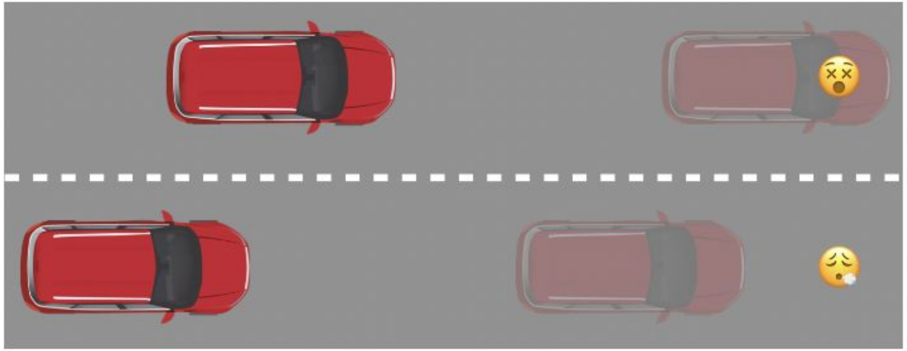
Safe Reinforcement Learning by Imagining the Near Future

Garrett Thomas, Yuping Luo, Tengyu Ma

Berk Can Özmen



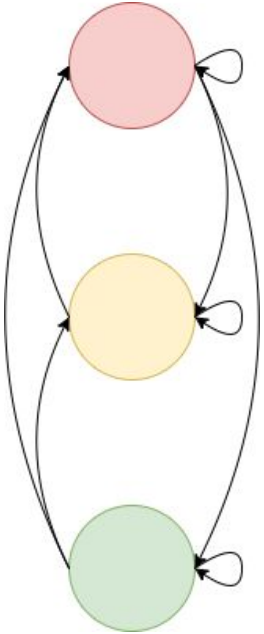
In a nutshell

If:

- irrecoverable and unsafe states are known
- there exists a safe policy

we can guarantee choosing a safe policy

Irrecoverable state

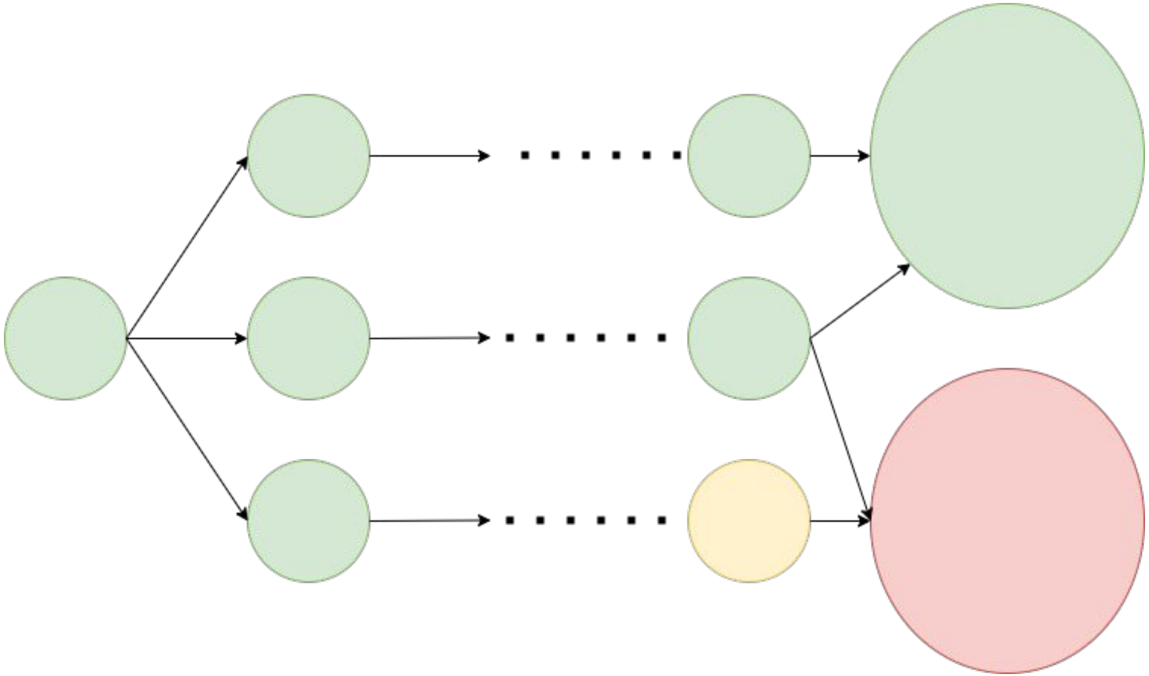


$$s \in S_{unsafe}$$

$$s \notin S_{unsafe}, \text{ given } s_{t+1} = \mathcal{T}(s_t, a_t) \text{ with } s_0 = s,$$
$$\forall t \in \mathbb{N},\ \mathcal{T} \text{ satisfies } s_{\bar{t}} \in S_{unsafe} \text{ for some } \bar{t} \in \mathbb{N}$$

Assumption 3.1. *There exists a horizon $H^* \in \mathbb{N}$ such that, for any irrecoverable states s, any sequence of actions $a_0, \ldots, a_{H^*-1}$ will lead to an unsafe state. That is, if $s_0 = s$ and $s_{t+1} = T(s_t, a_t)$ for all $t \in \{0, \ldots, H^* - 1\}$, then $s_{\bar{t}} \in \mathcal{S}_{\text{unsafe}}$ for some $\bar{t} \in \{1, \ldots, H^*\}$.*

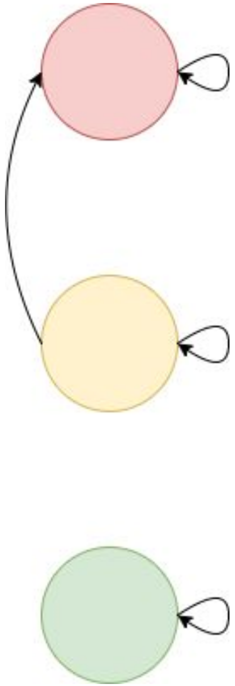
Idea



$$\{\mathcal{T} \mid \forall t \in \{1, .., H^*\}, \mathcal{T}(s_t, a_t) \in S_{safe}\}$$

$$\{\mathcal{T} \mid \exists t \in \{1, .., H^*\}, \mathcal{T}(s_t, a_t) \notin S_{safe}\}$$

Reward Penalty Framework



$$(\tilde{r}(s,a), \widetilde{T}(s,a)) = \begin{cases} (r(s,a), T(s,a)) & s \notin \mathcal{S}_{\text{unsafe}} \\ (-C, s) & s \in \mathcal{S}_{\text{unsafe}} \end{cases}$$

With big enough C

$$\sum_{t=0}^{H^*-1} \gamma^t r_{max} - \sum_{t=H^*}^{\infty} \gamma^t C = \frac{r_{max}(1-\gamma^{H*}) - C\gamma^{H*}}{1-\gamma}$$

$$\sum_{t=0}^{\infty} \gamma^t r_{min} = \frac{r_{min}}{1-\gamma}$$

$$\frac{r_{max}(1-\gamma^{H*}) - C\gamma^{H*}}{1-\gamma} < \frac{r_{min}}{1-\gamma}$$

Known model assumptions

$$C > \frac{r_{max} - r_{min}}{\gamma^{H*}} - r_{max}$$

Unknown model assumptions

$$C > \frac{r_{max} - r_{min}}{\gamma^{H*}} - r_{max}$$

$\hat{T} : S \times A \rightarrow \mathcal{P}(S)$ is **calibrated** if:
$T(s, a) \in \hat{T}(s, a) \quad \forall (s, a) \in (S \times A)$

1 - Bellmin Operator

➢ $$\underline{\mathcal{B}}^{*}Q(s,a) = \tilde{r}(s,a) + \gamma \min_{s' \in \hat{T}(s,a)} max_{a'} Q(s', a')$$

- $\underline{\mathcal{B}}^{*}$ is a $\gamma - contraction$
  in the $\infty - norm$

- Banach's fixed-point theorem

  ★ $\underline{\mathcal{B}}^{*}$ has a unique fixed point $\underline{Q}^{*}$

2 - Optimal Q

- for a **calibrated** $\hat{T}$, $\underline{Q}^*(s,a) \leq \tilde{Q}^*(s,a)$ for all (s,a)

  - $\underline{\mathcal{B}}^* Q(s,a) = \tilde{r}(s,a) + \gamma \min_{s' \in \hat{T}(s,a)} max_{a'} Q(s', a')$

  - $\mathcal{B}^* Q = r(s,a) + \gamma \max_{a'} Q(s', a')$

3 - Safety

- If there is a safe action a at state s, then $argmax_a \underline{Q}^*(s, a)$ is a safe action

  - $$a \in A_{Unsafe} \Rightarrow \underline{Q}^* \leq \underline{\tilde{Q}}^* \leq \frac{r_{max}(1-\gamma^{H*}) - C\gamma^{H*}}{1-\gamma}$$

  - $$a \in A_{Safe} \Rightarrow \frac{r_{min}}{1-\gamma} \leq \underline{Q}^* \leq \underline{\tilde{Q}}^*$$

Algorithm

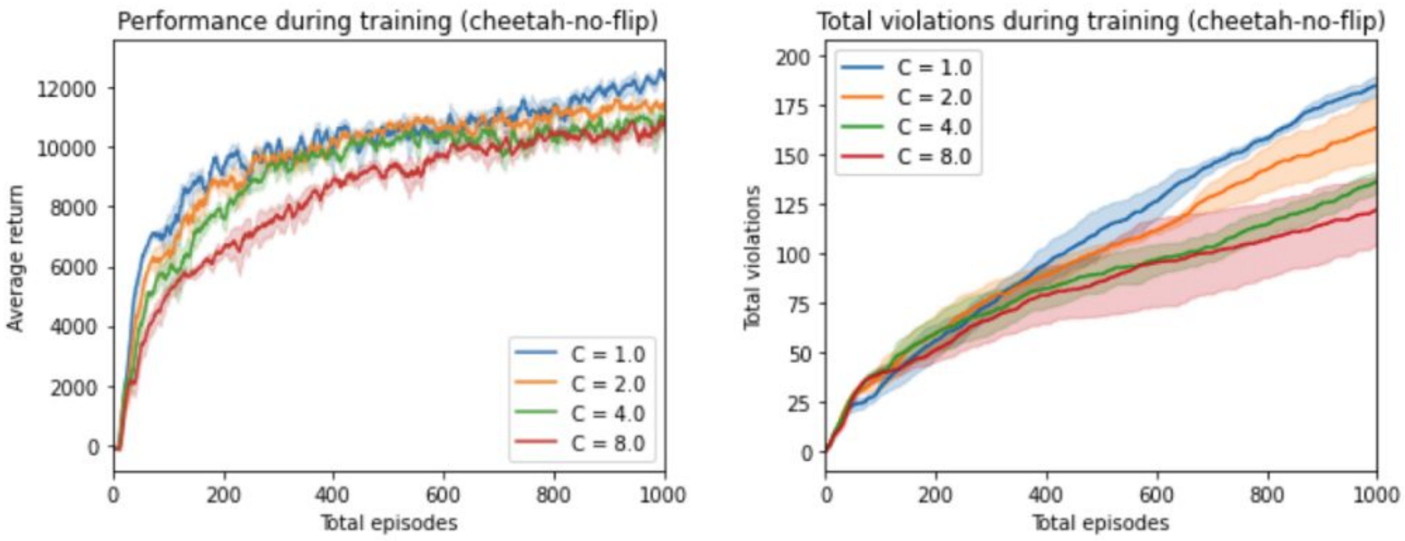
Algorithm 1 Safe Model-Based Policy Optimization (SMBPO)

Require: Horizon H

Initialize empty buffers $\mathcal{D}$ and $\widehat{\mathcal{D}}$, an ensemble of probabilistic dynamics $\{\widehat{T}_{\theta_i}\}_{i=1}^{N}$, policy π_ϕ, critic Q_ψ.
Collect initial data using random policy, add to $\mathcal{D}$.
for episode $1, 2, \ldots$ **do**
    Collect episode using π_ϕ; add the samples to $\mathcal{D}$. Let ℓ be the length of the episode.
    Re-fit models $\{\widehat{T}_{\theta_i}\}_{i=1}^{N}$ by several epochs of SGD on $L_{\widehat{T}}(\theta_i)$ defined in (9)
    Compute empirical r_{min} and r_{max}, and update C according to (3).
    for ℓ times **do**
        for n_{rollout} times (in parallel) **do**
            Sample $s \sim \mathcal{D}$.
            Startin from s, roll out H steps using π_ϕ and $\{\widehat{T}_{\theta_i}\}$; add the samples to $\widehat{\mathcal{D}}$.
        for n_{actor} times **do**
            Draw samples from $\mathcal{D} \cup \widehat{\mathcal{D}}$.
            Update Q_ψ by SGD on $L_Q(\psi)$ defined in (10) and target parameters $\bar{\psi}$ according to (12).
            Update π_ϕ by SGD on $L_\pi(\phi)$ defined in (13).

Based on MBPO (Janner et al., 2019) and Soft Actor Critic (Haarnoja et al., 2018)

Parameter C



(a) Performance with varying C

(b) Cumulative safety violations with varying C

Experiments - exploration

$$L_Q(\psi) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}\cup\widehat{\mathcal{D}}}[(Q_\psi(s,a) - (r + \gamma V_{\bar{\psi}}(s'))^2]$$

$$V_{\bar{\psi}}(s') = \begin{cases} -C/(1-\gamma) & s' \in \mathcal{S}_{\text{unsafe}} \\ \mathbb{E}_{a'\sim\pi(s')}[Q_{\bar{\psi}}(s',a') - \alpha \log \pi_\phi(a' \mid s')] & s' \notin \mathcal{S}_{\text{unsafe}} \end{cases}$$

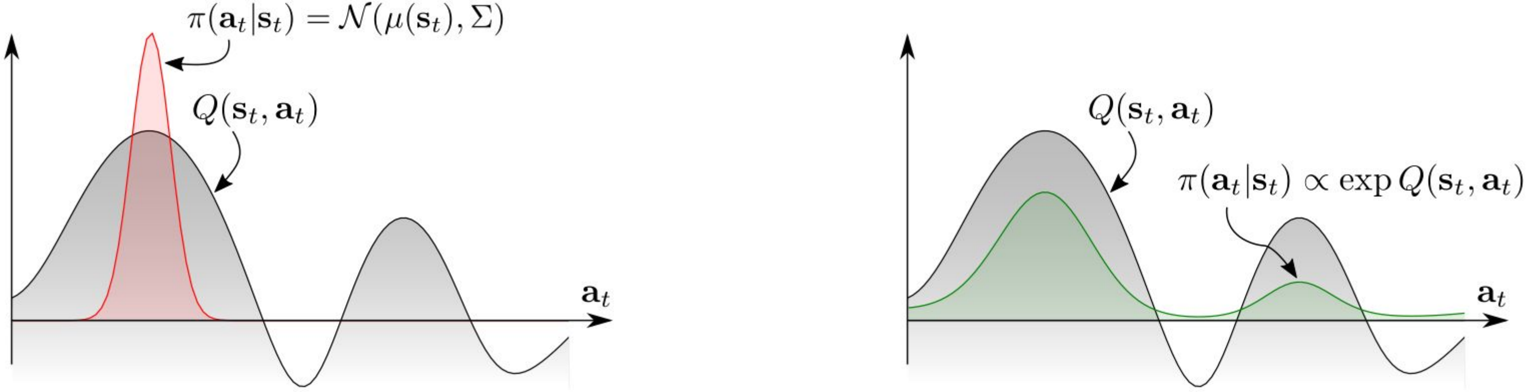
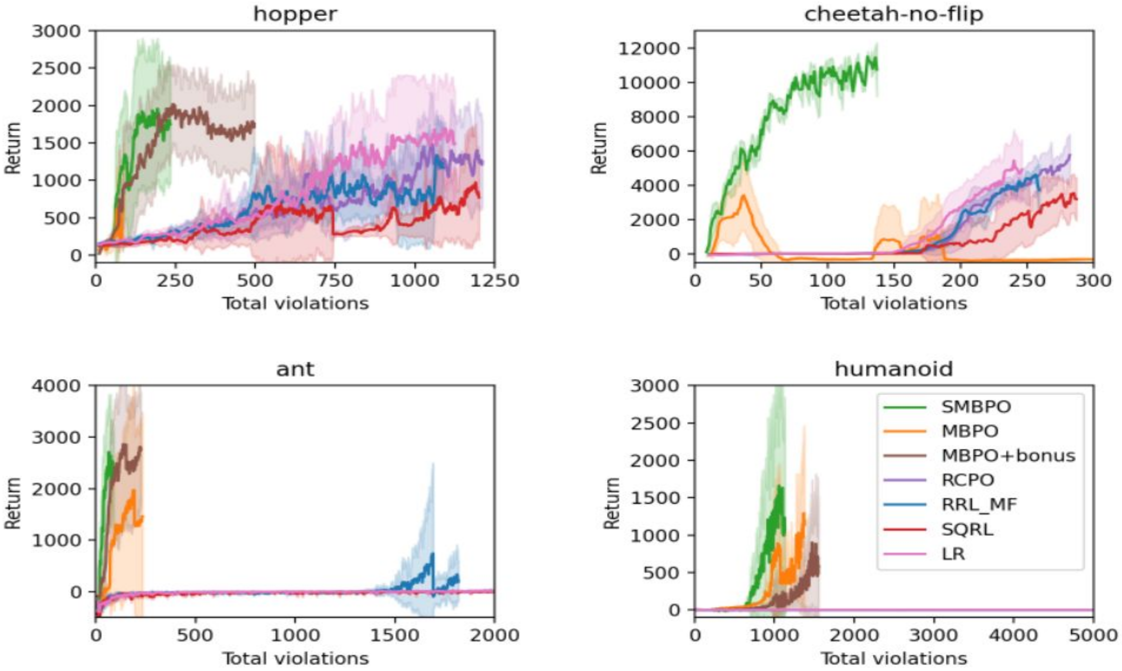


Figure from (Tang & Haarnoja, 2017)

Experiments



References

G. Thomas, Y. Luo, and T. Ma, “Safe reinforcement learning by imagining the near future,” 2022.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. arXiv preprint arXiv:1906.08253, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International Conference on Machine Learning, pages 1861–1870, 2018a.

Haoran Tang, and Tuomas Haarnoja. Learning Diverse Skills via Maximum Entropy Deep Reinforcement Learning. https://bair.berkeley.edu/blog/2017/10/06/soft-q-learning/. 2017