때문에 환경과의 상호작용으로 인한 위험을 완전히 배제한다. 오프라인 강화 학습은 추가적인 환경과의 상호작용 없이 수집된 데이터 만을 통해 효과적인 정책을 학습하는 기법이다[71]. 환경과의 상호작용 없이 사전 확보된 데이터만을 통해 정책을 학습하기 때문에 학습 중 위험에 노출되는 기존의 문제는 사라지고, 학습이 끝난 뒤의 실행 단계에서의 안전성만 고려한다. 이러한 오프라인 강화 학습에서의 주된 어려움으로 distributional shift 문제가 있다. 이는 정책을 학습하기 위해 수집된 데이터의 분포가 이후 정책을 평가하기 위한 데이터의 분포와 크게 떨어져 있을 때 발생한다. 안전을 고려한 강화 학습에서는 이러한 오프라인 강화 학습의 문제를 개선하기 위해 Q-learning 기반의 방법, DICE (Distribution Correction Estimation) 기반의 방법, 순차 모델링(sequential modeling) 기반의 방법, 그리고 확산 모델(diffusion model) 기반의 방법이 있다 [72].

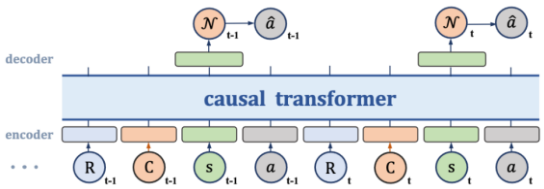

그림 8. 안전을 고려한 오프라인 강화 학습 방법[79].
Fig. 8. Safe offline reinforcement learning architecture [79].

3.1 Q-learning 기반

CPQ (Constraints Penalized Q-learning) [73]은 BCQ (Batch-Constrained Q-learning) [74]를 기반으로 하여 안전한 오프라인 강화 학습 방식을 제안하였다. Out-of-distribution과 불안전한 입력에 높은 제약조건 비용 값을 할당하며 안전한 상태-입력으로만 Q-가치 함수를 라그랑지안 방식을 통해 업데이트 한다. 하지만 가치 함수가 왜곡되고 일반화 성능이 낮은 단점이 있다[75].

3.2 DICE 기반

이전의 Q-learning 기반의 오프라인 강화 학습은 불확실한 상태에 대한 과대평가를 줄이기 위해 하이퍼파라메터를 정교하게 설정해야 하는 단점이 존재한다. 하지만 [76]에서는 데이터의 분포와 최적 정책의 분포 간의 차이를 줄일 수 있는 분포 비율을 추정한다. 이를 통해 이전의 Q-함수나 정책의 공간에서 최적화를 하던 것과 달리 stationary distribution 공간에서 최적화가 가능하게 되었다. COptiDICE [77]는 OptiDICE [76]에 안전에 관한 제약조건을 추가하였다. State-action stationary distribution을 직접 최적화한 뒤 stationary distribution으로부터 importance-weighted behavioral cloning을 통해 정책을 추출한다.

3.3 순차 모델링 기반

[78]의 DT (Decision Transformer)는 오프라인 강화 학습 문제를 순차 모델링을 통해 해결하는 접근법이다. 기존의 방식들이 하나의 상태를 받는 정책 $\pi(a|s)$를 파라메터화 하던 것과 달리 DT는 보상, 상태, 행동들의 시퀀스를 입력 토큰으로 받고 예측된 행동을 출력한다. [79]은 DT 기반의 방법으로서 안전에 관한 제약조건을 고려한 CDT (Constrained Decision Transformer)를 제안하였다. CDT는 그림 8과 같이 DT에 stochastic policy with entropy regularization과 data augmentation by return relabeling을 적용하여 안전성과 강건성을 개선하였다. 이를 통해 CDT는 추가적인 학습 없이 다른 제약조건 임계값에서도 zero-shot 적응(adaptation)이 가능하다. 하지만 Transformer 구조로 인해 연산량이 큰 것과 안전에 대한 이론적인 보장이 부족한 단점이 있다.

3.4 확산 모델 기반

Diffuser [80]은 확산 모델[81] 기반으로 경로 최적화를 하는 방법이다. 기존의 모델 기반의 planner는 미래의 시점을 autoregressive 방식으로 예측하는 반면, Diffuser는 모든 미래의 시점을 동시에 예측한다. 이러한 확산 모델의 경로 샘플링 과정에서 reward guidance를 통해 생성되는 경로에 제약조건을 주는 것이 가능하다. [82]에서 제안한 TREBI (Trajectory-based Real-time Budget Inference) 방법에서 Diffuser를 백본 모델 삼아서 행동하는 정책의 경로 분포는 학습하는 동안 오프라인

데이터셋을 통해 근사 되고, real-time budget을 위한 adaptive response는 추론 단계에서 budget-related trajectory planning을 함으로써 달성한다. Episodic 보상과 비용에 대한 추론의 에러 경계를 증명하여 TREBI의 성능 보장을 입증하였다. 또한 autoregressive 방식으로 경로를 생성하는 기존의 방법들은 long-term prediction의 rollout error 문제를 겪는 반면 [83], TREBI는 경로 전체를 동시에 생성하여 해당 문제를 해소한다.

IV. 강화 학습과 제어 이론을 결합한 방법

안전을 고려한 강화 학습은 정책을 학습하는 탐색 과정에서 제약조건을 만족하는 방향으로 발전되고 있다. 하지만 대부분의 안전을 고려한 강화 학습은 CMDP [2]로 문제가 정의되어 있으며, 제약조건 비용의 합이 특정 임계값보다 낮도록 하는 것을 제약조건으로 두고 있다. 그러나 이러한 방법들은 정책이 수렴하기 전까지 안전을 보장하지 못하는 단점이 있으며, 매 시점마다 안전을 보장하지는 못하는 한계가 있다. 이를 극복하기 위해 제어 이론의 개념을 결합하여 매 시점마다 hard constraint를 만족하는 방법들이 제시되었다.

모델 프리 강화 학습에서 [60,84]는 기존의 CMDP문제가 제약조건 위반 횟수를 0으로 만들 수 없음을 지적하며, 새로운 barrier certificate를 이용한 hard constraint 방법을 제안하였다. 이를 통해 거의 0에 근접한 제약조건 위반을 달성하였다. 모델 기반 강화 학습에서 [85]과 [86]는 각각 barrier certificate와 barrier function을 통해 hard safety constraint를 적용하였다. 특히 [86]는 생성 모델 기반의 soft barrier function을 통해 hard safety chance constraint를 적용하였다. 또한 새로운 bi-level 최적화 문제를 구성하여 생성모델, soft barrier function, 그리고 정책을 함께 학습할 수 있게 되었다. 이를 통해 CMDP 기반의 방법에 비해 안전한 비율을 향상시키고 안전한 확률에 대한 하한을 제공하였다. 또한 [87]는 제어 이론의 Lyapunov function을 이용하여 안정성을 보장할 수 있는 방법을 제시하였다. 오프라인 강화 학습에서는 [88]의 연구를 예로 들 수 있다. 제어 이론의 reachability 분석을 통해, hard safety constraint는 오프라인 데이터셋이 주어졌을 때 largest feasible region을 식별하는 문제로 변할 수 있음을 발견했다. 상태의 feasibility를 특징짓기 위해 Hamilton-jacobi reachability로 오프라인 데이터에서 largest feasible region을 정의하였다. 그림 9와 같이 feasibility-dependent 한 목적함수를 구성 후 학습하였고, hard safety constraint를 만족하는 방법을 제안하였다.

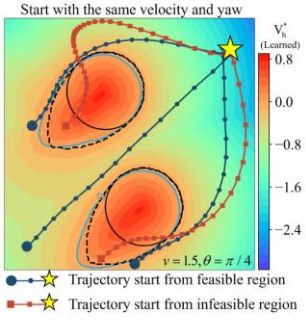

그림 9. Feasibility 를 결합한 오프라인 강화 학습 방법[88].
Fig. 9. Safe offline reinforcement learning with feasibility [88].

V. 결론

본 논문에서는 안전을 고려한 의사 결정 방법으로 모델 예측 제어와 강화 학습, 그리고 강화 학습과 제어 이론의 결합 방법을 소개하였다. 안전을 고려한 모델 예측 제어 방법은 모델을 기반으로 미래 상태를 예측하여 제약조건을 고려한 최적의 방법을 찾기에 적합하지만, 모델에 의존적이라는 한계로 인해 불확실한 모델과 제약조건을 다루기 위한 연구들이 진행되었다. 안전을 고려한 강화 학습은 CMDP 형태의 문제로 정의되어 제약조건을 고려한 정책 최적화를 통해 최종 정책이 안전하게 동작할 수 있는 방법이다. 태스크의 목적에 대한 보상을 최대화하는 측면에서 높은 성능을 보일 수 있지만, 안전하게 동작하기 위한 제약조건을 위반하는 경우가 많으며, 대부분 안전에 대한 보장이 부족한 것이 단점이다. 이 문제를 해결하기 위해 강화 학습에 제어 이론의 개념을 결합하여 매 시점 별 안전을 보장하거나 수렴한 정책의 안전성을 이론적으로 보장할 수 있는 방법들이 연구되고 있다. 그러나 현재의 방법들은 특정 가정에 의존하거나 여전히 실제 환경이 아닌 시뮬레이션 환경에서 검증되는 한계가 있다. 따라서 현실 환경에서 안전하고 신뢰할 수 있는 의사 결정을 위해서는 제어 이론의 강한 검증력과 강화 학습의 유연한 성능을 결합하여 학습과 실행 전 과정에서 안전성을 확보할 수 있는 새로운 방법이 요구된다.

REFERENCES

[1] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, “Safe learning in robotics: From learning-based control to safe reinforcement learning,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, pp. 411-444, 2022.

[2] E. Altman, *Constrained Markov decision processes,* Routledge, 2021.

[3] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model predictive control: theory, computation, and design*, Madison, WI: Nob Hill Publishing, 2017.

[4] E. F. Camacho and C. Bordons, *Constrained model predictive control*, Springer London, 2007.

[5] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert, “Constrained model predictive control: Stability and optimality,” *Automatica*, vol. 36, no. 6, pp 789-814, 2000.

[6] A. Bemporad, and M. Morari, “Robust model predictive control: A survey,” *In Robustness in identification and control*, Springer London, pp. 207-226, 2007.

[7] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Society for industrial and applied mathematics, 1994.

[8] M. V. Kothare, V. Balakrishnan, and M. Morari, “Robust constrained model predictive control using linear matrix inequalities,” *Automatica*, vol. 32, no. 10, pp. 1361-1379, 1996.

[9] F. Fontes and L. Magni, “Min-max model predictive control of nonlinear systems using discontinuous feedbacks,” *IEEE Transactions on Automatic Control*, vol. 48, no. 10, pp. 1750-1755, 2003.

[10] D. Limón, T. Alamo, F. Salas, and E. F. Camacho, “Input to state stability of min–max MPC controllers for nonlinear systems with bounded uncertainties,” *Automatica*, vol. 42, no. 5, pp. 797-803,

2006.
[11] Y. Xie, J. Berberich, and F. Allgöwer, "Data-driven min-max MPC for linear systems," *Proc. of the 2024 American Control Conference*, pp. 3184-3189, 2024.
[12] D. M. Raimondo, D. Limón, M. Lazar, L. Magni, and E. F. Camacho, "Min-max model predictive control of nonlinear systems: A unifying overview on stability," *European Journal of Control*, vol. 15, no. 1, pp. 5-21, 2009.
[13] D. Q. Mayne, "Robust and stochastic mpc: Are we going in the right direction?," *IFAC-PapersOnLine*, vol. 48, no. 23, pp. 1-8, 2015.
[14] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216-1226, 2013.
[15] D. D. Fan, A. Agha-mohammadi, and E. A. Theodorou, "Deep learning tubes for tube mpc," *Robotics: Science and Systems*, 2020.
[16] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Systems Magazine*, vol. 36, no. 6, pp. 30-44, 2016.
[17] M. Farina, L. Giulioni, and R. Scattolini, "Stochastic linear model predictive control with chance constraints–a review," *Journal of Process Control*, vol. 44, pp. 53-67, 2016.
[18] M. Ono, "Joint chance-constrained model predictive control with probabilistic resolvability," *Proc. of the 2012 American Control Conference*, pp. 435-441, 2012.
[19] M. Ono, "Closed-loop chance-constrained MPC with probabilistic resolvability," *Proc. of 2012 51st IEEE Conference on Decision and Control*, pp. 2611-2618, 2012.
[20] K. Ren, C. Chen, H. Sung, H. Ahn, I. Mitchell, and M. Kamgarpour, "Safe chance-constrained model predictive control under gaussian mixture model uncertainty," arXiv preprint arXiv:2401.03799, 2024.
[21] A. D. Bonzanini, A. Mesbah, and S. Di Cairano, "Perception-aware chance-constrained model predictive control for uncertain environments," *Proc. of the 2021 American Control Conference*, pp. 2082-2087, 2021.
[22] A. D. Bonzanini, A. Mesbah, and S. Di Cairano, "Multi-stage perception-aware chance-constrained MPC with applications to automated driving," *Proc. of the 2022 American Control Conference*, pp. 1697-1702, 2022.
[23] G. C. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on automatic control*, vol. 51, no. 5, pp. 742-753, 2006.
[24] M. Prandini, S. Garatti, and J. Lygeros, "A randomized approach to stochastic model predictive control," *Proc. of 2012 51st IEEE Conference on Decision and Control*, pp. 7315-7320, 2012.
[25] G. C. Calafiore and L. Fagiano, "Stochastic model predictive control of LPV systems via scenario optimization," *Automatica*, vol. 49, no. 6, pp. 1861-1866, 2013.
[26] D. Bernardini and A. Bemporad, "Scenario-based model predictive control of stochastic constrained linear systems," *Proc. of 48th IEEE Conference on Decision and Control*, pp. 6333-6338, 2009.
[27] F. Micheli and J. Lygeros, "Scenario-based stochastic mpc for systems with uncertain dynamics," *Proc. of the 2022 European Control Conference*, pp. 833-838, 2022.
[28] P. Tooranjipour, B. Kiumarsi, and H. Modares, "Risk-aware Stochastic MPC for Chance-constrained Linear Systems," *IEEE Open Journal of Control Systems*, vol. 3, pp. 282-294, 2024.
[29] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21-42, 2000.
[30] A. Hakobyan, G. C. Kim, and I. Yang, "Risk-aware motion planning and control using CVaR-constrained optimization," *IEEE Robotics and Automation letters*, vol. 4, no. 4, pp. 3924-3931, 2019.
[31] Z. Yi, C. Pan, G. He, G. Qu, and G. Shi, "CoVO-MPC: Theoretical Analysis of Sampling-based MPC and Optimal Covariance Design," *Proc. of Machine Learning Research*, vol. 242, pp. 1122-1135, 2024.
[32] T. Wang, X. Bao, I. Clavera, J. Hoang, Y. Wen, E. Langlois, S. Zhang, G. Zhang, P. Abbeel, and J. Ba, "Benchmarking model-based reinforcement learning," arXiv preprint arXiv:1907.02057, 2019.
[33] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," *Proc. of the 2018 IEEE International Conference on Robotics and Automation*, pp. 7559-7566, 2018.
[34] A. G. Richards, "Robust constrained model predictive control", PhD thesis, Massachusetts Institute of Technology, 2005.
[35] A. V. Rao, "A survey of numerical methods for optimal control," *Advances in the astronautical Sciences*, vol. 135, no. 1, pp. 497-528, 2010.
[36] G. Williams, A. Aldrich, and E. Theodorou, "Model predictive path integral control using covariance variable importance sampling," arXiv preprint arXiv:*1509.01149*, 2015.
[37] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," *Proc. of the 2016 IEEE International Conference on Robotics and Automation*, pp. 1433-1440, 2016.
[38] G. Williams, B. Goldfain, P. Drews, K. Saigol, J. M. Rehg, and E. A. Theodorou, "Robust sampling based model predictive control with sparse objective information," *Robotics: Science and Systems*, 2018.
[39] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic mpc for model-based reinforcement learning," *Proc. of the 2017 IEEE International Conference on Robotics and Automation*, pp. 1714-1721, 2017.
[40] S. Mannor, R. Y. Rubinstein, and Y. Gat, "The cross entropy method for fast policy search," *Proc. of the 20th International Conference on Machine Learning*, pp. 512-519, 2003.
[41] Z. I. Botev, D. P. Kroese, R. Y. Rubinstein, and P. L'Ecuyer, *The Cross-entropy Method for Optimization,* In Handbook of statistics, vol. 31, pp. 35-59, Elsevier, 2013.
[42] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[43] C. Pinneri, S. Sawant, S. Blaes, J. Achterhold, J. Stueckler, M. Rolinek, and G. Martius, "Sample-efficient cross-entropy method for real-time planning," *Conference on Robot Learning, PMLR*, pp. 1049-1065, 2021.
[44] Z. Liu, H. Zhou, B. Chen, S. Zhong, M. Hebert, and D. Zhao, "Constrained model-based reinforcement learning with robust

cross-entropy method," arXiv preprint arXiv:2010.07968, 2020.

[45] J. G. Hwang and J. G. Park, "A base station placement method for high-precision positioning using reinforcement learning," *Journal of Institute of Control, Robotics and Systems (in Korean),* vol. 29, no. 11, pp. 836-840, Nov. 2023.

[46] K. Lee, S. Baek, P. Jung, T. H. Kim, and J. H. Jeon, "Cooperative multi-agent reinforcement learning for multiple anti-aircraft target surveillance," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 6, pp. 587-595, Jun. 2024.

[47] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, ... and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484-489, 2016.

[48] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, ... and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354-359, 2017.

[49] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by pid lagrangian methods," Proc. of the *International Conference on Machine Learning, PMLR*, pp. 9133-9143, 2020.

[50] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic press, 2014.

[51] Y. Liu, A. Halev, and X. Liu, "Policy learning with constraints in model-free reinforcement learning: A survey," *The 30th International Joint Conference on Artificial Intelligence*, 2021.

[52] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, vol. 18, no. 167, pp. 1-51, 2018.

[53] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," arXiv preprint arXiv:1910.01708, 2019.

[54] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[55] J. Schulman, "Trust Region Policy Optimization," arXiv preprint arXiv:1502.05477, 2015.

[56] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *Proc. of the International Conference on Learning Representations*, 2019.

[57] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 22-31, 2017.

[58] T. Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," *Proc. of the International Conference on Learning Representations*, 2020.

[59] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15338-15349, 2020.

[60] H. Ma, C. Liu, S. E. Li, S. Zheng, and J. Chen, "Joint synthesis of safety certificate and safe control policy using constrained reinforcement learning," *Learning for Dynamics and Control Conference, PMLR*, pp. 97-109, 2022.

[61] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, 1-118, 2023.

[62] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM Sigart Bulletin*, vol. 2, no. 4, 160-163, 1991.

[63] M. A. Zanger, K. Daaboul, and J. M. Zöllner, "Safe continuous control with constrained model-based policy optimization," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3512-3519, 2021.

[64] A. I. Cowen-Rivers, D. Palenicek, V. Moens, M. A. Abdullah, A. Sootla, J. Wang, and H. Bou-Ammar, "Samba: Safe model-based & active reinforcement learning," *Machine Learning*, vol. 111, no. 1, pp. 173-203, 2022.

[65] M. Deisenroth and C. E. Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," *Proc. of the 28th International Conference on Machine Learning,* pp. 465-472, 2011.

[66] W. Huang, J. Ji, B. Zhang, C. Xia, and Y. Yang, "Safe dreamerv3: Safe reinforcement learning with world models," *Proc. of the International Conference on Learning Representations*, 2024.

[67] K. Lowrey, A. Rajeswaran, S. Kakade, E. Todorov, and I. Mordatch, "Plan online, learn offline: Efficient learning and exploration via model-based control," *Proc. of the International Conference on Learning Representations*, 2019.

[68] N. Hansen, X. Wang, and H. Su, "Temporal difference learning for model predictive control," *Proc. of the International Conference on Machine Learning, PMLR*, 2022.

[69] H. Sikchi, W. Zhou, and D. Held, "Learning off-policy with online planning," *Conference on Robot Learning, PMLR*, pp. 1622-1633, 2022.

[70] A. Agha, B. Kayalibay, A. Mirchev, P. van der Smagt, and J. Bayer, "Exploring under constraints with model-based actor-critic and safety filters," *8th Annual Conference on Robot Learning,* 2024.

[71] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv preprint arXiv:2005.01643, 2020.

[72] Z. Liu, Z. Guo, H. Lin, Y. Yao, J. Zhu, Z. Cen, H. Hu, W. Yu, T. Zhang, J. Tan and D. Zhao, "Datasets and benchmarks for offline safe reinforcement learning," *Journal of Data-centric Machine Learning Research*, vol.1, no. 12, 2024.

[73] H. Xu, X. Zhan, and X. Zhu, "Constraints penalized q-learning for safe offline reinforcement learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, pp. 8753-8760, 2022.

[74] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 2052-2062, 2019.

[75] J. Li, X. Zhan, H. Xu, X. Zhu, J. Liu, and Y. Q. Zhang, "When data geometry meets deep function: Generalizing offline reinforcement learning," *Proc. of the International Conference on Learning Representations*, 2022.

[76] J. Lee, W. Jeon, B. Lee, J. Pineau, and K. E. Kim, "Optidice: Offline policy optimization via stationary distribution correction estimation," *Proc. of the International Conference on Machine Learning, PMLR,* pp. 6120-6130, 2021.

[77] J. Lee, C. Paduraru, D. J. Mankowitz, N. Heess, D. Precup, K. E. Kim, and A. Guez, "Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation," *Proc. of the International Conference on Learning Representations*,

2022.
[78] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15084-15097, 2021.
[79] Z. Liu, Z. Guo, Y. Yao, Z. Cen, W. Yu, T. Zhang, and D. Zhao, "Constrained decision transformer for offline safe reinforcement learning," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 21611-21630, 2023.
[80] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 9902-9915, 2023.
[81] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840-6851, 2020.
[82] Q. Lin, B. Tang, Z. Wu, C. Yu, S. Mao, Q. Xie, X. Wang, and D. Wang, "Safe offline reinforcement learning with real-time budget constraints," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 21127-21152, 2023.
[83] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[84] Y. Yang, Y. Jiang, Y. Liu, J. Chen, and S. E. Li, "Model-free safe reinforcement learning through neural barrier certificate," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1295-1302, 2023.
[85] Y. Luo and T. Ma, "Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25621-25632, 2021.
[86] Y. Wang, S. S. Zhan, R. Jiao, Z. Wang, W. Jin, Z. Yang, Z. Wang, C. Huang, and Q. Zhu, "Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments," *Proc. of the International Conference on Machine Learning, PMLR*, pp. 36593-36604, 2023.
[87] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in neural information processing systems*, vol. 30, 2017.
[88] Y. Zheng, J. Li, D. Yu, Y. Yang, S. E. Li, X. Zhan, and J. Liu, "Safe offline reinforcement learning with feasibility-guided diffusion model," *Proc. of the International Conference on Learning Representations*, 2024.

함 형 찬

2021 년 울산과학기술원 전기 전자컴퓨터공학부 졸업. 관심분야는 로봇 제어, 환경인식.

안 희 진

2012 년 서울대학교 기계항공공학부 졸업 (공학사). 2014 년, 2018 년 메사추세츠 공과대학 기계공학과 졸업(공학 석사, 공학 박사). 2022 년~현재 한국과학기술원 전기및전자공학부 조교수. 관심분야는 안전 보장 제어, 최적화 기반 제어, 및 지능형 교통시스템으로의 응용.