Some Finitely Additive Dynamic Programming

Bill Sudderth
University of Minnesota

Discounted Dynamic Programming

Five ingredients: S, A, r, q, β.

S - state space

A - set of actions

$q(\cdot|s,a)$ - law of motion

$r(s,a)$ - daily reward function (bounded, real-valued)

$\beta \in [0,1)$ - discount factor

Play of the game

You begin at some state $s_1 \in S$, select an action $a_1 \in A$, and receive a reward $r(s_1, a_1)$.

You then move to a new state s_2 with distribution $q(\cdot|s_1, a_1)$, select $a_2 \in A$, and receive $\beta \cdot r(s_2, a_2)$.

Then you move to s_3 with distribution $q(\cdot|s_2, a_2)$, select $a_3 \in A$, receive $\beta^2 \cdot r(s_3, a_3)$. And so on.

Your total reward is the expected value of

$$\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n).$$

Plans and Rewards

A **plan** π selects each action a_n, possibly at random, as a function of the history $(s_1, a_1, \ldots, a_{n-1}, s_n)$. The **reward** from π at the initial state
$s_1 = s$ is

$$V(\pi)(s) = E_{\pi,s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)].$$

Given $s_1 = s$ and $a_1 = a$, the conditional plan $\pi[s, a]$ is just the continuation of π and

$$V(\pi)(s) = \int [r(s,a) + \beta \int V(\pi[s,a])(t)\, q(dt|s,a)]\pi(s)(da).$$

The Optimal Reward and the Bellman Equation

The **optimal reward** at s is

$$V^*(s) = \sup_{\pi} V(\pi)(s).$$

The **Bellman Equation** for V^* is

$$V^*(s) = \sup_{a} [r(s, a) + \beta \int V^*(t)\, q(dt|s, a)].$$

I will sketch the proof for S and A countable.

Proof of $\leq$:

For every plan π and $s \in S$,

$$\begin{aligned} V(\pi)(s) &= \int [r(s,a) + \beta \int V(\pi[s,a])(t)\, q(dt|s,a)]\pi(s)(da) \\ &\leq \sup_{a'} [r(s,a') + \beta \int V(\pi[s,a'])(t)\, q(dt|s,a')] \\ &\leq \sup_{a'} [r(s,a') + \beta \int V^*(t)\, q(dt|s,a')]. \end{aligned}$$

Now take the sup over π.

Proof of $\geq$: Fix $\epsilon > 0$.

For every state $t \in S$, select a plan π_t such that

$$V(\pi_t)(t) \geq V^*(t) - \epsilon/2.$$

Fix a state s and choose an action a such that

$$r(s,a) + \beta \int V^*(t)\, q(dt|s,a) \geq \sup_{a'} [r(s,a') + \beta \int V^*(t)\, q(dt|s,a')] - \epsilon/2.$$

Define the plan π at $s_1 = s$ to have first action a and conditional plans $\pi[s,a](t) = \pi_t$. Then

$$\begin{aligned} V^*(s) \geq V(\pi)(s) &= r(s,a) + \beta \int V(\pi_t)(t)\, q(dt|s,a) \\ &\geq \sup_{a'} [r(s,a') + \beta \int V^*(t)\, q(dt|s,a')] - \epsilon. \end{aligned}$$

Measurable Dynamic Programming

The first formulation of dynamic programming in a general measure theoretic setting was given by Blackwell (1965). He assumed:

1. S and A are Borel subsets of a Polish space (say, a Euclidean space).

2. The reward function $r(s, a)$ is Borel measurable.

3. The law of motion $q(\cdot|s, a)$ is a regular conditional distribution.

Plans are required to select actions in a Borel measurable way.

Measurability Problems

In his 1965 paper, Blackwell showed by example that for a Borel measurable dynamic programming problem:

The optimal reward function $V^*(\cdot)$ need not be Borel measurable and good Borel measurable plans need not exist.

This led to nontrivial work by a number of mathematicians including R. Strauch, D. Freedman, M. Orkin, D. Bertsekas, S. Shreve, and Blackwell himself. It follows from their work that for a Borel problem:

The optimal reward function $V^*(\cdot)$ is universally measurable and that there do exist good universally measurable plans.

The Bellman Equation Again

The equation still holds, but a proof requires a lot of measure theory. See, for example, chapter 7 of Bertsekas and Shreve (1978) - about 85 pages.

Some additional results are needed to measurably select the π_t in the proof of $\geq$. See Feinberg (1996).

The proof works exactly as given in a finitely additive setting, and it works for general sets S and A.

Finitely Additive Probability

Let γ be a finitely additive probability defined on a sigma-field of subsets of some set F. The integral

$$\int \phi \, d\gamma$$

of a simple function is defined in the usual way. The integral

$$\int \psi \, d\gamma$$

of a bounded, measurable function ψ is defined by squeezing with simple functions.

If γ is defined on the sigma-field $\mathcal{F}$ of **all** subsets of F, it is called a **gamble** and $\int \psi \, d\gamma$ is defined for all bounded, real-valued functions ψ.

Finitely Additive Processes

Let $G(F)$ be the set of all gambles on F. A **strategy** σ is a sequence $\sigma_1, \sigma_2, \ldots$ such that $\sigma_1 \in G(F)$ and for $n \geq 2$, σ_n is a mapping from F^{n-1} to $G(F)$. Every strategy σ naturally determines a finitely additive probability P_σ on the product sigma-field $\mathcal{F}^{\mathbb{N}}$. (Dubins and Savage (1965), Dubins (1974), and Purves and Sudderth (1976))

P_σ is regarded as the distribution of a random sequence

$$f_1, f_2, \ldots, f_n, \ldots.$$

Here f_1 has distribution σ_1 and, given $f_1, f_2, \ldots, f_{n-1}$, the conditional distribution of f_n is $\sigma_n(f_1, f_2, \ldots, f_{n-1})$.

Finitely Additive Dynamic Programming

For each (s, a), $q(\cdot|s, a)$ is a gamble on S. A plan π chooses actions using gambles on A.

Each π together with q and an initial state $s_1 = s$ determines a strategy $\sigma = \sigma(s, \pi)$ on $(A \times S)^{\mathbb{N}}$. For $D \subseteq A \times S$,

$$\sigma_1(D) = \int q(D_a|s, a)\, \pi_1(da)$$

and

$$\sigma_{n-1}(a_1, s_2, \ldots, a_{n-1}, s_n)(D) = \int q(D_a|s_n, a)\, \pi(a_1, s_2, \ldots, a_{n-1}, s_n)(da).$$

Let

$$P_{\pi,s} = P_\sigma.$$

Rewards and the Bellman Equation

For any bounded, real-valued reward function r, the reward for a plan π is well-defined by the same formula as before:

$$V(\pi)(s) = E_{\pi,s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)].$$

Also as before, the optimal reward function is

$$V^*(s) = \sup_{\pi} V(\pi)(s).$$

The Bellman equation

$$V^*(s) = \sup_{a} [r(s,a) + \beta \int V^*(t)\, q(dt|s,a)].$$

can be proved **exactly** as in the discrete case.

Blackwell Operators

Let $\mathbb{B}$ be the Banach space of bounded functions $x : S \mapsto \mathbb{R}$ equipped with the supremum norm.

For each function $f : S \mapsto A$, define the operator T_f for elements $x \in \mathbb{B}$ by

$$(T_f x)(s) = r(s, f(s)) + \beta \int x(s')\, q(ds'|s, f(s)).$$

Also define the operator T^* by

$$(T^* x)(s) = \sup_a [r(s, a) + \beta \int x(s')\, q(ds'|s, a)].$$

This definition of T^* makes sense in the finitely additive case, and in the countably additive case when S is countable. There is trouble in the general measurable case.

Fixed Points

The operators T_f and T^* are β-contractions. By a theorem of Banach, they have unique fixed points.

The fixed point of T^* is the optimal reward function V^*. The equality

$$V^*(s) = (T^* V^*)(s)$$

is just the Bellman equation

$$V^*(s) = \sup_a [r(s,a) + \beta \int V^*(t)\, q(dt|s,a)].$$

Stationary Plans

A plan π is **stationary** if there is a function $f : S \mapsto A$ such that $\pi(s_1, a_1, \ldots, a_{n-1}, s_n) = f(s_n)$ for all $(s_1, a_1, \ldots, a_{n-1}, s_n)$.

Notation: $\pi = f^\infty$.

The fixed point of T_f is the reward function $V(\pi)(\cdot)$ for the stationary plan $\pi = f^\infty$.

$$V(\pi)(s) = r(s, f(s)) + \beta \int V(\pi)(t)\, q(dt|s, f(s)) = (T_f V(\pi))(s)$$

Fundamental Question: Do optimal or nearly optimal stationary plans exist?

Existence of Good Stationary Plans

Fix $\epsilon > 0$. For each s, choose $f(s)$ such that

$$(T_f V^*)(s) \geq V^*(s) - \epsilon(1-\beta).$$

Let $\pi = f^\infty$. An easy induction shows that

$$(T_f^n V^*)(s) \geq V^*(s) - \epsilon, \text{ for all } s \text{ and } n.$$

But, by Banach's Theorem,

$$(T_f^n V^*)(s) \to V(\pi)(s).$$

So the stationary plan π is ϵ - optimal.

The Measurable Case: Trouble for T^*

T^* does not preserve Borel measurability.

T^* does not preserve universal measurability.

T^* does preserve “upper semianalytic” functions, but these do not form a Banach space.

Good stationary plans do exist, but the proof is more complicated.

Finitely Additive Extensions of Measurable Problems

Every probability measure on an algebra of subsets of a set F can be extended to a gamble on F, that is, a finitely additive probability defined on all subsets of F. (The extension is typically **not unique**.)

Thus a measurable, discounted problem S, A, r, q, β can be extended to a finitely additive problem $S, A, r, \widehat{q}, \beta$ where $\widehat{q}(\cdot|s, a)$ is a gamble on S that extends $q(\cdot|s, a)$ for every s, a.

Questions: Is the optimal reward the same for both problems? Can a player do better by using non-measurable plans?

Reward Functions for Measurable and for Finitely Additive Plans

For a measurable plan π, the reward

$$V_M(\pi)(s) = E_{\pi,s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)]$$

is the expectation under the countably additive probability $P_{\pi,s}$.

Each measurable π can be extended to a finitely additive plan $\widehat{\pi}$ with reward

$$V(\widehat{\pi})(s) = E_{\widehat{\pi},s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)]$$

calculated under the finitely additive probability $P_{\widehat{\pi},s}$.

Fact: $V_M(\pi)(s) = V(\widehat{\pi})(s)$.

Optimal Rewards

For a measurable problem, let

$$V_M^*(s) = \sup V_M(\pi)(s),$$

where the sup is over all measurable plans π, and let

$$V^*(s) = \sup V(\pi)(s),$$

where the sup is over all plans π in some finitely additive extension.

Theorem: $V_M^*(s) = V^*(s)$.

Proof: The Bellman equation is known to hold in the measurable theory:

$$V_M^*(s) = \sup_a [r(s,a) + \beta \int V_M^*(t)\, q(dt|s,a)].$$

In other terms

$$V_M^*(s) = (T^* V_M^*)(s).$$

But V^* is the unique fixed point of T^*.

Positive Dynamic Programming

Assume the daily reward function r is nonnegative and that the discount factor $\beta = 1$. Let

$$V(\pi)(s) = E_{\pi,s}[\sum_{n=1}^{\infty} r(s_n, a_n)].$$

In a measurable setting

$$V(\pi)(s) = \lim_{\beta \to 1} E_{\pi,s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)]$$

by the monotone convergence theorem. Blackwell (1967) used this equality to prove, for example,

Theorem. In a measurable positive dynamic programming problem, there always exists, for each $\epsilon > 0$ and $s \in S$ such that $V^*(s) < \infty$, an ϵ-optimal stationary plan at s.

Finitely Additive Positive Dynamic Programming

The monotone convergence theorem fails for finitely additive measures. An example with S equal to the set of ordinals less than or equal to the first uncountable ordinal (Dubins and Sudderth, 1975) shows that good stationary plans need not exist.

There is also a countably additive counterexample with a much larger state space (Ornstein, 1969).

References: Countably Additive Dynamic Programming

D. Blackwell (1965).Discounted dynamic programming. *Ann. Math. Statist.* 36 226-235.

D. Blackwell, D. Freedman and M. Orkin (1974). The optimal reward operator in dynamic programming. *Ann. Prob.* 2 926-941.

D. Bertsekas and S. Shreve (1978). *Stochastic Optimal Control: The Discrete Time Case*. Academic Press.

E. Feinberg (1996). On measurability and representation of strategic measures in Markov decision theory. *Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell*, editors T. S. Ferguson, L.S. Shapley, J. B. MacQueen. IMS Lecture Notes-Monograph Series 30 29-44.

D. Ornstein (1969). On the existence of stationary optimal strategies. *Proc. Amer. Math Soc.* 20 563-569.

References: Gambling and Finite Additivity

L. Dubins (1974). On Lebesgue-like extensions of finitely additive measures. *Ann. Prob.* 2 226-241.

L. E. Dubins and L. J. Savage (1965). *How to Gamble If You Must: Inequalities for Stochastic Processes*. McGraw-Hill.

L. E. Dubins and W. Sudderth (1975). An example in which stationary strategies are not adequate. *Ann. Prob.* 3 722-725.

R. Purves and W. Sudderth (1976). Some finitely additive probability theory. *Ann. Prob.* 4 259-276.