End-to-End Risk-Aware Planning by Gradient Descent

Noah Patton, Jihwan Jeong, Michael Gimelfarb*, Scott Sanner*

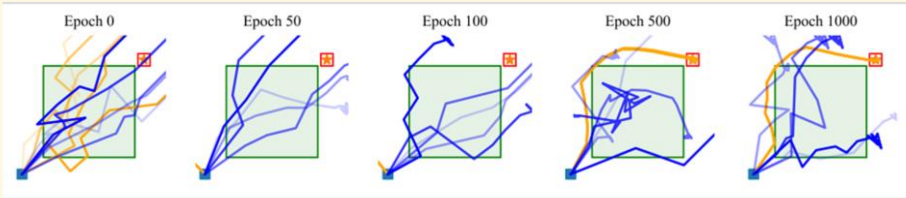
Department of Mechanical and Industrial Engineering, University of Toronto, Canada

** Affiliate to Vector Institute, Toronto, Canada*

ICAPS PRL 2021

MOTIVATION

- Consider trajectory planning in a 2-dimensional environment with a highly stochastic region (green).
- Given that entering the stochastic region increases the probability of failure in a planning framework,
- **We aim to reduce cumulative reward variance, while maintaining high cumulative reward.**



OBJECTIVES

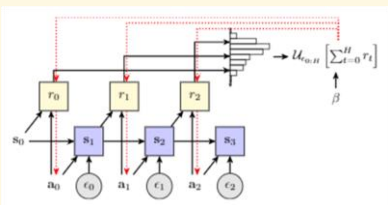
- Enable risk-aware planning using mean-variance approximation of entropic utility.
- Leverage auto-differentiation improve action sequences directly, in an end-to-end manner.
- Avoid computational difficulties of using the Bellman Principle explicitly

RELATED WORK

- Safety is a concern of machine learning models deployed in the real world (Faria 2018; Pereira and Thomas 2020).
- Optimizing expected cumulative reward can lead to excessive risk taking in sequential stochastic decision-making (Moldovan 2014)
- This problem can be addressed by optimizing risk measures with favorable mathematical properties (Ruszczyński 2010)
- Much of the existing scalable end-to-end planning frameworks do not incorporate risk, such as **BackpropPlan** (Wu, Say and Sanner 2017)

MAIN IDEA

- We can sample independent noise and use it to reparametrize stochastic transitions into deterministic transitions with added noise (black arrows).



- For each batch of forward passes we estimate the sufficient statistics of the cumulative reward and use them to calculate the utility objective.
- Notably, due to the reparametrized transitions, we can leverage auto-differentiation to update the sequence of actions (red arrows).

Reparameterization & Forward Sampling

- Suppose you have a stochastic node s_{t+1}
$$s_{t+1} \sim p(\cdot \mid s_t, a_t)$$
- s_{t+1} *blocks* the gradient of the objective from backpropogating to s_t.
- Reparameterization transforms s_{t+1} into:
$$s_{t+1} = \phi(s_t, a_t, \varepsilon_t),\ \ \varepsilon_t \sim p(\varepsilon_t)$$
- where $\phi(\cdot)$ is a deterministic function that is differentiable w.r.t. a_t and s_t.
- Now sampling $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_H)$ we can generate samples of $\sum_{t=0}^{H} r(s_t, a_t)$
- which can be used to estimate sufficient statistics and in turn the utility objective and its gradient.

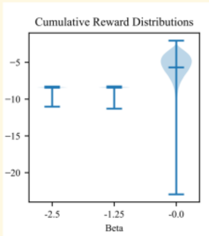
Straight-line Utility Objective

- Update actions based on straight-line utility objective:
$$u_{SL}(s_0) \coloneqq \max_{a_{0:H}} U_{\varepsilon_{0:H}}\left(\sum_{t=0}^{H} r(s_t, a_t)\right)$$
- Where $U_{\varepsilon_{0:H}}$ is the entropic utility:
$$U(X) \coloneqq \frac{1}{\beta} \log \mathbb{E}\left[e^{\beta X}\right]$$
- Using Taylor expansion, it can be written in mean variance form:
$$U(X) = \mathbb{E}[X] + \frac{\beta}{2} Var[X] + O(\beta^2)$$
- Now β can be interpreted as a risk aversion parameter:
  I. $\beta = 0$ induces risk neutral behavior
  II. $\beta > 0$ ($\beta < 0$) induces risk-seeking (risk-averse) behaviors.

- **Theoretical results**:
  I. For any R.V. X, Y if P (X ≥Y) = 1 then U(X) ≥U(Y).
  II. If c is deterministic then U (X + c) = U(X) + c
  III. Due to the recursive property of entropic utility (Osogami 2012; Dowson, Morton, and Pagnoncelli 2020), the optimal utility satisfies the Bellman equation:
$$U_h^*(s_h) = \max_{a_h \in \mathcal{A}} U_{s_{h+1}}\left(r(s_h, a_h) + U_{h+1}^*(s_{h+1})\right)$$
  IV. By result I and III, the optimal utility $U_0^*(s_0)$ satisfies:
$$U_0^*(s_0) \geq u_{SL}(s_0)$$

EXPERIMENTAL EVALUATION

- Tested on two environments:
  I. Navigation (Faulwasser and Findeisen 2009):

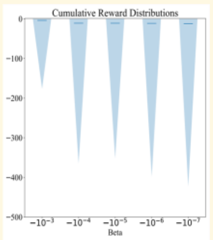


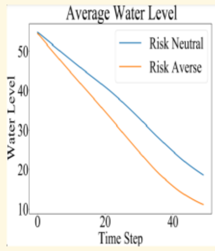
β	Misses (%)
0.0	92.29
-1.25	0.17
-2.5	0.09

Observations:

I. Lower β (more risk-aware) yields lower cumulative reward variance
II. Lower β leads to smaller cumulative reward
III. Lower β leads to less goal state misses
IV. Risk-aware navigation results in agent avoiding highly stochastic region.

  II. Reservoir Control (Yeh 1985):





β	Overflows (%)
-10^{-7}	0.67
-10^{-6}	0.65
-10^{-5}	0.54
-10^{-4}	0.61
-10^{-3}	0.17

Observations:

I. Lower β (more risk-aware) leads to less variance cumulative reward
II. Lower β increased cumulative reward
III. Lower β leads to less overflows in the reservoir domain
IV. Risk-aware reservoir sets the water levels lower.