Risk-averse and robust control of Markov Decision Processes

Yann Le Tallec
letallec@mit.edu
(joint work with Duncan Simester and John Tsitsiklis)

Massachusetts Institute of Technology

CORS 2006

Outline

1. Introduction
   - An example from revenue management
   - Literature review
2. Convex dynamically consistent Markovian risk measure
   - Convex single-period risk measure
   - Dynamically consistent multi-period risk measures
   - Markovian multi-period risk measures
   - Markovian risk minimization
3. Robust control of MDPs
   - The classical robust control of MDPs
   - A “new” robust control formulation

A classical revenue management problem

Airline company selling N seats at 2 prices.
At each time period, it observes how many seats remain unsold, and decides how many seats are available in the low price class.

Company wants to maximize its profit, but uncertain purchase process.

Position $X \in \mathbb{R}^\Omega$: $X(\omega)$ cost for uncertain outcome $\omega \in \Omega$
(Ω finite in this talk).

Risk measure $\rho : \mathbb{R}^\Omega \to \mathbb{R}$: preference among positions.

The decision maker solves $\inf_{a \in \mathcal{A}} \rho(X_a)$.

Relevant literature

- Risk-averse control of MDPs with exponential utility
  [Howard and Matheson, *MS*, 1972]

  [Runolfsson, *IEEE Conf. on Decision and Ctrl*, 1993], [Hernández-Hernández and Marcus, *Systems and Ctrl Lett.*, 1996]
- Convex multi-period risk measures
  [Artzner et al, *Math. Fin.*, 1999], and [Föllmer and Schied, *Fin. and Stocha.*, 2002]

  Artzner et al (tech. report 2003), Ruszczynski and Shapiro (tech. report, 2004), Roorda et al (*Math. Fin.*, 2004), Riedel (*Stocha. Proc and Appli.*, 2004), Detlefsen (tech. report, 2005)
- Worst-case robust control of uncertain MDPs
  [Satia and Lave, *OR*, 1973], [Nilim and El Ghaoui, *OR*, 2005], and [Iyengar *Math. OR*, 2005]

Coherent and convex risk measures

Definition

A convex risk measure on the position space $\mathbb{R}^{\Omega}$ is a functional $\rho : \mathbb{R}^{\Omega} \to \mathbb{R}$ that is

1. Monotonic: If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\rho(X) \leq \rho(Y)$.
2. Invariant by translation: $\rho(X + m) = \rho(X) + m$.
3. Convex: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y), \ \lambda \in [0, 1]$.

If for all $\lambda \geq 0$, $\rho(\lambda X) = \lambda\rho(X)$, then ρ is a coherent risk measure.

Given probability space $(\Omega, \mathcal{F}, \mu)$,
the exponential utility function $\rho(X) = -\gamma^{-1} \log E_{\mu} \exp(-\gamma X)$ is a convex risk measure.

Representation theorem for convex risk measures

Theorem

Let ρ be a convex risk measure on $\mathbb{R}^\Omega$. Define a penalty function $\phi(P) = \sup_X E_P[X] - \rho(X)$ and the set of probability measures on Ω $\mathcal{P} := \{P|\phi(P) < +\infty\}$. Then,

$$\rho(X) = \sup_{P\in\mathcal{P}} E_P[X] - \phi(P). \tag{1}$$

If ρ is coherent, $\phi(\mathcal{P}) = 0$.
Conversely, if ρ has the form (1), then ρ is a convex risk measure.

$$\rho(X) = -\gamma^{-1} \log E_\mu \exp(-\gamma X) = \sup_P \left(E_P[X] - D(P\|\mu)\right)$$

Dynamically consistent multi-period risk measure

From now on, multi-period sample space: $\omega = (v_1, \ldots, v_T)$ with finite time horizon T, realized sequentially. At time t, only partial history $\omega_t = (v_1, \ldots, v_t)$ (partially) observed.

A convex multi-period risk measure on $\mathbb{R}^\Omega$ is an arbitrary family of convex risk measures $\rho(\cdot|\omega_t)$ on $\mathbb{R}^\Omega$ indexed by the partial histories ω_t.

Definition

A convex multi-period risk measure is dynamically consistent if

$$[\forall v_{t+1},\ \rho(X|\omega_t v_{t+1}) \leq \rho(Y|\omega_t v_{t+1})] \ \Rightarrow\ \rho(X|\omega_t) \leq \rho(Y|\omega_t).$$

Our model for dynamical system

- Finite *state space* $\mathcal{X}$
- Finite *action space* $\mathcal{A}$
- Finite time horizon T

Sample space $\omega = (v_1, \ldots, v_T) \in \Omega$. v_t describes for all conceivable current state and action at time t the outcome for the next state and transition cost.

Markovian policy π: maps current time and state to an action.
Policy π with ω defines an observed state-action-cost trajectory $(S_t^\pi, A_t^\pi, Q_t^\pi)_{t \geq 1}$.

Tail cost of policy π: $C_\tau^\pi = \sum_{t=\tau}^{T} Q_t^\pi$

Controller minimizes his a priori risk given initial state s_1

$$\inf_{\pi \in \Pi_{m,d}} \rho \left(\sum_{t=1}^{T} Q_t^\pi \,|s_1 \right).$$

Markovian multi-period risk measures

Definition

A multi-period risk measure ρ is called Markovian if for any fixed Markovian policy π and time t,

$$\rho(C_t^\pi | \omega_t) = \rho(C_t^\pi | \bar{\omega}_t) =: \rho(C_t^\pi | s, a)$$

whenever $S_t^\pi(\omega) = S_t^\pi(\bar{\omega}) = s$ and $A_t^\pi(\omega) = A_t^\pi(\bar{\omega}) = a$.

- Can extend state space from all partial histories to any “sufficient statistics” for system’s behavior and risk perception.
- Leverage state space to efficiently optimize over large policy space, and even make infinite horizon problems solvable.

Finite horizon risk minimization theorem

Theorem

Let ρ be a convex dynamically consistent Markovian risk measure. Then $\inf_{\pi\in\Pi_{m,d}} \rho(C_1^\pi | s_1) = V(1, s_1)$ *where*

$$V(T,j) = \min_{a\in\mathcal{A}_j} \sup_{P_{T,j,a}\in\mathcal{P}^{T,j,a}} \left[(\bar{q}_{T,j,a} - \phi_{T,j,a})(P_{T,j,a})\right],$$

$$V(t,j) = \min_{a\in\mathcal{A}_j} \sup_{P_{t,j,a}\in\mathcal{P}^{t,j,a}} \left[(\bar{q}_{t,j,a} - \phi_{t,j,a})(P_{t,j,a}) + \sum_{k\in\mathcal{X}} P(k|t,j,a)V(t+1,k)\right]$$

Shapley-Bellman equations

Optimal to select action that achieves minimum in Shapley equation.

Comments on finite horizon risk minimization

- Risk evaluation can be performed by the theorem recursion for any fixed Markovian policy π.
- Minimizing convex dynamically consistent Markovian risk is equivalent to solving a zero-sum sequential Markov game.
- The minimal risk and optimal policy can be found in $O(\mathcal{X}T)$ vs. $O(\exp(T))$ by enumeration or in the literature

The classical robust control of MDPs

[Satia and Lave, *OR*, 1973], [Nilim and El Ghaoui, *OR*, 2005], and [Iyengar, *Math of OR*, 2005] :
Uncertain transition probabilities $(P(\cdot|t,j,a))_{t,j,a}$ in uncertainty set $\mathcal{P}$.
Nature picks the worst uncertain parameters for the controller's policy.
Zero-sum game between controller and nature

With "rectangular" uncertainty set (i.e. $\mathcal{P} = \prod_{t,j,a} \mathcal{P}_{t,j,a}$):

- sequential Markov game
- Shapley equations

Equivalent to minimizing coherent dynamically consistent Markovian risk ρ with $\rho(C_t^\pi|\omega_t) := \sup_{P\in\mathcal{P}(\omega_t)} E_P[C_t^\pi]$,

$$\mathcal{P}(\omega_t) = \left\{ \left(\prod_{\tau=t}^{T-1} Z_{\omega_\tau} \right) Z_{\omega_T} \text{ with } Z_{\omega_\tau} \in \prod_{(j,a)\in\mathcal{X}\mathcal{A}} \mathcal{P}_{\tau,j,a},\ \tau \geq t \right\}.$$

From convex risk to robust control of MDPs

Classical robust control: nature picks parameters on the boundary of uncertainty set $\Rightarrow$ sensitive to uncertainty set design and most unlikely parameters.

Define risk measure $\rho(\cdot|\omega_t) = \sup_{P\in\mathcal{P}(\omega_t)} E_P[\cdot] - \phi_{\omega_t}(P)$ with penalty functions $\phi_{\omega_r}(P) = E_P \sum_{t=r}^{T} \sum_{(j,a)\in\mathcal{X}\mathcal{A}} \phi_{t,j,a}(P^{t,j,a}_{|\omega_t})$.
ρ is a convex dynamically consistent Markovian risk measure.

“New” Robust control formulation: $\inf_\pi \sup_{P\in\mathcal{P}} E_P[C_t^\pi] - \phi(P)$

Markovian risk minimization theorem applies:

- Equivalence with a zero-sum sequential Markov game
- Shapley-Bellman equations

$$V(t,j) = \inf_{a\in\mathcal{A}_j} \sup_{P_{t,j,a}\in\mathcal{P}^{t,j,a}} \left[(\bar{q}_{t,j,a} - \phi_{t,j,a})(P_{t,j,a}) + \sum_{k\in\mathcal{X}} P(k|t,j,a)V(t+1,k) \right]$$

- Robust control can be solved efficiently

Contributions presented in this talk

- Defined and motivated Markovian multi-period risk measures
- Optimized efficiently convex dynamically consistent Markovian risk measures of sample cost.
- Showed the equivalence of this problem with zero-sum sequential Markov game against nature.
- Proposed a tractable formulation for the robust control of MDPs founded in risk-sensitive decision theory.