# Approximate Dynamic Programming via Iterated Bellman Inequalities

Yang Wang and Stephen Boyd
*Information Systems Laboratory, Electrical Engineering, Stanford University*

## Stochastic Control

- discrete-time dynamics: $x_{t+1} = f(x_t, u_t, w_t)$, $t = 0, 1, \ldots,$
  - $x_t \in \mathcal{X}$ is the state
  - $u_t \in \mathcal{U}$ is the input
  - $w_t \in \mathcal{W}$ is IID noise
  - $x_0$ is random, independent of $w_t$
- state feedback control policy: $u_t = \psi(x_t)$, $t = 0, 1, \ldots,$
  - $\psi : \mathcal{X} \to \mathcal{U}$ is the state feedback function
- infinite horizon discounted cost:

$$J = \mathbf{E} \sum_{t=0}^{\infty} \gamma^t \ell(x_t, u_t)$$

  - $\ell : \mathcal{X} \times \mathcal{U} \to \mathbf{R} \cup \{+\infty\}$ is stage cost function
  - $\gamma \in (0, 1)$ is discount factor
- use infinite values of $\ell$ to encode constraints on state and input:

$$(x_t, u_t) \in \mathcal{C} = \{(x, u) \mid \ell(x, u) < \infty\} \text{ a.s.}$$

- stochastic control problem: *choose feedback function $\psi$ to minimize $J$*
- infinite dimensional nonconvex problem: very hard to solve in general
- $\psi^\star$ and $J^\star$ denote optimal feedback function and objective value

## Dynamic Programming

- can characterize optimal solution via dynamic programming
- find (unique) $V^\star : \mathcal{X} \to \mathbf{R}$ that satisfies *Bellman equation*

$$V^\star(z) = \min_{v \in \mathcal{U}} \left(\ell(z, v) + \gamma \, \mathbf{E} \, V^\star(f(z, v, w_t))\right), \quad \forall z \in \mathcal{X}$$

  (expectation is over $w_t$)
- optimal feedback function is then

$$\psi^\star(z) = \operatorname*{argmin}_{v \in \mathcal{U}} \left(\ell(z, v) + \gamma \, \mathbf{E} \, V^\star(f(z, v, w_t))\right)$$

- optimal value of stochastic control problem is $J^\star = \mathbf{E} \, V^\star(x_0)$
- define *Bellman operator* $\mathcal{T}$, for $h : \mathcal{X} \to \mathbf{R}$

$$(\mathcal{T}h)(z) = \inf_{v \in \mathcal{U}} \{\ell(z, v) + \gamma \, \mathbf{E} \, h(f(z, v, w_t))\}$$

- Bellman equation is then $V^\star = \mathcal{T}V^\star$
- $\mathcal{T}$ is monotone: $V_1 \le V_2 \;\Rightarrow\; \mathcal{T}V_1 \le \mathcal{T}V_2$
- $\mathcal{T}$ is $\gamma$-contraction (w.r.t. $\sup$-norm) so for any $V : \mathcal{X} \to \mathbf{R}$

$$V^\star = \lim_{k \to \infty} \mathcal{T}^k V$$

  ('value iteration' converges)

## Bounds and Approximate Policies

- we consider problems for which we cannot compute $V^\star$, $\psi^\star$, $J^\star$
- **performance bound**:
  - find effectively computable lower bound $J^{\mathrm{lb}}$ on $J^\star$
- **approximate policy**:
  - find effectively computable policy $\hat{\psi}$ with performance $\hat{J}$
- if $\hat{J} - J^{\mathrm{lb}}$ small, we're done (for this problem), since $J^{\mathrm{lb}} \le J^\star \le \hat{J}$
- 'effectively computable' means, *e.g.*, by solving tractable convex optimization problem(s)

## Iterated Bellman inequality

- suppose $\hat{V}$ satisfies *iterated Bellman inequality*

$$\hat{V} \le \mathcal{T}^M \hat{V}$$

- then $\hat{V} \le \mathcal{T}^M \hat{V} \le \mathcal{T}^{2M} \hat{V}$ (by monotonicity of $\mathcal{T}$)
- thus we get

$$\hat{V}(x) \le \lim_{k \to \infty} (\mathcal{T}^{kM} \hat{V})(x) = V^\star(x), \quad \forall x \in \mathcal{X}$$

  (since $\mathcal{T}$ is a $\gamma$-contraction)
- iterated Bellman inequality is a sufficient condition for $\hat{V} \le V^\star$
- $\hat{V}$ is a value function *underestimator*
- then (by monotonicity of expectation) $\mathbf{E} \, \hat{V}(x_0) \le \mathbf{E} \, V^\star(x_0) = J^\star$
- so value function underestimator $\hat{V}$ yields performance bound $J^{\mathrm{lb}} = \mathbf{E} \, \hat{V}(x_0)$
- can get better bounds with larger $M$

## Bound Optimization

- a sufficient condition for the iterated Bellman inequality is

$$\hat{V}_{i-1} \le \mathcal{T}\hat{V}_i, \quad i = 1, \ldots, M$$

  where we define $\hat{V}_0 = \hat{V}_M = \hat{V}$
- each $\hat{V}_i$ is restricted to a finite-dimensional subspace

$$\hat{V}_i = \sum_{j=1}^{K} \alpha_{ij} V^{(j)}$$

  - $\alpha_{ij} \in \mathbf{R}$ are coefficients; $V^{(j)} : \mathcal{X} \to \mathbf{R}$ are (pre-selected) basis functions
- optimize performance bound by solving

$$\begin{array}{ll} \text{maximize} & \mathbf{E} \, \hat{V}(x_0) = \alpha_1 \, \mathbf{E} \, V^{(1)}(x_0) + \cdots + \alpha_K \, \mathbf{E} \, V^{(K)}(x_0) \\ \text{subject to} & \hat{V}_{i-1} \le \mathcal{T}\hat{V}_i, \quad i = 1, \ldots, M \end{array}$$

  - convex optimization problem in variables $\alpha_{ij}$

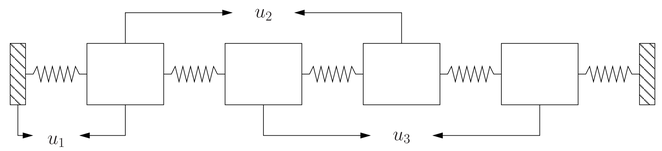
## ADP Policy

- approximate dynamic programming (ADP) policy is

$$\psi^{\mathrm{adp}}(z) = \operatorname*{argmin}_{v \in \mathcal{U}} \left(\ell(z, v) + \gamma \, \mathbf{E} \, V^{\mathrm{adp}}(f(z, v, w_t))\right)$$

  - $V^{\mathrm{adp}} : \mathcal{X} \to \mathbf{R}$ (which is to be chosen) is the *approximate value function*
  - when $V^{\mathrm{adp}} = V^\star$, this is optimal policy
- often performs well, even when $V^{\mathrm{adp}}$ is not close to $V^\star$
- the underestimator $\hat{V}$ is a natural choice for $V^{\mathrm{adp}}$

## Example: Discretized Mechanical System




- 4 masses connected by springs
- 3 input forces, with $|u_i| \le 1$
- $w_t \sim \mathcal{N}(0, 0.1I)$
- $Q = 0.1I$, $R = 0.01I$, $\gamma = 0.95$

| Policy/Bound | Objective/Value |
|---|---|
| ADP ($V^{\mathrm{adp}} = \hat{V}$ from iterated bound) | 162.8 |
| $J^{\mathrm{lb}}$ (iterated Bellman with $M = 50$) | 154.8 |
| $J^{\mathrm{lb}}$ (iterated Bellman with $M = 1$) | 111.1 |
| Unconstrained LQR | 55.7 |

- iterated Bellman inequality gives better bounds
- performance of ADP policy very close to bound
- can close gap even further with a few simple methods ...
- ADP policy can be evaluated in *microseconds*

## Summary

- applies to many other problem families with *e.g.*, polynomial dynamics and stage costs, switching systems, multiplicative disturbances, ...
- in many cases, gives everything you want:
  - a provable lower bound on performance
  - a policy that comes close in performance, can be evaluated quickly