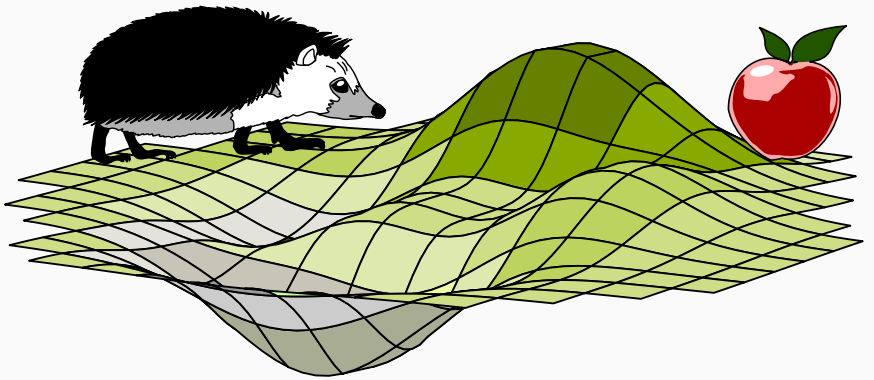
REINFORCEMENT LEARNING

PREDICTIVE ALGORITHMS

Pavel Osinenko



Predictive algorithms

The picture of reinforcement learning is incomplete without model-predictive control which is in most cases a synonym of online, finite-horizon optimal control. The related problem also, like the ∞-horizon one, admits an HJB (although time-dependent). However, unlike the ∞-horizon optimal control problem, the finite-horizon one is computationally tractable.

Model-predictive control (MPC) is listed within some popular RL libraries (such as SLM Lab) and has a number of advantages

Predictive algorithms

A fairly general MPC setup reads:

$$\min_{\{u_{i|k}\}_i^N} \mathbb{E}\left[\sum_{i=1}^{N} \rho(\hat{X}_{i|k}, u_{i|k}) \,\middle|\, X_{1|k} = x_k \right]$$

$$\text{s.t.} \quad dX_t = f(X_t, u^\delta)dt + \sigma(X_t, u^\delta)dB_t,$$

$$u^\delta(t) \equiv u_{i|k}, \quad t \in [(k+i-1)\delta, (k+i)\delta),$$

$$\hat{X}_{i+1|k} = Q^\delta(\hat{X}_{i|k}, u_{i|k}),$$

$$\mathbb{P}[X_{\delta(k+i-1)} \in \mathcal{X}] \geq 1 - \varepsilon_x,$$

$$u_{i|k} \in \mathcal{U}$$

Predictive algorithms

(cont.)

Description:

- sequence notation: $\{\bullet_{i|k}\}_i^N = \{\bullet_k, \ldots, \bullet_{k+N-1}\}$

  $\bullet_{i|k} = \bullet_{k+i-1}$

- N : horizon (length)
- u^δ : sampled control that is piece-wise constant on δ-intervals of time
- Q^δ : a digital model of the environment that produces estimated states $\hat{X}_{i|k}$ at time nodes $(k+i-1)\delta$

Predictive algorithms

(cont.)

If the environment description, i.e., the functions f, σ, is known, then Q^δ may, e.g., be the Euler-Maruyama numerical solver:

$$Q^\delta(\hat{X}_{ilk}, u_{ilk}) = \hat{X}_{ilk} + f(\hat{X}_{ilk}, u_{ilk})\delta + \sigma(\hat{X}_{ilk}, u_{ilk})\Delta B_{ilk},$$

where $\{\Delta B_{ilk}\}_i$ are the Wiener increments, i.e., i.i.d. normal zero-mean random variables with variance δ

Predictive algorithms

(cont.)

Otherwise, an environment model may be learned from experience during exploration

- $\mathcal{X}, \mathcal{U}$ are the sets of allowed states and, respectively, actions. These could, e.g., be interpreted as safe state and action sets. The expression $u_{i|k} \in \mathcal{U}, \forall i \in [N]$ is also called input constraint (here, $[N]$ means $\{1, \ldots, N\}$). The probabilistic type of a state constraint $\mathbb{P}[X_{s(k+i-1)} \in \mathcal{X}] \geq 1 - \varepsilon_x$ is also called chance state constraint with a user-defined tolerance ε_x

Predictive algorithms

MPC has a number of advantages compared to RL, especially in its online neural-network-based variants. Namely, it has an apparatus to guarantee constraint satisfaction at all times (this is related to the so-called recursive feasibility in the MPC theory). That is a strong feature that makes MPC an industrial standard. Furthermore, MPC has a strong theory of environment stabilization, whereas the latter is a major pain for online RL.
In the following, we will study how such a stabilization works on the example of an MDP with a specific structure

Predictive algorithms

Consider the following particular form of an MDP:

$$X_{k+1} = f(X_k, U_k, W_k),$$

$f(0,0,0) = 0$, $W_k \in \mathbb{W} \subseteq \mathbb{R}^d$ are i.i.d., 0-mean with a PDF P_W

State prediction is done via a direct use of the ground truth model (although this is not the only option):

$$X_{i+1|k} = f(X_{i|k}, U_{i|k}, W_{i|k})$$

Reference:

Lorenzen, M., Müller, M. A., & Allgöwer, F. (2019). Stochastic model predictive control without terminal constraints. International Journal of Robust and Nonlinear Control, 29(15), 4987-5001.

Predictive algorithms

Denote: $$P_\infty := \bigotimes_{k=0}^{\infty} P_W ,$$

$$P_k(\bullet) := \mathbb{P}[\bullet \mid X_k = x_k] ,$$

$$\mathbb{E}_k[\bullet] := \mathbb{E}[\bullet \mid X_k = x_k] .$$

$P_\infty(\bullet) = 1$ will also be abbreviated „P_∞ a.s.“

We will assume input and state constraints in the format described above.

Also, the stage cost ρ is assumed continuous, positive-definite and radially unbounded.

Let's denote the objective as:

$$J_N(x_k \mid \{u_{i|k}\}_i^N) := \mathbb{E}_k\left[\sum_{i=1}^{N} \rho(X_{i|k}, u_{i|k})\right]$$

Predictive algorithms

The corresponding value function reads:

$$V_N(x_k) := \min_{\{u_{i|k}\}_i^N} J_N(x_k \mid \{u_{i|k}\}_i^N)$$

s.t. system dynamics,
state prediction,
input & state constraints

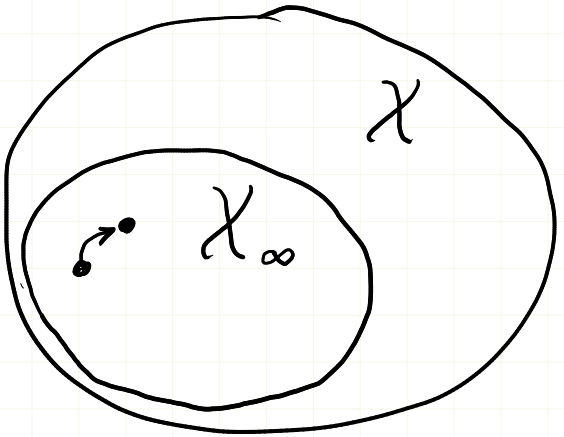
Let's denote the corresponding minimizer as $\{u^*_{i|k}\}_i^N$. Accordingly, the state sequence will be denoted $\{x^*_{i|k}\}_i^N$

Predictive algorithms

The MPC policy is $K_N(x_k) := u^*_{1|k} = u^*_k$.

For the forthcoming analysis to work, we need the following **viability** assumption:

there is a set $\mathcal{X}_\infty \subseteq \mathcal{X}$ s.t. $\forall x_k \in \mathcal{X}_\infty$, $\forall N$ the value function is attained (i.e., a solution $K_N(x_k)$ to the MPC problem exists), and $\forall w \in \mathbb{W}$ $f(x_k, K_N(x_k), w) \in \mathcal{X}_\infty$.



Such a property that the future state lie in the same set is called **positive invariance**. Under control and noise, it's **robust controlled invariance**

Predictive algorithms

The idea of viability is that inside $\mathcal{X}_\infty$, there is always a way (via MPC) to stay inside, no matter what value the noise assumes.

The next assumption describes the properties of the stage cost. First, some notation:

consider, for any N,

$$\{v^*_{i|0}\}_i^N := \arg\min_{\{u_{i|0}\}_i^N} J(0 \mid \{u_{i|0}\}_i^N),$$

the minimizer at zero initial state. Let the corresponding state sequence be $\{z^*_{i|0}\}_i^N$.

Denote

$$\bar{\rho}_i^N := \mathbb{E}_0\left[\rho(z^*_{i|0}, v^*_{i|0})\right],$$

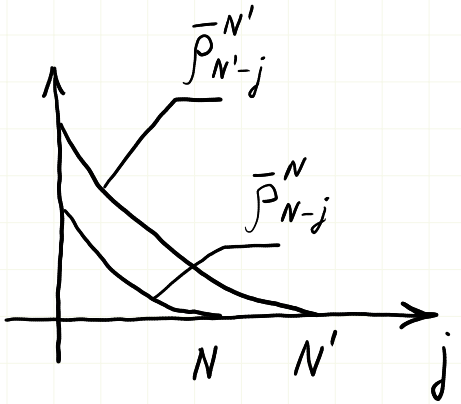
the ith expected stage cost

Predictive algorithms

(cont.)

With this at hand, we assume:

1. $\bar{\rho}_1^N = 0$
2. $\forall\, N' \geq N \quad \forall\, j \qquad \bar{\rho}_{N-j}^N \leq \bar{\rho}_{N'-j}^{N'}$



3. $\forall x_k \qquad \mathbb{E}_k\left[\rho\left(x_{i|k}^*, u_{i|k}^*\right)\right] \geq \bar{\rho}_i^N$

4! $\exists$ p.-d., continuous, radially unbounded function $\nu : \mathbb{R}^n \longrightarrow \mathbb{R}_{\geq 0}$ and $\alpha > 0$ s.t.

$$V_N(x) \leq \alpha\, \nu(x) + \sum_{i=1}^{N} \bar{\rho}_i^N,$$

$$\rho(x,u) \geq \nu(x), \quad \forall\, x \in \mathbb{R}^n,\ u \in \mathcal{U}$$

Predictive algorithms

We will utilize the following variant of the DPP:

let $V_0(x) = 0$, then, for $\ell \in \mathbb{N}$, $x \in \mathcal{X}_\infty$, it holds that

$$V_\ell(x) = \min_u \left\{ \rho(x,u) + \mathbb{E}_{P_W}\left[V_{\ell-1}(X_+^u) \right] \right\}$$

$$\text{s.t.} \quad u \in \mathcal{U}$$

$$\mathbb{P}[X_+^u \in \mathcal{X}] \geq 1 - \varepsilon$$

This can be shown by recursion. Recall we are assuming attainability of the value functions. So, for brevity, we will omit the constraints. First, for $\ell = 1$, we have the relation trivially

Predictive algorithms

(cont.)

Take $l=2$. We have:

$$V_1(x) = \min_u \rho(x,u) ,$$

$$V_2(x) = \min_{u_0,u_1} \left\{ \rho(x,u_0) + \mathbb{E}_{P_W}\left[\rho(X_+^{u_0}, u_1) \mid X=x \right] \right\} .$$

By optimality of V_1,

$$\min_{u_0,u_1} \left\{ \rho(x,u_0) + \mathbb{E}_{P_W}\left[\rho(X_+^{u_0}, u_1) \mid X=x \right] \right\} \geqslant$$

$$\min_{u_0} \left\{ \rho(x,u_0) + \mathbb{E}_{P_W}\left[V_1(X_+^{u_0}) \mid X=x \right] \right\}$$

Predictive algorithms

(cont.)

On the other hand, by optimality of u_0^*, u_1^*, no matter what tail action we take — be it the minimizer corresponding to V_1 — it must hold that

$$\min_{u_0,u_1}\left\{\rho(x,u_0)+\mathbb{E}_{P_W}\left[\rho(X_+^{u_0},u_1)\,|\,X=x\right]\right\} \leqslant$$

$$\min_{u_0}\left\{\rho(x,u_0)+\mathbb{E}_{P_W}\left[V_1(X_+^{u_0})\,|\,X=x\right]\right\},$$

whence the required identity

Predictive algorithms

(cont.)

Suppose the identity holds for $\ell-1$, so

$$V_{\ell-1}(x) = \min_u \left\{ \rho(x,u) + \mathbb{E}_{P_W}\left[V_{\ell-2}(X_+^u) \right] \right\} .$$

Unwrap the next one:

$$V_\ell(x) = \min_{\{u_i\}_{i=0}^{\ell-1}} \left\{ \mathbb{E}\left[\sum_{i=0}^{\ell-1} \rho(X_{+i}^{u_{i-1}}, u_i) \,\middle|\, X = x \right] \right\}$$

$$= \min_{\{u_i\}_{i=0}^{\ell-1}} \left\{ \rho(x,u_0) + \mathbb{E}\left[\sum_{i=1}^{\ell-1} \rho(X_{+i}^{u_{i-1}}, u_i) \,\middle|\, X = x \right] \right\},$$

and apply the same optimality argument observing that

$$V_{\ell-1}\left(X_+^{u_0}\right) = \min_{\{u_i\}_{i=1}^{\ell-1}} \mathbb{E}\left[\sum_{i=1}^{\ell-1} \rho(X_{+i}^{u_{i-1}}, u_i) \,\middle|\, X = x \right]$$

Predictive algorithms

(cont.)

As a corollary, we may state the DPP with an arbitrary head:

$$V_N(x_k) = \min_{\{u_{i|k}\}_i^{N-\ell}} \left\{ \mathbb{E}\left[\sum_{i=1}^{N-\ell} \rho(x_{i|k}, u_{i|k}) + V_\ell\left(x_{N-\ell+1|k}\right) \right] \right\},$$

for any $\ell \in \mathbb{N}$, $\ell < N$, $x_k \in \mathcal{X}_\infty$.

We will use this version in what follows

Predictive algorithms

Now, we can state the main stability result:

under the stated assumptions, Let $x_0 \in \mathcal{X}_\infty$. Then:

$\forall k$: 1. the constraints are satisfied

2. $\exists C_1 \geq 0$ s.t.

$$\mathbb{E}_k\left[V_N(X^*_{k+1})\right] - V_N(x_k) \leq -\rho(x_k, u^*_k) + \frac{a^2}{N-1} v(x) + C_1 .$$

Furthermore, $\exists N \in \mathbb{N}$ s.t. $\forall x_k \in \mathcal{X}_\infty$

$$\mathbb{E}_k\left[V^\circ_N(X^*_{k+1})\right] \leq \lambda V^\circ_N(x_k) + C_2$$

with $\lambda \in (0,1)$, $C_2 \geq 0$ and

$$V^\circ_N(x) := V_N(x) - \sum_{i=1}^{N} \bar{\rho}^N_i$$

To prove this claim, let the current state to have assumed a value x_k and consider the expected difference in value:

$$\mathbb{E}_k\left[V_N\left(X^*_{k+1}\right)\right] - V_N(x_k) \,.$$

Using the non-trivial head version of the DPP, we may write (for $j < N$):

$$\mathbb{E}_k\left[V_N\left(X^*_{k+1}\right)\right] = \mathbb{E}_k\left[\mathbb{E}_{k+1}\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i|k+1}, u^*_{i|k+1}\right) + V_j\left(X^*_{N-j|k+1}\right)\right]\right]$$

Plugging in the optimal action sequence $\{u^*_{i|k+1}\}_i^N$ because it's the minimizer corresponding to $V_N\left(X^*_{k+1}\right)$

Predictive algorithms

(cont.)

Next,

$$\mathbb{E}_k\left[V_N\left(X^*_{k+1}\right)\right] = \mathbb{E}_k\left[\mathbb{E}_{k+1}\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i|k+1}, u^*_{i|k+1}\right) + V_j\left(X^*_{N-j|k+1}\right)\right]\right]$$

$$= \mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i|k+1}, u^*_{i|k+1}\right) + V_j\left(X^*_{N-j|k+1}\right)\right]$$

since the condition of the inner expectation is already determined by the condition of the outer one

Predictive algorithms

(cont.)

Now, we consider, at step $k+1$, a feasible candidate for the head $\{u^*_{i+1|k}\}_i^{N-j-1}$.

Why is it feasible?

Well, $\{u^*_{i|k}\}_i^N$ was feasible at step k and we just make a cyclic shift at step $k+1$ remembering that we're inside $\mathcal{X}_\infty$ which „neutralizes" the noise.

With this feasible candidate, we may write:

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(x^*_{i|k+1}, u^*_{i|k+1}\right) + V_j\left(x^*_{N-j|k+1}\right)\right] \leq$$

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(x^*_{i+1|k}, u^*_{i+1|k}\right) + V_j\left(x^*_{N-j+1|k}\right)\right]$$

(be aware of the difference in the state sequences!)

Predictive algorithms

(cont.)

Now, since $\sum_{i=N-j+1}^{N} \bar{\rho}_i^N - \sum_{i=1}^{j} \bar{\rho}_i^j$ is non-negative*, we may add it on the right-hand side of the last inequality:

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i|k+1}, U^*_{i|k+1}\right) + V_j\left(X^*_{N-j|k+1}\right)\right] \leq$$

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i+1|k}, U^*_{i+1|k}\right) + V_j\left(X^*_{N-j+1|k}\right)\right] +$$

$$\sum_{i=N-j+1}^{N} \bar{\rho}_i^N - \sum_{i=1}^{j} \bar{\rho}_i^j$$

* By item 2. of the standing assumption, each $\bar{\rho}^N_{N-i} \geq \bar{\rho}^j_{j-i}$

Predictive algorithms

(cont.)

Plugging this back into the expected difference in value and unwrapping the pivot yields:

$$\mathbb{E}_k\left[V_N\left(X^*_{k+1}\right)\right] - V_N(x_k) \leqslant \mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i+1|k}, u^*_{i+1|k}\right) + V_j\left(X^*_{N-j+1|k}\right)\right] + \sum_{i=N-j+1}^{N} \bar{\rho}_i^N - \sum_{i=1}^{j} \bar{\rho}_i^j - \mathbb{E}_k\left[\sum_{i=1}^{N} \rho\left(X^*_{i|k}, u^*_{i|k}\right)\right].$$

By the standing assumption (item 3.),

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1} \rho\left(X^*_{i+1|k}, u^*_{i+1|k}\right)\right] + \sum_{i=N-j+1}^{N} \bar{\rho}_i^N \leqslant \mathbb{E}_k\left[\sum_{i=1}^{N} \rho\left(X^*_{\cdot+1|k}, u^*_{i+1|k}\right)\right]$$

Predictive algorithms

(cont.)

Therefore,

$$\mathbb{E}_k\left[\sum_{i=1}^{N-j-1}\rho(X^*_{i+1|k},U^*_{i+1|k})\right]+\sum_{i=N-j+1}^{N}\bar{\rho}_i^N-\mathbb{E}_k\left[\sum_{i=1}^{N}\rho(X^*_{i|k},U^*_{i|k})\right]\leqslant$$

$$-\rho(x_k,u_k^*),$$

whence

$$\mathbb{E}_k[V_N(X^*_{k+1})]-V_N(x_k)\leqslant-\rho(x_k,u_k^*)-\sum_{i=1}^{j}\bar{\rho}_i^j+$$

$$\mathbb{E}_k\left[V_j(X^*_{N-j+1|k})\right]$$

Predictive algorithms

(cont.)

Again, by the standing assumptin (item 4), we may write:

$$\mathbb{E}_k\left[V_N\left(X^*_{k+1}\right)\right] - V_N(x_k) \leq -\rho(x_k, u^*_k) - \sum_{i=1}^{j} \bar{\rho}_i^j + \mathbb{E}_k\left[V_j\left(X^*_{N-j+1|k}\right)\right]$$

$$\leq -\rho(x_k, u^*_k) + \alpha\, \mathbb{E}_k\left[\nu\left(X^*_{N-j+1|k}\right)\right].$$

Predictive algorithms

(cont.)

Now, by the same item 1. of the assumption, we have

$$\mathbb{E}_k\left[\sum_{i=2}^{N} \rho(X^*_{i|k}, u^*_{i|k})\right] \overset{*}{\leq} V_N(x_k) \leq \alpha V(x_k) + \sum_{i=1}^{N} \bar{\rho}_i^N .$$

From this, we may deduce, using $\forall x, u \in \mathcal{U}\ \rho(x,u) \geq V(x)$, that there is an $\ell \in [2:N]$ s.t.

$$\mathbb{E}_k\left[V(X^*_{\ell|k})\right] \leq \mathbb{E}_k\left[\rho(X^*_{\ell|k}, u^*_{\ell|k})\right] \leq \frac{\alpha}{N-1} V(x_k) + \frac{1}{N-1} \sum_{i=1}^{N} \bar{\rho}_i^N$$

(there is a term that is not greater than the average among the $N-1$ ones)

* This is just dropping the first term

Predictive algorithms

(cont.)

Going back to the expected difference in value:

$$\mathbb{E}_k\left[V_N(X^*_{k+1})\right] - V_N(x_k) \leq -\rho(x_k, u^*_k) + \alpha\, \mathbb{E}_k\left[\nu\left(X^*_{N-j+1|k}\right)\right],$$

and using the last derived inequality (from the previous slide), we may write:

$$\mathbb{E}_k\left[V_N(X^*_{k+1})\right] - V_N(x_k) \leq -\rho(x_k, u^*_k) + \frac{\alpha^2}{N-1}\nu(x_k) + \frac{\alpha}{N-1}\sum_{i=1}^{N}\bar{\rho}_i^N$$

(assuming we took $j = N-l+1$ from the start).

Setting $c_1 := \frac{\alpha}{N-1}\sum_{i=1}^{N}\bar{\rho}_i^N$ validates the first part of the original claim

Predictive algorithms

(cont.)

As for the second, observe that if $N \geq a^2 + 1$,

$$\frac{a^2}{N-1} V(x_k) \leq V(x_k)$$

and so

$$\frac{a^2}{N-1} V(x_k) \leq \rho(x_k, u), \quad \forall u \in \mathcal{U},$$

whence

$$-\rho(x_k, u_k^*) + \frac{a^2}{N-1} V(x_k) + c_1 \leq \left(\frac{a^2}{N-1} - 1\right) V(x_k) + c_1$$

Predictive algorithms

(cont.)

Recall

$$V_N^0(x_k) = V_N(x_k) - \sum_{i=1}^{N} \bar{\rho}_i^N \quad \text{and} \quad c_1 = \frac{\alpha}{N-1} \sum_{i=1}^{N} \bar{\rho}_i^N$$

Also, $V_N(x_k) \le \alpha\, v(x_k) + \frac{N-1}{\alpha} c_1$,

whence $-v(x_k) \le -\frac{1}{\alpha} V_N(x_k) + \frac{N-1}{\alpha^2} c_1 = -\frac{1}{\alpha} V_N^0(x_k)$

and so

$$\left(\frac{\alpha^2}{N-1} - 1\right) v(x_k) + c_1 \le \left(\frac{\alpha}{N-1} - \frac{1}{\alpha}\right) V_N^0(x_k) + c_1 .$$

Predictive algorithms

(cont.)

Now, since $\mathbb{E}_k[V_N(X^*_{k+1})] - V_N(x_k) = \mathbb{E}_k[V^0_N(X^*_{k+1})] - V^0_N(x_k)$,

we have:

$$\mathbb{E}_k[V^0_N(X^*_{k+1})] - V^0_N(x_k) \leq \left(\frac{a}{N-1} - \frac{1}{a}\right) V^0_N(x_k) + C_1,$$

whence

$$\mathbb{E}_k[V^0_N(X^*_{k+1})] \leq \left(\frac{a}{N-1} - \frac{1}{a} + 1\right) V^0_N(x_k) + C_1$$

which validates the second part of the claim with

$$\lambda = \frac{a}{N-1} - \frac{1}{a} + 1 \quad \text{and} \quad C_2 = C_1 = \frac{a}{N-1} \sum_{i=1}^{N} \bar{\rho}_i^N$$

Predictive algorithms

(cont.)

If $N > \alpha^2 + 1$, then

$$\lambda = \frac{\alpha}{N-1} - \frac{1}{\alpha} + 1 = \frac{\alpha^2 - N + 1 + (N-1)\alpha}{(N-1)\alpha} = \frac{\alpha^2 + (\alpha-1)(N-1)}{(N-1)\alpha} < 1$$

which shows that $\mathbb{E}_k\left[V_N^0(X_{k+1}^*)\right]$ stays bounded for any k (in fact, it'd decay asymptotically to zero if c_1 were zero — so is the case if $f(0,0,w)=0$ for any $w \in \mathbb{W}$).

Predictive algorithms

To actually claim stochastic stability (in mean), we need to relate

$$\mathbb{E}_k\left[V_N^0(X_{k+1}^*)\right]$$

to the mean state.

Assume, for simplicity, that indeed $f(0,0,w)=0$ for any $w \in \mathbb{W}$. Then, zero control at zero initial state yields zero-mean cost, i.e., $\bar{\rho}_i^N = 0 \ \forall i$, whence $V_N^0 = V_N$.

We have $\mathbb{E}_k\left[V_N(X_{k+1}^*)\right] \leq \lambda^k V_N(x_0) \leq \lambda^k \alpha V(x_0)$

Predictive algorithms

(cont.)

So,

$$\mathbb{E}_k\left[V_N(X^*_{k+1})\right] \leq \lambda^k a\, v(x_0) \ .$$

Since $\forall\, x, u \in \mathcal{U} \quad v(x) \leq p(x,u)$, deducting all but the first stage cost from the value function on the left yields:

$$\mathbb{E}_k\left[v(X^*_{k+1})\right] \leq \lambda^k a\, v(x_0) \ .$$

So, if, say $v(x) = b\|x\|^2$ for some $b > 0$, we may claim mean-square stability of the (norm) state and, moreover, its asymptotic convergence to zero

Predictive algorithms

Discussion

- We considered environment stabilization for an MDP, but the results could be extended to an SDE in the sample-and hold setup (as described at the beginning). To this end, one may employ moment-based accuracy bounds of the Euler-Maruyama or some other numerical integration method
- In the studied result, stabilization was achieved for all sufficiently long horizons. MPC theory actually offers further machinery for stabilization, such as terminal constraints and terminal costs