Risk-Aware Decision Making and Dynamic Programming

Boris Defourny, Damien Ernst and Louis Wehenkel
Department of Electrical Engineering and Computer Science
University of Liège, Belgium
{Boris.Defourny, dernst, L.Wehenkel}@ulg.ac.be

Abstract

This paper considers sequential decision making problems under uncertainty, the tradeoff between the expected return and the risk of high loss, and methods that use dynamic programming to find optimal policies. It is argued that using Bellman Principle determines how risk considerations on the return can be incorporated. The discussion centers around returns generated by Markov Decision Processes and conclusions concern a large class of methods in Reinforcement Learning.

1 Introduction

Incorporating risk sensitivity in Reinforcement Learning (RL for short) can serve different purposes: to balance exploration versus exploitation for fast convergence, to protect the agent during the learning process, to increase policy robustness by limiting confidence in the model, or to prevent rare undesirable events. This paper deals with the latter aspect.

Many popular methods in RL such as Temporal Difference learning [27] or Q-learning [29] base the selection of actions on average rewards-to-go, following principles from Monte Carlo estimation [21] and Dynamic Programming [4].
The present paper focuses on the dynamic programming part. It discusses how the use of Bellman Principle restrains the user from imposing arbitrary requirements on the distribution of the return generated by a Markov Decision Process [25] (MDP).
It is assumed that the parameters of the MDP are known. Restrictions in that setup will also hold in setups involving estimation issues, observability issues, or complexity issues.

The paper attracts the attention to limitations of dynamic programming beyond the curse of dimensionality, and brings insights on the structure of risk-aware policies. The second point is also of potential interest for non dynamic-programming-based methods, such as policy search methods [20] or scenario-tree-based methods [10].

The paper is organized as follows. Section 2 provides background material on the considered processes, returns and objectives. Section 3 defines an abstract Bellman Principle, with the motivation of identifying in Section 4 limitations that also hold for powerful versions of the Bellman Principle. Section 5 considers the relaxation of clear objectives on the return distribution at the light of the preceding discussion, and Section 6 concludes.

2 Markov Decision Process and Return Risk

2.1 MDP Formulation

The return-generating process is formalized as a finite or an infinite horizon Markov Decision Process with a finite state space $S = \{1, \ldots, |S|\}$, a finite action space $A = \{1, \ldots, |A|\}$, a transition

probability matrix $P \in \mathbb{R}^{|S|\times|A|\times|S|}$ with $P_{ijk} = \mathbb{P}(s' = k|s = i, a = j)$, an initial distribution of state q, and a reward matrix $R \in \mathbb{R}^{|S|\times|A|\times|S|}$ with $R_{ijk} = r(s = i, a = j, s' = k)$.

The return obtained during one realization of the process is (under usual measurability conditions) the random variable

$$\mathcal{R}^{q,\pi} = \sum_{t=0}^{\infty} \beta^t r(x_t, u_t, x_{t+1}), \tag{1}$$

with $\beta \in (0, 1)$ the discount factor. In finite horizon problems the sum is truncated and $\beta \in (0, 1]$. The random variable depends on the initial distribution q and on the policy π of the agent. The superscripts q, π stress that dependence but are omitted in the sequel when the context is clear.

The agent applies a Stochastic Markov policy $\pi \in \Sigma$. Stochastic policies are often relevant in a risk-sensitive context: see Section 4.1. The policy may be time-varying. Even if the process is observable, in many risk-sensitive setups, it is necessary to extend the state space and define auxiliary variables (see Section 2.2). Therefore, let I refer to such an extended state space, and let I be identified to the information space of the agent. The random selection of an action in A according to a distribution conditioned on the information state $i_t \in I$ is written

$$u_t \sim \pi_t(i_t). \tag{2}$$

Provided I is finite and has $|I|$ elements, the policy at time t can be represented by a matrix $\Pi^{(t)} \in \mathbb{R}^{|I|\times|A|}$ with $\Pi^{(t)}_{ij} = \mathbb{P}(a = j|i_t = i)$. For deterministic policies, $\Pi^{(t)}$ has 0's and 1's.

Let Φ be an functional that maps random variables to $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. Examples of choice for Φ are given in Section 2.2. Φ can also be viewed as a mapping from return distributions to $\overline{\mathbb{R}}$.
The goal of the agent is the maximization of $\Phi(\mathcal{R}^{q,\pi})$ over his policy space:

$$\max_{\pi\in\Sigma} \Phi\left(\mathcal{R}|x_0 \sim q, \pi\right). \tag{3}$$

2.2 Popular Objectives in the MDP Literature

When $\Phi(\mathcal{R}) = \mathbb{E}\{\mathcal{R}\}$, the objective reduces to the risk-neutral maximization of the expected return.

A first important choice in risk-sensitive MDP originally proposed by [17] (undiscounted case) and further studied by [18, 8] (discounted case) is

$$\Phi(\mathcal{R}) = -\gamma^{-1} \log(\mathbb{E}\{\exp(-\gamma\mathcal{R})\}). \tag{4}$$

The parameter γ models a constant risk sensitivity that corresponds to risk aversion if $\gamma > 0$. Constant risk sensitivity means that the sensitivity does not vary with the current wealth of the agent, here defined as the current partial sum of rewards. Solving methods often exploit dynamic programming (but see [11] for a counterexample), and this explains why (4) is often investigated in various extensions of the basic MDP setup [13, 23].

Another important choice initiated by [6] and refined by [30] is

$$\Phi(\mathcal{R}) = \mathbb{P}(\mathcal{R} \geq a) = \mathbb{E}\{I_{\geq a}(\mathcal{R})\}. \tag{5}$$

The parameter a is a target level for the return and $I_{\geq a}(x)$ denotes the 0-1 indicator function for the event $x \geq a$. Optimal policies have generally to take into account the agent's wealth:

$$R_t = \sum_{\tau=0}^{t-1} \beta^\tau r(x_\tau, u_\tau, x_{\tau+1}) \quad \in \mathbb{R},$$

allowing the agent to convert rewards-to-go to returns.
Several variants of the setup exploit (5), e.g. optimal stopping problems [22].

Some authors have considered more complex functionals, but finding optimal policies seems challenging. For instance, [19] studies optimality conditions for the maximization of $\mathbb{E}\{g(\mathcal{R})\}$ subject to $\mathbb{E}\{h_j(\mathcal{R})\} \leq \alpha_j$, where g and h_j, $1 \leq j \leq k$, are utility functions, and α_j some thresholds. This amounts to choose

$$\Phi(\mathcal{R}) = \begin{cases} \mathbb{E}\{g(\mathcal{R})\} & \text{if } \mathbb{E}\{h_j(\mathcal{R})\} \leq \alpha_j, \quad 1 \leq j \leq k \\ -\infty & \text{otherwise.} \end{cases} \tag{6}$$

2.3 Return Distribution

The distribution function of the return $\mathcal{R}^\pi$ generated by the MDP with policy π and initial state s,

$$F_s(y) \triangleq \mathbb{P}(\mathcal{R}^\pi \leq y | x_0 = s),$$

can be estimated by simulating trajectories of the process and evaluating the corresponding returns. Empirical distributions converge to the theoretical distribution as the number of samples grows to infinity [14].

For the particular case of an infinite horizon MDP, nonnegative bounded rewards, and a stationary policy, it holds (see [8]) that $\boldsymbol{F}(\cdot) \triangleq [F_1(\cdot), \ldots, F_{|S|}(\cdot)]$ is the unique fixed point of a *nonexpansive* mapping M applied on distribution functions $\boldsymbol{G}(\cdot) \triangleq [G_1(\cdot), \ldots, G_{|S|}(\cdot)]$ and defined by

$$(M\boldsymbol{G})_i(y) = \sum_{k=1}^{|S|} P_{ijk}\ G_k(\frac{y - R_{ijk}}{\beta})\Big|_{j=\pi(i)}$$

with $G_i(y) \triangleq 0$ if $y < 0$, and $G_i(y) = 1$ if $y > \sup_{ijk}\{R_{ijk}\}/(1-\beta)$. The mapping is not contractive in general but iterative methods adapted to nonexpansive mappings are able to converge to the fixed point (see again [8]). Obviously, in practical computations $\boldsymbol{G}$ is evaluated on a finite set of values spanning the domain of y.

3 Bellman Principle

3.1 Formulation for the Risk-Neutral MDP

For the maximization of the expected return of an infinite horizon MDP with discount factor β, the dynamic programming principle (*parts of an optimal path are also optimal*) translates into iterations running backwards in time with respect to the state transitions:

$$\begin{aligned} Q(s,a) &\leftarrow \mathbb{E}\{r(s,a,s') + \beta V(s')\} && \forall\ s,\ a \in A(s), \\ V(s) &\leftarrow \max_{a \in A} Q(s,a) && \forall\ s. \end{aligned}$$

The iterations act as a contractive mapping, the value function V converges to its unique fixed point [3], and an optimal policy can be derived from the Q-function: $\pi^*(s) \in \arg\max_a Q(s,a)$.

For the MDP over a finite horizon T, the value function is initialized from the terminal rewards and indexed by time T, and an optimal policy can be derived from time-indexed Q-functions obtained through T iterations, using $\pi_t(s) \in \arg\max_a Q_t(s,a)$.

3.2 Abstract Formulation

It will be fruitful, besides the usual contraction mapping abstraction, to view the application of the Bellman Principle as a step-by-step, recursive optimization of a function $\mathcal{V}_t : I \rightarrow \Upsilon$ with I the information space of the agent, and Υ standing for $\overline{\mathbb{R}}$ (in the conventional setup), or $\overline{\mathbb{R}}^k$ (in the spirit of the multi-criterion setup of [15]), or even the infinite-dimensional space of probability distributions (section 4.1). The space of such functions is noted $S_{\mathcal{V}}$.

An intermediary function $\mathcal{Q}_t : I \times A \rightarrow \Upsilon$, belonging to a space $S_{\mathcal{Q}}$, is derived pointwisely from $\mathcal{V}_{t+1}$, using a family of operators $\psi_{i,a} : S_{\mathcal{V}} \rightarrow \Upsilon$ indexed by $(i,a) \in I \times A$.

Another family of operators $\phi_{i,a} : S_{\mathcal{Q}} \rightarrow \overline{\mathbb{R}}$ indexed by $(i,a) \in I \times A$ is also introduced. The goal is to induce for each state in I a complete ordering of the actions in A. Usually, $\phi_{i,a}$ simply extracts from $\mathcal{Q}_t$ the value $\mathcal{Q}_t(i,a) \in \Upsilon$ and maps it to a scalar score. With the more abstract domain definition on $S_{\mathcal{Q}}$, a joint optimization over the actions $a(i)$ using all the values $\mathcal{Q}_t(i,a)$ is allowed.

After all these preliminary definitions, the Bellman iteration can now be expressed as

$$\begin{aligned} \mathcal{Q}_t(i,a) &\leftarrow \psi_{i,a}(\mathcal{V}_{t+1}) && \forall\ i,\ a \in A(i) \\ a^*(i) &\leftarrow \arg\sup_{a \in A(i)} \phi_{i,a}(\mathcal{Q}_t) && \forall\ i \\ \mathcal{V}_t(i) &\leftarrow \mathcal{Q}_t(i, a^*(i)). \end{aligned} \tag{7}$$

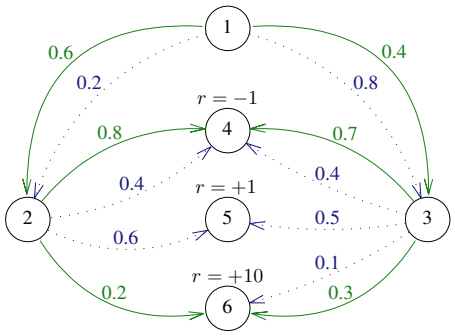

Figure 1: Transition probabilities and returns of the finite horizon MDP of section 4.1. Plain lines are relative to state transitions under action $a = 1$, dotted lines concern $a = 2$.

The information state $i = i(t)$ encodes information on the current history $[x_0, u_0, r_0, x_1, u_1, r_1, \ldots, x_t]$ of *one* particular realization of the decision process.
$\mathcal{V}_{t+1}$ encodes information on the reward-to-go distribution, and thus on the return distribution, conditionally to $i(t+1)$.
The best actions $a^*(i)$ are selected through the operators $\phi_{i,a}$ with ties broken arbitrarily.

A last comment concerns the actions a themselves. It is possible to redefine a as a particular assignment of probabilities of actions in A. By such an extension of the action space, stochastic policies can be derived from the Q-functions.

4 Limitation of Bellman Agents

An agent that solves the abstract recursion (7) step by step is referred to as a Bellman Agent. An example developed in the formalism of Section 3.2 will first illustrate how arbitrary risk constraints may prevent an agent from being a Bellman Agent. More general situations are then discussed in Section 4.2

4.1 Toy Example

Consider the process depicted on Figure 1 with 6 states and 2 actions. The process starts from state 1 at time $t = 0$. The agent takes decisions at times $t = 0, 1$, and obtains a return $\mathcal{R}^\pi$ when he enters at time $t = 2$ one of the terminal states $s = 4, 5, 6$ (resp. $r = -1, +1, +10$ as shown on the figure). Let the agent maximize $\mathbb{E}\{\mathcal{R}^\pi\}$ subject to $\mathbb{P}(\mathcal{R}^\pi > 0) \geq 0.5$.

First, deterministic policies are considered. A direct policy search over the 8 possible policies (an assignment of actions 1 or 2 to states 1,2,3) indicates that the best choice is

$$\pi^\dagger(1) = 2, \pi^\dagger(2) = 1, \pi^\dagger(3) = 2.$$

The policy yields $\mathbb{E}\{\mathcal{R}^{\pi^\dagger}\} = 1.12$ with $\mathbb{P}(\mathcal{R}^{\pi^\dagger} > 0) = 0.52 \geq 0.5$.

Next, stochastic policies are considered. An exact calculation developed below shows that an optimal policy is

$$\pi^\star(1) = 2, \pi^\star(2) = 2, \pi^\star(3) = a_3$$

with a_3 a random action in {1,2} such that $\mathbb{P}\{a_3 = 1\} = 5/12$.
The policy yields $\mathbb{E}\{\mathcal{R}^{\pi^\star}\} = 1.32$ with $\mathbb{P}(\mathcal{R}^{\pi^\star} > 0) = 0.5$.
There is an improvement with respect to the deterministic policy.

In this example, exact optimal policies are beyond the rationality of the Bellman agent. To see this, let $\mathcal{Q}_t(j, p_j)$ be the reward-to-go *distribution* from state j, computed at the iteration relative to time t, with the "action" consisting in selecting $a = 1$ with probability p_j and $a = 2$ with probability $(1 - p_j)$. Deterministic policies have $p_j \in \{0, 1\}$ and stochastic policies have $p_j \in [0, 1]$. Remark that here the distributions $\mathcal{Q}_t(j, p_j)$ also correspond to return distributions.

At the iteration relative to time $t = 1$,

$$\mathcal{Q}_1(1, p_1) = 0 \quad \forall\, p_1 \text{ (rewards from successor states have not been propagated yet)},$$

$$\mathcal{Q}_1(2, p_2) = \begin{cases} -1 & \text{with probability} \quad 0.8\ p_2 + 0.4\ (1 - p_2) \\ +1 & \qquad\qquad\quad 0.6\ (1 - p_2) \\ +10 & \qquad\qquad\quad 0.2\ p_2, \end{cases}$$

$$\mathcal{Q}_1(3, p_3) = \begin{cases} -1 & \text{with probability} \quad 0.7\ p_3 + 0.4\ (1 - p_3) \\ +1 & \qquad\qquad\quad 0.5\ (1 - p_3) \\ +10 & \qquad\qquad\quad 0.3\ p_3 + 0.1\ (1 - p_3). \end{cases}$$

But it is easy to realize that there is no choice for p_2 and p_3, even made jointly, that can ensure optimality independently of p_1. The constrained problem has to be solved entirely at the iteration relative to time $t = 0$. By mixing the distributions $\mathcal{Q}_1(2, p_2)$ and $\mathcal{Q}_1(3, p_3)$ with weights $(0.6\ p_1 + 0.2\ (1 - p_1))$ and $(0.4\ p_1 + 0.8(1 - p_1))$ respectively, we get the return distribution conditionally to $s = 1$ and the choice of parameters p_j:

$$\mathcal{Q}_0(1, \{p_j\}_{1 \leq j \leq 3}) = \begin{cases} -1 & \text{with probability} \quad (10 + 2\ p_2 + 4\ p_1\ p_2 + 6\ p_3 - 3\ p_1\ p_3)/25 \\ +1 & \qquad\qquad\quad (13 + p_1 - 3\ p_2 - 6\ p_1\ p_2 - 10\ p_3 + 5\ p_1\ p_3)/25 \\ +10 & \qquad\qquad\quad (2 - p_1 + p_2 + 2\ p_1\ p_2 + 4\ p_3 - 2\ p_1\ p_3). \end{cases}$$

From $\mathcal{Q}_0(s, \{p_j\}_{1 \leq j \leq 3})$ and the initial state distribution concentrated on $s = 1$, comes the complete program for policy search over the policy parameters p_1, p_2, p_3:

$$\begin{aligned} \text{maximize } \mathbb{E}\{\mathcal{R}\} \equiv & (23 - 9\ p_1 + 5\ p_2 + 10\ p_1\ p_2 + 24\ p_3 - 12\ p_1\ p_3)/25 \\ \text{subject to } \mathbb{P}\{\mathcal{R} \geq 0\} \equiv & (15 - 2\ p_2 - 4\ p_1\ p_2 - 6\ p_3 + 3\ p_1\ p_3)/25 \geq 1/2 \\ & p_1, p_2, p_3 \in \{0, 1\} \text{ for deterministic policies} \\ & p_1, p_2, p_3 \in [0, 1] \text{ for stochastic policies.} \end{aligned}$$

This nonlinear program can be solved, but the purpose of Bellman principle would have been to decompose the problem.

4.2 Discussion

The Bellman decomposition fails in the preceding example because the probability that the process reaches the states $s = 2$ and $s = 3$ at time $t = 1$ influences the ranking of actions.

4.2.1 Bellman Compatibility Condition

Now let $\mathbb{P}(i_t = i')$ denote the prior probability that the agent reaches some information state $i' \in I$ at time t. In general, that probability is function of the initial state distribution and of the policy at times $0, 1, \ldots, t - 1$. Obviously, the selection of actions (represented by $\phi_{i,a}$ in (7)) cannot depend on $\mathbb{P}(i_t = i')$, since the Bellman iteration proceeds backwards in time to optimize the actions. The restriction can be expressed by the compact notation

$$\phi_{i,a} \perp \mathbb{P}(i_t = i') \quad \forall\, t > 0,\ \forall i' \in I. \tag{8}$$

In other words the actions optimized at time t can at most reflect preferences over return distributions *conditioned on* $i_t = i$. Those distributions are noted $\mathcal{R}|i_t = i$.

In short, the risk-sensitive objective Φ in (3) must be compatible with a greedy selection of the conditional return distributions $\mathcal{R}|i_t = i$.

For the objective (5) of Section 2.2, selecting the conditional return distribution $\mathcal{R}|i_t = i$ that maximizes $\mathbb{P}(\mathcal{R} \geq a \mid i_t = i)$ will also maximize $\mathbb{P}(\mathcal{R} \geq a)$. In contrast, for the objective (6), having $\mathbb{E}\{h_j(\mathcal{R})|i_t = i)\} \leq \alpha_j$ for all i ensures that $\mathbb{E}\{h_j(\mathcal{R})\} \leq \alpha_j$ but in a conservative way. Fixing for each i the balance between increasing $\mathbb{E}\{g(\mathcal{R})|i_t = i\}$ and decreasing $\mathbb{E}\{h_j(\mathcal{R})|i_t = i\}$ has a variable effect on $\mathbb{E}\{h_j(\mathcal{R})\}$ which depends on $\mathbb{P}(i_t = i')$, $i' \in I$. And the optimal balance cannot be guessed using the Bellman Principle.

4.2.2 Class of Admissible Functionals

Condition (8) turns out to be highly restrictive with respect to candidate risk-averse criteria for Φ. Usual requirements against tail events, such as those introduced in [26, 1], that have been well welcomed by practitioners [28] and are well adapted to convex programming methods [7], are beyond the rationality of the Bellman Agent.

Recent work [2] in the formalism of Stochastic Processes suggests that the applicability of Bellman Principle for the risk-aware optimization of a random final value under a non-anticipative policy is restricted to a small class of functionals Φ. A final value refers in principle to a unique final reward, but state augmentation techniques can generalize the scope of the statement. There is a set of equivalent conditions or properties that define the class of admissible functionals: see [2]. At the end of the day, the class mainly consists of functionals of the form

$$\Phi(\mathcal{R}_T) = f^{-1}(\mathbb{E}\{f(\mathcal{R}_T)\}) \tag{9}$$

with $\mathcal{R}_T$ the random final value, f a strictly increasing function, and f^{-1} its reciprocal (see also [24]).
This does not leave many degrees of freedom, besides the choice of f, to try to express risk sensitivity.

4.2.3 Critique of Arbitrary Risk Criteria on the Return

On the other hand, a risk-averse criterion that ignores the condition (8) is questionable. Such a criterion induces decisions sensitive to events $\{i_t = i'\}$, with $i' \neq i$, that have a null probability (i.e. *won't* happen) when the information state is i. It may seem paradoxical that an agent being for sure in some information state at time t needs to consider the a priori distribution of i_t (i.e. ignoring current information). There are in fact works in utility theory that take the recursive property of the Bellman principle as an axiomatic time-consistency property [12]. The functional Φ is not defined explicitly; instead, it is assumed that the value V_t of a reward trajectory $r_\omega \triangleq [r_0(\omega)\ r_1(\omega)\ \ldots]$ always satisfies a set of recursive relations

$$V_t(r_\omega) = \min_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\{\sum_{s=t}^{\tau-1} \beta^{s-t} g(r_t(\omega)) + \beta^{\tau-t} V_\tau(r_\omega)|i_t\}, \qquad \forall \tau > t \tag{10}$$

with $\mathcal{P}$ the set of distributions for the reward process compatible with the state dynamics and the admissible policies. The optimization of a policy would then take place without a clear target criterion on the return distribution, with the utility g serving as a heuristic.

5 Bellman Decision Process

Section 4.2.2 suggests that optimizing functionals of the return distribution by dynamic programming is impossible at the exception of some particular and already well-studied cases, such as the expectation or the exponential-utility-based functional (4).
At the same time, Section 4.2.3 emphasizes that other functionals of the return distribution induce inconsistent behaviors. The recognition of this fact is important because these functionals are commonly optimized in rolling horizon setups, using direct policy search or tools from convex programming. The resulting situation is frustrating. In the context of portfolio management applications for instance, [5] critics common risk measures in the name of time consistency, but ends up proposing the well-studied criterion (5).

Yet there might still be room for alternative time-consistent risk-aware strategies if one is ready to consider suboptimal decision processes, dropping explicit requirements on the distribution of the return.

5.1 Proposal

Following the spirit of (10), let a *Bellman Decision Process* be defined as a process where the agent optimizes in a backward fashion arbitrary functionals of the reward-to-go distributions conditioned on the information states. If the state comprises the current wealth, rewards-to-go can be translated to returns. The process serves as a heuristic to induce risk-aware behaviors in a more consistent way than the direct optimization of a functional of the return distribution as in (3).

5.2 Toy Example (Continued)

Coming back to the example of Section 4.1, let us greedily maximize $\mathbb{E}\{\mathcal{Q}_1(j, p_j)\}$ subject to $\mathbb{P}(\mathcal{Q}_1(j, p_j) > 0) \geq 0.5$ with $j = 2, 3$. The motivation behind the conservative constraints imposed on every state at time $t = 1$ is the automatic satisfaction of the constraint on the return at time

$t = 0$, irrespective of the probabilities of states 2 and 3. The optimal solutions are $p_2 = 1/4$ and $p_3 = 1/3$. Then let us maximize $\mathbb{E}\{\mathcal{Q}_0(1, p_1)\}$ over p_1. The constraint $\mathbb{P}(\mathcal{Q}_0(1, \cdot) > 0) \geq 0.5$ is always satisfied. The optimal solution is $p_1 = 0$. The resulting stochastic policy is

$$\pi^B(1) = 2, \pi^B(2) = a_2, \pi^B(3) = a_3,$$

with a_2, a_3 random actions such that $\mathbb{P}(a_2 = 1) = 1/4$ and $\mathbb{P}(a_3 = 1) = 1/3$.
The policy yields $\mathbb{E}\{\mathcal{R}^{\pi^B}\} = 1.29$ with $\mathbb{P}(\mathcal{R}^{\pi^B} > 0) = 0.5$.
This performance is close to the value 1.32 of Section 4.1.

If only deterministic policies are considered, the restriction $p_j \in \{0, 1\}$ leads to $p_1 = p_2 = p_3 = 0$. The corresponding policy yields 0.92. Here the regret with respect to the deterministic policy $\pi^\dagger$ of Section 4.1 is substantial (but recall that the optimization criterion was altered).

5.3 Approximate Distributions and Particles

The Bellman Decision Process assumes that conditional distributions can be computed by the agent during the optimization of the policy. In a practical implementation, the distributions can be approximated by a set of N samples, with the parameter N bounding the complexity of the representation. The same technique applied to the nonlinear filtering problem is known as particle filtering [9].

The description of the implementation is put asides due to the lack of space and for the sake of a higher level discussion. Experiments on a capital growth problem inspired from [31] allowed us to assess the feasibility and interest of the particle approach. Various unexpected issues were solved by resorting to variance reduction techniques [16].

6 Conclusions

Surprisingly, there are few sound ways for optimizing returns generated by a dynamical process in a risk-aware manner. The limitations are theoretical for the essential part.

These findings suggest that custom risk control requirements should be mostly enforced heuristically, by altering policy optimization procedures and checking the compliance of the policies with the initial requirements. The regret caused by such a procedure is offset by the improved time consistency properties of the policy.

Numerical experiments using the ideas developed in Section 5 appear as extremely promising. Like the MDP, the proposed Bellman Decision Process framework has its complexity independent of the time horizon, giving it an advantage over methods that cannot use time decomposition for optimizing policies.

Acknowledgments

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. Damien Ernst is a Research Associate of the Belgian FNRS of which he acknowledges the financial support.

References

[1] C. Acerbi. Spectral measures of risk: a coherent representation of subjective risk aversion. *J. of Banking and Finance*, 26(7):1505–1518, 2002.

[2] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and Bellman principle. *Annals of Operations Research*, 152(1):5–22, 2007.

[3] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[4] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 3rd edition, 2005.

[5] K. Boda and J.A. Filar. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63:169–186, 2006.

[6] M. Bouakiz and Y. Kebir. Target-level criterion in Markov decision processes. *J. of Optimization Theory and Applications*, 86(1):1–15, 1995.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[8] K.-J. Chung and M. Sobel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM J. Control and Optimization*, 25(1):49–62, 1987.

[9] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans. Signal Processing*, 50(3):736–746, 2002.

[10] B. Defourny, D. Ernst, and L. Wehenkel. Lazy planning under uncertainty by optimizing decisions on an ensemble of incomplete disturbance trees. In *Recent Advances in Reinforcement Learning, 8th European Workshop, EWRL'08*, LNCS (LNAI) 5323. Springer, 2008.

[11] E.V. Denardo and U.G. Rothblum. Optimal stopping, exponential utility, and linear programming. *Math. Programming*, 16:228–244, 1979.

[12] L. Epstein and M. Schneider. Recursive multiple-priors. *J. of Economic Theory*, 113:1–13, 2003.

[13] E. Fernandez-Gaucherand and S. Marcus. Risk-sensitive optimal control of hidden Markov models: structural results. *IEEE Transactions on Automatic Control*, 42(10):1418–1422, 1997.

[14] R. Fortet and E. Mourier. Convergence de la répartition empirique vers la répartition théorique. *Annales scientifiques de l'Ecole Normale Supérieure, Sér. 3*, 70(3):267–285, 1953.

[15] Z. Gabor, Z. Kalmar, and C. Szepesvari. Multi-criteria reinforcement learning. In *Proc. of the 15th Int. Conf. on Machine Learning*, pages 197–205. Morgan Kaufmann, 1998.

[16] P. Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53 of *Applications of Mathematics (Stochastic Modelling and Applied Probability)*. Springer, 2004.

[17] R. Howard and J. Matheson. Risk-sensitive Markov Decision Processes. *Management Science*, 18(7):356–369, 1972.

[18] S.C. Jacquette. A utility criterion for Markov Decision Processes. *Management Science*, 23(1):43–49, 1976.

[19] Y. Kadota, M. Kurano, and M. Yasuda. Discounted Markov decision processes with utility constraints. *Computers and Mathematics with Applications*, 51(2):279–284, 2006.

[20] S. Mannor, R. Rubinstein, and Y. Gat. The cross entropy method for fast policy search. In *Proc. of the 20th Int. Conf. on Machine Learning*, pages 512–519. Morgan Kaufmann, 2003.

[21] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Stat. Assoc.*, 44(247):335–341, 1949.

[22] Y. Ohtsubo. Value iteration methods in risk minimizing stopping problems. *J. of Computational and Applied Mathematics*, 152(1-2):427–439, 2003.

[23] S. Patek. On terminating Markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.

[24] G.Ch. Pflug and W. Römisch. *Modeling, Measuring and Managing Risk*. World Scientific Publishing Company, 2007.

[25] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[26] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *J. of Risk*, 2(3):21–41, 2000.

[27] R.S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[28] N. Topaloglou, H. Vladimirou, and S. Zenios. CVaR models with selective hedging for international asset allocation. *J. of Banking and Finance*, 26(7):1535–1561, 2002.

[29] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[30] C. Wu and Y. Lin. Minimizing risk models in Markov decision processes with policies depending on target values. *J. of Mathematical Analysis and Applications*, 231:47–67, 1999.

[31] R. Ziemba and W.T. Ziemba. *Scenarios for risk management and global investment strategies*. Wiley Finance, 2007.