

The Curse of Uncertainty in Dynamic Programming and How to Fix It

Laurent El Ghaoui

co-author: Arnab Nilim

EECS Dept., UC Berkeley, USA

LIDS Seminar—November 2004

Overview

- *The curse of uncertainty in Markov Decision Processes*
- Robust control formulation
- Robust dynamic programming recursion
- Entropy-based uncertainty models
- Application: air traffic routing

Markov decision processes

- Consider a **finite-state, finite-action** discrete-time process:
 - ▷ state set \mathcal{X} (cardinality: n),
 - ▷ action set \mathcal{A} (cardinality: m).
- At each time step $t \in T := \{0, \dots, N - 1\}$, the states make transitions according to a **transition matrix** $P_t^{a_t}$, where a_t is the current action.
- An action a taken at time t in state i incurs a **cost** $c_t(i, a)$.
- The decision maker's **policy** is represented as $\pi = (\mathbf{a}_0, \dots, \mathbf{a}_{N-1})$, where \mathbf{a}_t 's are functions of the state.

Nominal problem

Finite-horizon problem:

$$\phi_N(\Pi, \tau) := \min_{\pi \in \Pi} C_N(\pi, \tau),$$

where

- $\Pi := \mathcal{A}^N$ the decision maker's policy space,
- $\tau := (P_t^a)_{a \in \mathcal{A}, t \in T}$ denotes nature's policy—it is **fixed** for now.
- C_N is the associated γ -discounted total expected cost:

$$C_N(\pi, \tau) := \mathbf{E} \left(\sum_{t=0}^{N-1} \gamma^t c_t(i_t, \mathbf{a}_t(i_t)) + c_N(i_N) \right).$$

Bellman's recursion

- The **optimal value** of the nominal problem is $\phi_N(\Pi, \tau) = q^T v_0$, where q is the distribution of the initial state, and the value function $(v_t)_{t \in T}$ is the solution to the dynamic programming recursion:

$$v_t(i) = \min_{a \in \mathcal{A}} (c_t(i, a) + \gamma(p_i^a)^T v_{t+1}), \quad i \in \mathcal{X}, \quad t \in T,$$

where p_i^a is the i -th row of P^a .

- An **optimal policy** $\pi^* = (\mathbf{a}_0^*, \dots, \mathbf{a}_{N-1}^*)$ is obtained by setting:

$$\mathbf{a}_t^*(i) \in \arg \min_{a \in \mathcal{A}} (c_t(i, a) + \gamma(p_i^a)^T v_{t+1}), \quad i \in \mathcal{X}, \quad t \in T.$$

Complexity of Bellman's recursion

Computational complexity for computing an ϵ -suboptimal policy:

- Finite horizon case: $O(mn^2N)$. (N : horizon length.)
- Infinite horizon: $O\left(mn^2 \log \frac{\gamma}{(1-\gamma)\epsilon}\right)$. (γ : discount factor.)

Complexity of Bellman's recursion

Computational complexity for computing an ϵ -suboptimal policy:

- Finite horizon case: $O(mn^2N)$. (N : horizon length.)

By exhaustive search: $O(m^{n^N})!$

- Infinite horizon: $O\left(mn^2 \log \frac{\gamma}{(1-\gamma)\epsilon}\right)$. (γ : discount factor.)

The curse of uncertainty

- In real-world problems, transition probabilities are very often **inaccurate**, due to estimation errors.
- The optimal solution may be very **sensitive** w.r.t. these probabilities.

The curse of uncertainty

- In real-world problems, transition probabilities are very often **inaccurate**, due to estimation errors.
- The optimal solution may be very **sensitive** w.r.t. these probabilities.
- We need to find **robust policies**: policies that perform well even in the presence of estimation errors.
- We want to obtain robustness at **low extra computational cost**.

Previous work

Many authors have recognized the need for taking into account uncertainty in transition matrix, including (but not limited to!):

- White & Eldeib, Satia & Lave, Givan, Leach & Dean Bagnell, Ng & Schneider (robot path planning), Abbad & Filar (control problems), Kalyanasundaram & Chong (call admission in a network), Epstein & Schneider (dynamic portfolio choice model), Madanat & Kuhn (infrastructure management), ...
- Most models assume component-wise uncertainty, i.e. the transition matrix is an **interval matrix**:
 - ▷ such models lead to overly conservative policies;
 - ▷ they do not capture possible asymmetries in the confidence of estimates.

Previous work (follow'd)

- More recently, Iyengar has built on an initial report of ours and gave an independent proof of the "robust recursion".
- Previous results do not provide complexity estimates and most of them rely on heuristics.

Contributions

- **Formulating** uncertain MDP problem as a robust control problem;

Contributions

- **Formulating** uncertain MDP problem as a robust control problem;
- **Proving** a robust Bellman recursion for both finite and infinite horizon problems;

Contributions

- **Formulating** uncertain MDP problem as a robust control problem;
- **Proving** a robust Bellman recursion for both finite and infinite horizon problems;
- Identifying a class of **statistically accurate** uncertainty models for the transition matrices . . .

Contributions

- **Formulating** uncertain MDP problem as a robust control problem;
- **Proving** a robust Bellman recursion for both finite and infinite horizon problems;
- Identifying a class of **statistically accurate** uncertainty models for the transition matrices ...
- ...for which the robust recursion can be implemented via a simple bisection algorithm, **at no extra computational cost** w.r.t. the classical recursion.

Overview

- The curse of uncertainty in Markov Decision Processes
- *Robust control formulation*
- Robust dynamic programming recursion
- Entropy-based uncertainty models
- Application: air traffic routing

Describing uncertainty on transition matrices

- We will assume that transition matrices are unknown-but-bounded:

$$\forall a \in \mathcal{A}, P^a \in \mathcal{P}^a,$$

where \mathcal{P}^a is a given subset of the set of transition matrices of $\mathbf{R}^{n \times n}$.

Describing uncertainty on transition matrices

- We will assume that transition matrices are unknown-but-bounded:

$$\forall a \in \mathcal{A}, P^a \in \mathcal{P}^a,$$

where \mathcal{P}^a is a given subset of the set of transition matrices of $\mathbf{R}^{n \times n}$.

- We can assume two kinds of behavior for nature:
 - ▷ **Time-varying uncertainty model:** nature picks the transition matrices once and for all time periods.
 - ▷ **Stationary uncertainty model:** nature picks a new set of transition matrices at each time period.

Two robust control problems

$$\phi_N(\Pi, \mathcal{U}) := \min_{\pi \in \Pi} \max_{\tau \in \mathcal{U}} C_N(\pi, \tau),$$

where the set \mathcal{U} represents nature's policy space:

- Time-varying uncertainty model: $\mathcal{U} = \mathcal{T}$, where

$$\mathcal{T} := \left(\bigotimes_{a \in \mathcal{A}} \mathcal{P}^a \right)^N.$$

- Stationary uncertainty model: $\mathcal{U} = \mathcal{T}_s$, where

$$\mathcal{T}_s := \{ \tau \in \mathcal{T} : P_t^a = P_\tau^a \text{ for every } t, \tau \in \mathcal{T}, a \in \mathcal{A} \}.$$

Between a rock and a hard place

- The **stationary** model is attractive for statistical reasons, as it is much easier to develop statistically accurate sets of confidence when the underlying process is time-invariant.

Between a rock and a hard place

- The **stationary** model is attractive for statistical reasons, as it is much easier to develop statistically accurate sets of confidence when the underlying process is time-invariant.

Difficulty: solving the corresponding robust control problem.

Between a rock and a hard place

- The **stationary** model is attractive for statistical reasons, as it is much easier to develop statistically accurate sets of confidence when the underlying process is time-invariant.

Difficulty: solving the corresponding robust control problem.

- The **time-varying** model is attractive, as one can solve the corresponding game using the robust Bellman recursion seen later.

Between a rock and a hard place

- The **stationary** model is attractive for statistical reasons, as it is much easier to develop statistically accurate sets of confidence when the underlying process is time-invariant.

Difficulty: solving the corresponding robust control problem.

- The **time-varying** model is attractive, as one can solve the corresponding game using the robust Bellman recursion seen later.

Difficulty: estimating a meaningful set of confidence for time-varying transition matrices P_t^a .

Our approach

We'll start with a **stationary model**, but **approximate** the control problem using a time-varying model—this will incur some degree of **suboptimality**.

Our approach

We'll start with a **stationary model**, but **approximate** the control problem using a time-varying model—this will incur some degree of **suboptimality**.

Theorem: With a γ -discounted cost function, the gap between the optimal values of the finite-horizon problems under stationary and time-varying uncertainty models goes to zero as the horizon length N goes to infinity, at a geometric rate γ :

$$0 \leq \phi_N(\Pi, \mathcal{T}) - \phi_N(\Pi, \mathcal{T}_s) \leq \text{Constant} \cdot \frac{\gamma^N}{1 - \gamma}.$$

Thus, there is **no gap** in infinite-horizon problems.

Overview

- The curse of uncertainty in Markov Decision Processes
- Robust control formulation
- *Robust dynamic programming recursion*
- Entropy-based uncertainty models
- Application: air traffic routing

Robust Bellman recursion

(Finite-horizon case; set $\gamma = 1$.)

Theorem: With a time-varying uncertainty model, *perfect duality* holds:

$$\phi_N(\Pi, \mathcal{T}) := \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} C_N(\pi, \tau) = \max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} C_N(\pi, \tau).$$

The problem can be solved via the recursion

$$v_t(i) = \min_{a \in \mathcal{A}} \left(c_t(i, a) + \max_{p \in \mathcal{P}_i^a} p^T v_{t+1} \right), \quad i \in \mathcal{X}, \quad t \in T, \quad (1)$$

where \mathcal{P}_i^a is the projection of \mathcal{P}^a onto the i -th state coordinates.

Worst-case value

Theorem (follow'd): The worst-case expected cost with a given policy $\pi = (\mathbf{a}_0, \dots, \mathbf{a}_{N-1})$:

$$\phi_N(\pi, \mathcal{T}) := \max_{\tau \in \mathcal{T}} C_N(\pi, \tau),$$

can be evaluated by the following recursion

$$v_t^\pi(i) = c_t(i, \mathbf{a}_t(i)) + \max_{p \in \mathcal{P}_i^{\mathbf{a}_t(i)}} p^T v_{t+1}^\pi, \quad i \in \mathcal{X}, \quad t \in T, \quad (2)$$

which provides the worst-case value function v^π for the policy π .

Optimal policy

Theorem (end): A worst-case optimal control policy $\pi^* = (\mathbf{a}_0^*, \dots, \mathbf{a}_{N-1}^*)$ is obtained by setting

$$\mathbf{a}_t^*(i) \in \arg \min_{a \in \mathcal{A}} \left(c_t(i, a) + \max_{p \in \mathcal{P}_i^a} p^T v_{t+1} \right), \quad i \in \mathcal{X}, \quad t \in T. \quad (3)$$

(A similar theorem holds in the [infinite-horizon](#) case.)

Sketch of proof (I)

- By **weak duality**, obtain a lower bound

$$\phi_N(\Pi, \mathcal{T}) = \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} C_N(\pi, \tau) \geq \max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} C_N(\pi, \tau).$$

- **Linear program** representation of nominal problem (Puterman, 1994):

$$\min_{\pi \in \Pi} C_N(\pi, \tau) = \max_{\mathbf{v}} q^T \mathbf{v}_0 : \mathbf{v}_t \leq c_t^a + P_t^a \mathbf{v}_{t+1}, \quad a \in \mathcal{A}, \quad t \in T,$$

where q is the distribution of the initial state, and $c_t^a(i) := c_t(i, a)$.

- Hence the lower bound can be formulated as a **nonlinear program**:

$$\max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} C_N(\pi, \tau) = \max_{\mathbf{v}, \tau \in \mathcal{T}} q^T \mathbf{v}_0 : \mathbf{v}_t \leq c_t^a + P_t^a \mathbf{v}_{t+1}, \quad a \in \mathcal{A}, \quad t \in T.$$

Sketch of proof (II)

- Likewise, the expected cost for a **given** controller policy $\pi = (\mathbf{a}_t)_{t \in T}$ is given by the LP

$$C_N(\pi, \tau) = \max_{\mathbf{v}} q^T \mathbf{v}_0 \quad : \quad \mathbf{v}_t \leq \mathbf{c}_t^{\mathbf{a}_t} + P_t^{\mathbf{a}_t} \mathbf{v}_{t+1}, \quad t \in T,$$

where $c_t^{\mathbf{a}_t}(i) := c_t(i, \mathbf{a}_t(i))$, $P_t^{\mathbf{a}_t}(i, j) := P_t^{\mathbf{a}_t(i)}(i, j)$.

- Hence, the **worst-case cost for a given policy** π is given by the nonlinear program

$$\max_{\tau \in \mathcal{I}} C_N(\pi, \tau) = \max_{\mathbf{v}, \tau \in \mathcal{I}} q^T \mathbf{v}_0 \quad : \quad \mathbf{v}_t \leq \mathbf{c}_t^{\mathbf{a}_t} + P_t^{\mathbf{a}_t} \mathbf{v}_{t+1}, \quad t \in T.$$

Sketch of proof (III)

- By a **monotonicity** argument, the two previous nonlinear problems can be solved by the recursions (1) and (2), respectively.
- The robust Bellman recursion (1) provides an optimal value function v^* and via (3), a deterministic policy π^* ; since v^* satisfies the recursion (2) for $\pi = \pi^*$, we have

$$\max_{\tau \in \mathcal{T}} \min_{\pi \in \Pi} C_N(\pi, \tau) = \max_{\tau \in \mathcal{T}} C_N(\pi^*, \tau).$$

- Since π^* is an admissible (that is, a deterministic) policy:

$$\max_{\tau \in \mathcal{T}} C_N(\pi^*, \tau) \geq \min_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} C_N(\pi, \tau) = \phi_N(\Pi, \mathcal{T}),$$

from which **perfect duality follows**. ■

Overview

- The curse of uncertainty in Markov Decision Processes
- Robust control formulation
- Robust dynamic programming recursion
- *Entropy-based uncertainty models*
- Application: air traffic routing

Solving the inner problem

We now consider the inner problem:

$$\max_{p \in \mathcal{P}} p^T v,$$

where v is given, and \mathcal{P} is a given subset of the probability simplex.

- The inner problem is **convex**, no matter what \mathcal{P} is!
- We need to solve it **fast**—complexity depends heavily on the structure of \mathcal{P} .
- \mathcal{P} must represent a **statistically accurate** model of uncertainty on the state transition probabilities.

Accuracy issues

What is the accuracy at which we must solve the inner problem, so as to produce an ϵ -suboptimal controller?

Accuracy issues

What is the accuracy at which we must solve the inner problem, so as to produce an ϵ -suboptimal controller?

Theorem: To compute an ϵ -suboptimal controller, we need to compute the optimal value of each inner problem with an absolute accuracy of at most δ , where

- $\delta = \epsilon/N$ in the finite-horizon case;
- $\delta = (1 - \gamma)\epsilon/2\gamma$ in the infinite-horizon case.

Likelihood models

Consider the likelihood model

$$\mathcal{P} = \left\{ p \in \mathbf{R}^n, \quad p \geq 0, \quad p^T \mathbf{1} = 1, \quad \sum_j f(j) \log p(j) \geq \beta \right\},$$

where β is given (such that $\mathcal{P} \neq \emptyset$), and f contains the empirical probabilities.

- A very natural model of uncertainty, motivated by maximum likelihood methods.
- Such models are typically used to derive intervals or ellipsoids of confidence—they are much more **accurate** than intervals or ellipsoids.

Complexity

Theorem: In the likelihood model, the optimal value of the inner problem can be computed with absolute accuracy δ by bisection in $O(n) \log(1/\delta)$ time.

With respect to the nominal case, this represents a relative increase in computational complexity of $O(1)$ only!

Sketch of proof

(Assume $f > 0$ WLOG)

- The log-likelihood acts as a **barrier** for the probability simplex, hence we can safely drop the sign constraint on p .
- The dual problem then involves only two scalar variables, and we can reduce the dual to a **one-dimensional problem**.
- The 1-D problem can be solved by **bisection**—each step requires $O(n)$ time.
- Care must be taken to make sure we compute the optimal **value** of the problem with the required absolute accuracy δ , as opposed to merely locating a (dual) optimal variable with that accuracy.

Relative entropy models

Similar results hold for models based on **relative entropy** bounds:

$$\mathcal{P} = \{p \in \mathbf{R}^n, p \geq 0, p^T \mathbf{1} = 1, : D(p||f) \leq \gamma\},$$

where $\gamma > 0$ is fixed, $f > 0$ is a given distribution, and $D(p||f)$ denotes the Kullback-Leibler divergence from f to p :

$$D(p||f) := \sum_j p(j) \log \frac{p(j)}{f(j)}$$

(Likelihood models involve $D(f||p)$ instead of $D(p||f)$.)

The best of both worlds

Entropy models take **the best of both worlds**:

- Solving the inner problem incurs a **computational overhead of $O(1)$** with respect to the case with no uncertainty.
- They represent **statistically natural** ways to capture measurement uncertainty on the transition probabilities.

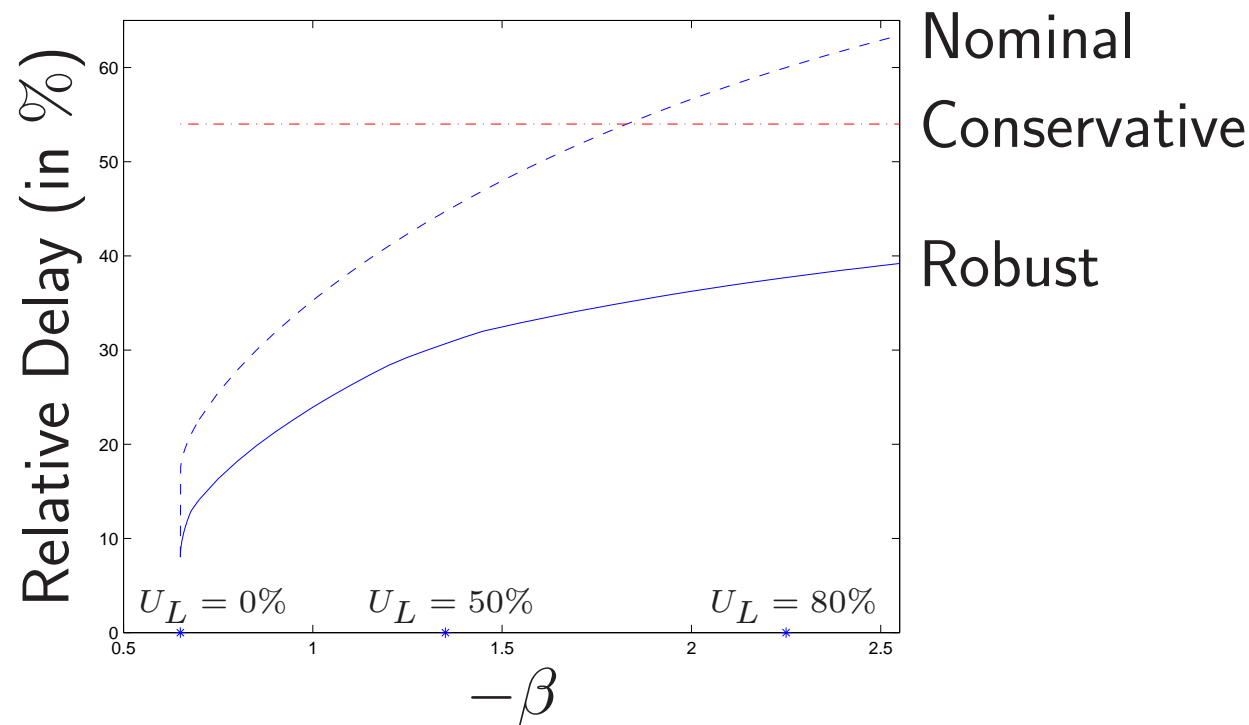
Complexity vs. uncertainty models

<i>Uncertainty model</i>	<i>Computational overhead (relative to nominal)</i>
Likelihood	$O(1)$
Entropy	$O(1)$
MAP	$O(1)$
Interval	$O(\log n)$
Ellipsoid \cap simplex	$O(n^{0.5})$
L Scenarios	$O(L)$

Overview

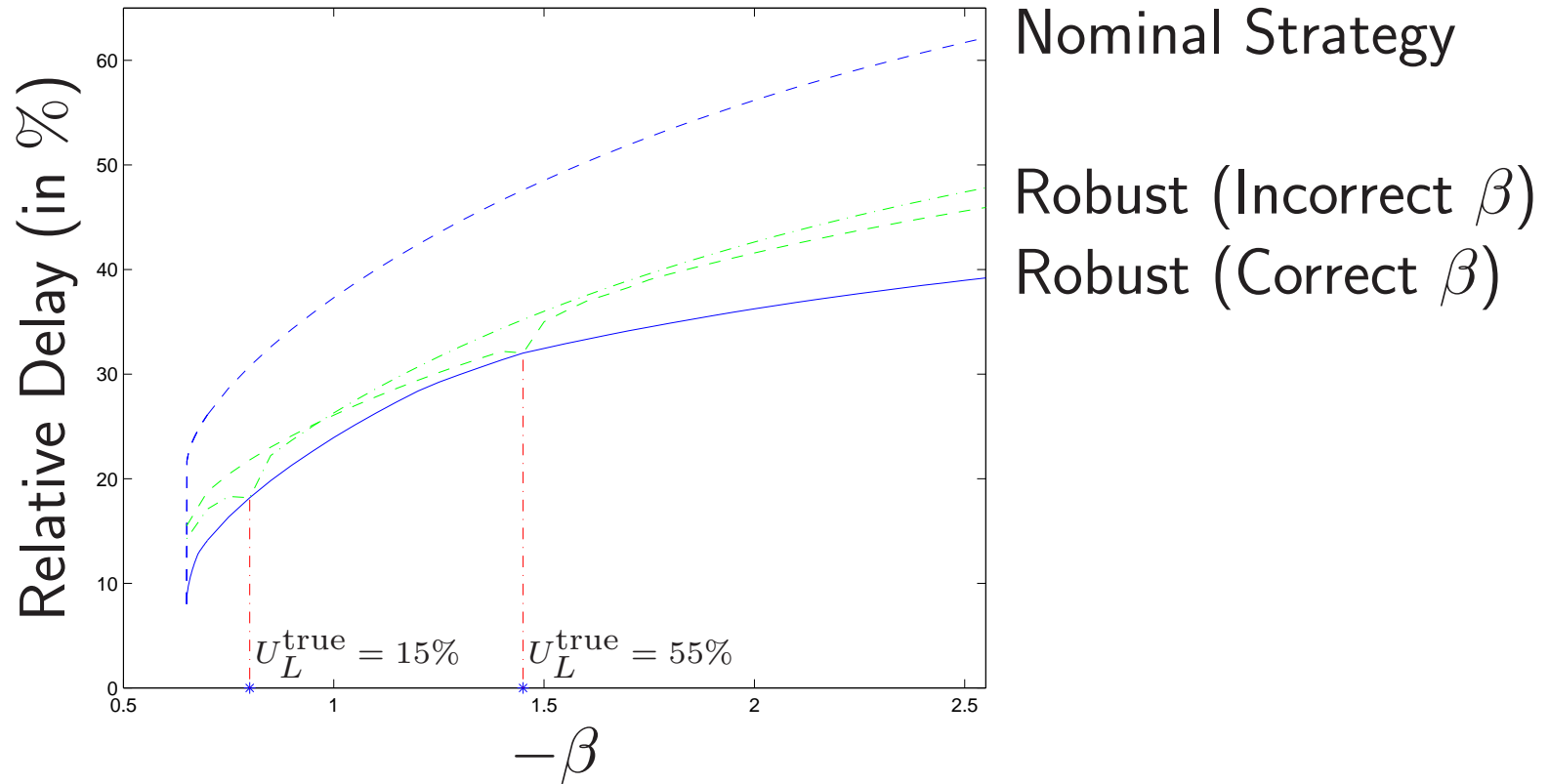
- The curse of uncertainty in Markov Decision Processes
- Robust control formulation
- Robust dynamic programming recursion
- Entropy-based uncertainty models
- *Application: air traffic routing*

Application: air traffic routing



Worst-case delay vs. uncertainty level β (lower bound on the log-likelihood function), for a conservative policy (red), the classical Bellman recursion (dotted blue), and its robust counterpart (solid blue). $1 - U_L$ is a confidence level derived from β using an asymptotically large sample approximation.

Uncertainty on uncertainty level



Worst-case delay vs. uncertainty level β , for the classical Bellman recursion (dotted blue) and its robust counterpart, with correct (solid blue) and incorrect predictions (green) for the uncertainty level β .

Concluding remarks

- We gave a rigorous framework for addressing **parameter uncertainty** in **MDPs**.
- We proved the corresponding **robust recursion** that enables a polynomial time algorithm.
- Uncertainty models based on entropy bounds are not only **statistically natural**—they result in a very **small computational overhead**.
- An air traffic routing example demonstrated that robustness is greatly improved at very little expense in optimality, even if the uncertainty level is only crudely guessed.
- Method applies whenever Bellman recursion is practical—larger-scale problems remain a challenge.