Shrinking-Horizon Dynamic Programming

Joëlle Skaf * Stephen Boyd † Assaf Zeevi ‡

Abstract

We describe a heuristic control policy, for a general finite-horizon stochastic control problem, that can be used when the current process disturbance is not conditionally independent of previous disturbances, given the current state. At each time step, we approximate the distribution of future disturbances (conditioned on what has been observed) by a product distribution with the same marginals. We then carry out dynamic programming, using this modified future disturbance distribution, to find an optimal policy, and in particular, the optimal current action. We then execute only the optimal current action. At the next step we update the conditional distribution, and repeat the process, this time with a horizon reduced by one step. (This explains the name 'shrinking horizon dynamic programming'.) We explain how the method can be thought of as an extension of model predictive control, and illustrate our method on two variations on a revenue management problem.

1 Introduction

We consider a general finite-horizon stochastic control problem, with full state information, but without the standard assumption that the current disturbance is conditionally independent of past disturbances, given the current state (see, *e.g.*, [1, 6, 8, 13, 19, 24]). When this assumption holds, standard dynamic programming (DP) can be used to find the optimal policy (see, *e.g.*, [6, 11, 20, 21]). While the curse of dimensionality renders DP impractical in many problems, there are still many other problems for which DP can be carried out effectively. These include, for example, the case in which the state and input spaces are finite, with modest cardinalities, and the case when they are continuous, with small dimension (one or two).

When the disturbances do not satisfy the conditional independence assumption, however, straightforward dynamic programming cannot be used. A general approach is to augment

*This material is based upon work supported by AFOSR award FA9550-06-1-0514 and by NSF award 0529426.

†Joëlle Skaf and Stephen Boyd are with the Information Systems Lab, Electrical Engineering Department, Stanford University, Stanford, CA 94305-9510 USA. Email: {jskaf,boyd}@stanford.edu.

‡Assaf Zeevi is with the Graduate School of Business, Columbia University, Uris Hall, Room 406, New York, NY 10027 USA. Email: assaf@gsb.columbia.edu.

the state to include all previous disturbances; with this augmented state, the conditional independence assumptions holds, so standard DP can be applied. Unless the time horizon is very small, however, this is not practical, since the augmented state is large (either in cardinality, if it is discrete, or in dimension, if it is continuous).

Approximate dynamic programming (ADP) methods, which are meant to handle unwieldy large state spaces, can be applied to find a suboptimal policy, using the augmented system [5, 6, 18]. These methods are based on using an estimate of the optimal value function, or optimal policy. Another general method for finding a suboptimal control policy is model predictive control (MPC) [4, 14, 17, 24], which goes by many other names, including certainty-equivalent model predictive control (CE-MPC), dynamic matrix control [10], rolling horizon planning [9], and dynamic linear programming (DLP) [22]. In MPC, the action or control is found as follows. At each step, we solve a deterministic optimal control problem, with the unknown future disturbances replaced with some kind of estimates available at the current time (such as conditional means). We can think this optimization as a planning exercise, working out the best sequence of actions to take, if the future disturbances were equal to our estimates. We then execute only the current action in this plan. At the next step, the same problem is solved, this time using the exact value of the current state, which is now known from the measurement, and an updated set of predictions.

In this note we introduce another suboptimal policy that can be used when straightforward DP would be practical, if the disturbances satisfied the conditional independence assumption. This includes, for example, systems with finite state, action, and noise spaces, with the product of their cardinalities no more than a million or so (say), for which we can directly compute the value function by recursion. Another example is systems with continuous state with low dimension (say, one or two), for which we can (accurately) discretize the state, and carry out the value function recursion numerically. Our method requires the solution of a DP problem, for the given (unaugmented) system, but with independent disturbances, at each step. Since the number of remaining time steps shrinks as time advances, we call the method *shrinking horizon dynamic programming* (SHDP). SHDP can be thought of as a variation on MPC, in which a (tractable) DP problem is solved each step, instead of a deterministic optimal control problem.

In §2 we describe a general finite horizon stochastic control problem, fixing our notation. In §3, we briefly describe DP, and DP with state augmentation. In §4, we describe three suboptimal policies, which grow in sophistication. In certainty-equivalent open-loop control (CE-OLC), we ignore all variation in the disturbances, and simply replace the disturbances with some fixed values, which yields a (deterministic) optimization problem. In CE-MPC, at each step, future disturbances are replaced with predictions, based on currently available information. And finally, in SHDP, at each step, we replace the future disturbance distribution with a product distribution with the same marginals, and then use DP to solve the resulting problem. In §5 we illustrate SHDP with two applications from revenue management.

2 Finite horizon stochastic control

We consider a discrete-time dynamic system, over the time interval $t = 1, \ldots, T$, with dynamics

$$x_{t+1} = f_t(x_t, u_t, w_t), \quad t = 1, \ldots, T-1, \tag{1}$$

where $x_t \in \mathcal{X}$ is the system state, $u_t \in \mathcal{U}$ is the control input or action, $w_t \in \mathcal{W}$ is the process noise or disturbance, all at time step t. The functions $f_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$ are the state transition functions. We assume that the initial state x_1 is known. The process noise trajectory $w_{1:T} \in \mathcal{W}^T$ is random, with a known distribution. Here we use the notation $z_{i:j}$ to denote $z_{i:j} = (z_i, z_{i+1}, \ldots, z_{j-1}, z_j)$.

We will consider causal control policies, in which $x_1, \ldots, x_t$ (*i.e.*, $x_{1:t}$) and $w_1, \ldots, w_{t-1}$ (*i.e.*, $w_{1:t-1}$) are available when the control input u_t must be chosen. Thus we have

$$u_t = \varphi_t(x_{1:t}, w_{1:t-1}), \quad t = 1, \ldots, T, \tag{2}$$

where the family of functions $\varphi_t : \mathcal{X}^t \times \mathcal{W}^{t-1} \to \mathcal{U}$, for $t = 1, \ldots, T$, is called the *control policy*. For fixed control policy, (1) and (2) can be used to express the control trajectory $u_{1:T}$ and the state trajectory $x_{1:T}$ as functions of $w_{1:T}$, so these are also random variables.

We can express the control policy in several other forms. In (2) the control input is expressed as a family of functions of the current and past states $x_{1:t}$ and past disturbances $w_{1:t-1}$. But it can just as well be expressed as a family of functions of the past disturbances $w_{1:t-1}$ alone, since the current and past states $x_{1:t}$ are a function of the past disturbances $w_{1:t-1}$ (given the control policy functions $\varphi_1, \ldots, \varphi_{t-1}$). It will also be convenient in the sequel to express the control policy as a family of functions of the current state and the past disturbances, as in

$$u_t = \phi_t(x_t, w_{1:t-1}), \qquad t = 1, \ldots, T.$$

When φ_t is a function only of x_t, *i.e.*, has the form

$$\varphi_t(x_{1:t}, w_{1:t-1}) = \psi_t(x_t), \tag{3}$$

where $\psi_t : \mathcal{X} \to \mathcal{U}$, we refer to the policy as a *state feedback policy*. When φ_t is a constant, *i.e.*, independent of any states or disturbances, we refer to it as a *constant policy* or *open-loop policy*.

The objective function has the form

$$J = \mathbf{E} \sum_{t=1}^{T} \ell_t(x_t, u_t, w_t), \tag{4}$$

where $\ell_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbf{R} \cup \{\infty\}$ is the stage cost at time step t, for $t = 1, \ldots, T$. (Thus, we are implicitly imposing the constraint on u_t that $\ell_t(x_t, u_t, w_t) < \infty$ almost surely.)

The objective J is a (very complex) function of the control policy. In the *stochastic control problem*, the goal is to choose the control policy so that J is minimized. The data in this problem are the initial state x_1, the state transition functions f_t, the distribution of $w_{1:T}$, and the state cost functions ℓ_t; the optimization variable is the control policy.

2.1 Prescient lower bound

We can obtain a lower bound on the optimal value of the stochastic control problem by relaxing the causality constraint on the policies: We allow *all* u_t to depend on $w_{1:T}$. In this case we can explicitly solve the problem: For each realization of $w_{1:T}$, the optimal control input sequence is found by solving the (deterministic) optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} \ell_t(x_t, u_t, w_t) \\ \text{subject to} & x_{t+1} = f_t(x_t, u_t, w_t), \quad t = 1, \ldots, T-1, \end{array} \tag{5}$$

with variables $x_2, \ldots, x_T$, $u_1, \ldots, u_T$. The optimal value of this optimization is a random variable, since it depends on $w_{1:T}$. The mean value of the optimal value of (5) is evidently a lower bound on the optimal value of the stochastic control problem. We call this bound the *prescient lower bound* since it is the optimal value of the problem when the control actions know the future disturbances exactly.

To evaluate this lower bound, we must be able to effectively solve the deterministic optimization problem (5). We can evaluate the mean by Monte Carlo, by generating many realizations of $w_{1:T}$, solving the problem (5) for each realization, and averaging the optimal values obtained.

There is no reason to believe that this bound should be close to the optimal value of the stochastic control problem. Indeed, the difference between these two numbers can be directly interpreted as the cost of not knowing the future.

3 Dynamic programming

In this section we consider the special case in which the following assumption about the disturbance distribution holds:

$$w_t \text{ is independent of } w_{1:t-1} \text{ given } x_t, \qquad t = 2, \ldots, T. \tag{6}$$

Since $x_1, \ldots, x_{t-1}$ are functions of $w_{1:t-1}$, this assumption implies that w_t is independent of $x_{1:t-1}$, given x_t. If the disturbances $w_1, \ldots, w_T$ are independent, then of course the assumption (6) holds.

It is well known that when (6) holds, the optimal control policy is a state feedback policy (3). Moreoever, an optimal policy can be found by dynamic programming (DP), using the Bellman recursion [2, 3]. We recursively define the cost-to-go or value functions $V_t : \mathcal{X} \to \mathbf{R}$, as

$$V_t(z) = \inf_{v \in \mathcal{U}} \mathbf{E}_{w_t} \left(\ell_t(z, v, w_t) + V_{t+1}(f_t(z, v, w_t)) \right), \tag{7}$$

for $t = T, \ldots, 1$, where we take $V_{T+1} = 0$; the expectation is conditioned on $x_t = z$. An optimal control policy is then given by

$$\psi_t^\star(x_t) = \arg\inf_{v \in \mathcal{U}} \mathbf{E}_{w_t} \left(\ell_t(x_t, v, w_t) + V_{t+1}(f_t(x_t, v, w_t)) \right), \tag{8}$$

for $t = 1, \ldots, T$. For more on dynamic programming, see [5, 6, 11, 13, 20, 21].

3.1 Dynamic programming with state augmentation

When w_t is not independent of $w_{1:t-1}$ given x_t, *i.e.*, the assumption (6) does not hold, the straightforward DP method described above cannot be used to find an optimal policy. One general approach in this case is to augment the state with all previous disturbances: We take states of the augmented system to be

$$\tilde{x}_1 = x_1, \qquad \tilde{x}_t = (x_t, w_{1:t-1}), \qquad t = 2, \ldots, T.$$

(These vary in size.) The state transition function can be extended to the augmented system in the obvious way. For the augmented system, the current disturbance w_t is now conditionally independent of $w_{1:t-1}$ given $\tilde{x}_t$ (since in this case w_{t-1} is deterministic). Standard DP can now be applied to this augmented (and much larger) system. For more on state augmentation, see [6, §1.4].

In some cases we can get away with a smaller state augmentation. If w_t can be expressed as a function of a Markov process with state $s_t \in \mathcal{S}$, we can augment the state as $\tilde{x}_t = (x_t, s_t)$, assuming we can have access to s_t in determining the control action at time step t.

Our focus in this paper is on systems for which the basic DP algorithm of §3 would be practical, if the disturbances satisfied the conditional independence assumption (6), but DP with state augmentation, as described here, is not.

4 Suboptimal policies

In this section we describe several methods for finding a good, if not optimal, policy, when the conditional independence assumption (6) does not hold. Each of the methods can be interpreted as finding an optimal policy for a modified problem, after a simplification of the disturbance model.

4.1 Certainty-equivalent open-loop control

A very simple constant or open-loop policy can be obtained as follows.

1. *Form constant approximations $\hat{w}_1, \ldots, \hat{w}_T$ of the disturbances.*

   These can be means, most likely values, or any other reasonable approximations of w_t.

2. *Solve the resulting problem using these approximate disturbance values.*

   In this case the stochastic control problem reduces to an ordinary optimization problem,

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} \ell_t(\tilde{x}_t, \tilde{u}_t, \hat{w}_t) \\ \text{subject to} & \tilde{x}_{t+1} = f_t(\tilde{x}_t, \tilde{u}_t, \hat{w}_t), \quad t = 1, \ldots, T-1 \\ & \tilde{x}_1 = x_1. \end{array} \tag{9}$$

   with variables $\tilde{x}_1, \ldots, \tilde{x}_T \in \mathcal{X}$, $\tilde{u}_1, \ldots, \tilde{u}_T \in \mathcal{U}$. Let $u_1^\star, \ldots, u_T^\star$ denote optimal values of $\tilde{u}_1, \ldots, \tilde{u}_T$.

3. *Use $u_t = u_t^\star$, $t = 1, \ldots, T$, as an (open-loop) policy.*

We refer to this as *certainty-equivalent open-loop control* (CE-OLC). In CE-OLC, we ignore all variation in the disturbances, since we assume that the disturbances are known, and take on the predicted values $\hat{w}_t$. CE-OLC requires the solution of the one optimization problem (9), which can be done ahead of time.

4.2 Certainty-equivalent model predictive control

In CE-MPC we calculate each input u_t by replacing the current and future disturbances with constant approximations $\hat{w}_{t|t}, \ldots, \hat{w}_{T|t}$, obtained using the most recent known data, and solving the resulting optimization problem over the remaining time period.

To find u_t, we proceed as follows.

1. *Form constant approximations $\hat{w}_{t|t}, \ldots, \hat{w}_{T|t}$ of the current and future disturbances.*

   Here $\hat{w}_{\tau|t}$ denotes our prediction of w_τ based on the information available at time period t, *i.e.*, $w_{1:t-1}$. These predictions could be, for example, the conditional means $\hat{w}_{\tau|t} = \mathbf{E}(w_\tau | w_{1:t-1})$, or the conditionally most likely values. These approximations are (in general) functions of $w_{1:t-1}$.

2. *Solve the resulting problem over the remaining period using these approximate disturbance values.*

   We solve the (deterministic) optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{\tau=t}^{T} \ell_\tau(\tilde{x}_\tau, \tilde{u}_\tau, \hat{w}_{\tau|t}) \\ \text{subject to} & \tilde{x}_{\tau+1} = f_\tau(\tilde{x}_\tau, \tilde{u}_\tau, \hat{w}_{\tau|t}), \quad \tau = t, \ldots, T-1 \\ & \tilde{x}_t = x_t, \end{array} \tag{10}$$

   with variables $\tilde{x}_t, \ldots, \tilde{x}_T \in \mathcal{X}$, $\tilde{u}_t, \ldots, \tilde{u}_T \in \mathcal{U}$. Let $u_t^\star, \ldots, u_T^\star$ denote optimal values of $\tilde{u}_t, \ldots, \tilde{u}_T$. These are (in general) functions of $w_{1:t-1}$ (through the predicted values $\hat{w}_{\tau|t}$) and x_t (through the equality constraint in (10).

3. *Use $u_t = u_t^\star$ as the current input.*

CE-MPC can be thought of as CE-OLC, where at each step we use the most up to date predictions of future disturbance values. The CE-MPC policy has recourse, *i.e.*, u_t is a function of the current state and past disturbances. Unlike CE-OLC, CE-MPC takes advantage of the measured values of past disturbances in its determination of a current action, through the generation of the predicted future disturbances. Its model of the future disturbances, however, is still rather unsophisticated, since the implicit assumption is that the future disturbances are known exactly. CE-MPC requires the solution of the optimization problem (10) at each time step.

4.3 Shrinking-horizon dynamic programming

We now come to the algorithm we propose. SHDP takes CE-MPC one step further, by taking into account variation in future disturbances. However, any dependency amoung the future disturbances is ignored, which makes it possible to solve the (modified) problem using DP.

In SHDP, the control input u_t is found as follows.

1. *Form approximate product measure for current and future disturbances.*

   Let $\mathcal{D}_t$ denote the distribution of $w_{t:T}$, conditioned on the observed $x_{1:t}$ and $w_{1:t-1}$. Let $\tilde{\mathcal{D}}_t$ denote the distribution on $w_{t:T}$ obtained from $\mathcal{D}_t$ by keeping the marginal distributions of $w_t, \ldots, w_T$, but otherwise making them independent. In other words: find the marginal distributions of $w_t, \ldots, w_T$ under $\mathcal{D}_t$, and then form $\tilde{\mathcal{D}}_t$ as the product of these measures.

2. *Solve the resulting stochastic control problem over the remaining period using this approximate measure on current and future disturbannces.*

   Use DP to an find optimal policy on the remaining time interval, for the modified future disturbance distribution. Define the Bellman functions $V_T, \ldots, V_{t+1}$ recursively as

   $$V_\tau(z) = \inf_{v \in \mathcal{U}} \mathbf{E}_{w_\tau} \left(\ell_\tau(z, v, w_\tau) + V_{\tau+1}(f_\tau(z, v, w_\tau)) \right), \tag{11}$$

   for $\tau = T, \ldots, t+1$, with $V_{T+1} = 0$, using the marginal conditional distributions for $w_t, \ldots, w_T$ (*i.e.*, $\tilde{\mathcal{D}}_t$). Let

   $$u_t^\star = \arg \inf_{v \in \mathcal{U}} \mathbf{E}_{w_t} \left(\ell_t(x_t, v, w_t) + V_{t+1}(f_t(x_t, v, w_t)) \right) \tag{12}$$

   denote an optimal input for the modified stochastic control problem. This (in general) depends on $x_{1:t}$ and $w_{1:t-1}$, via the conditional distribution $\mathcal{D}_t$.

3. *Use $u_t = u_t^\star$ as the current input.*

In CE-MPC, all uncertainty in the future disturbances is ignored; in SHDP, however, we retain information about uncertainty in future disturbances, but we ignore any dependence between the future disturbances. At each time step, SHDP requires the solution of a stochastic control problem (that satisfies the conditional independence assumption (6)), using (11) and (12).

When the disturbance in the original problem satisfies the conditional independence assumption (6)), the true conditional distribution $\mathcal{D}_t$ and the approximate product conditional distribution $\tilde{\mathcal{D}}_t$ are the same, and SHPD is an optimal control policy.

4.4 Implementing SHDP

We mention two methods that can be used to implement SHDP. In some cases $w_{1:T}$ has a parametrized distribution, for which the conditional distributions are easily formed. For

example, if the distribution is Gaussian or log-normal, so is the conditional distribution $\mathcal{D}_t$, whose parameters are easily computed. In these two cases, the marginal distributions of w_τ, conditioned on t, are also Gaussian or log-normal, with easily computed parameters.

In other cases we can use a sampling approximation for $\tilde{D}_t$. To do this we only need a method for generating a set of samples or realizations $w_{t:T}^{(i)}$, $i = 1, \ldots, N$, from the conditional distribution $\mathcal{D}_t$. We interpret $w_t^{(i)}, \ldots, w_T^{(i)}$ as a set of plausible future disturbance trajectories, given everything observed to date.

5 Revenue management

We now describe a general nonperishable revenue management (RM) problem and show how SHDP can be implemented for two variations on the problem. In RM, the goal is to maximize expected profit from sales of an asset that occur over T periods, denoted $t = 1, \ldots, T$. Let $x_t \geq 0$ denote the total amount of the asset remaining at period t, with $x_1 = B$ being the given initial quantity available. At each time period t, we must decide how much of the asset remaining to release for sale, which we denote u_t, with $0 \leq u_t \leq x_t$. We let $d_t \geq 0$ denote the demand for the asset in period t, with $d_{1:T}$ random from a known distribution. (This distribution might be obtained by a modeling step, from historical demand data.) When u_t is chosen, the previous demands $d_{1:t-1}$ are known; the current period and future demands $d_{t:T}$ are not. The amount of asset sold in period t is $s_t = \min\{u_t, d_t\}$, the minimum of the amount of asset released for sale and the demand. The asset is nonperishable, *i.e.*, any amount made available in a time period but not sold carries over to the next time period, so we have $x_{t+1} = x_t - s_t$. The price of the asset in time period t is p_t, which we assume is known. We will assume that the prices are positive and increasing, *i.e.*, $0 < p_1 < \cdots < p_T$. The total revenue is $R = \sum_{t=1}^T p_t s_t$. The RM problem is to find a policy

$$u_t = \phi_t(x_t, d_1, \ldots, d_{t-1}), \quad t = 1, \ldots, T,$$

that maximizes $\mathbf{E}R$. We will look at two versions of the RM problem: in the discrete (or indivisible) version, the asset quantities x_t, u_t, d_t are integers; in the continuous (or divisible) version, the asset quantities x_t, u_t, d_t are real numbers.

As we will see, the RM problem is readily solved when the demands satisfy the conditional independence assumption (6). But the dependence of the demands over time is a key point in RM: The demand in the first period, for example, tells us something about the demand in the later periods, when the price is higher, and therefore can strongly affect our early actions.

There are many variations on the RM problem described above. We can allow $x_t < 0$, which we interpret as backlog (which incurs some backlog cost). In the perishable RM problem, the amount of asset made available but not sold is lost, so $x_{t+1} = x_t - u_t$. In another version, the prices are random, with some known distribution. In yet another version, we do not know the demands $d_1, \ldots, d_{t-1}$; instead we only know the sales $s_1, \ldots, s_{t-1}$. If $s_\tau < u_\tau$, we know the demand exactly; if, however, we have $s_t = u_t$, we only know that $d_\tau \geq u_\tau$. For more on revenue management and for some of the latest work in the field, see [7, 12, 15, 16, 23].

Let us cast our discrete and continuous RM problems in the formulation presented in §2. The state is simply x_t, the remaining amount of asset, with $\mathcal{X} = \{0, 1, \ldots, B\}$ (in the continuous case, $\mathcal{X} = [0, B]$); the control input is u_t, the amount released, with $\mathcal{U} = \{0, 1, \ldots, B\}$ (in the continuous case, $\mathcal{U} = [0, B]$). The process disturbance w_t is just the demand d_t, with $\mathcal{W} = \mathbf{Z}_+$ ($\mathcal{W} = \mathbf{R}_+$ in the continuous case). The dynamics (1) are given by

$$x_{t+1} = f_t(x_t, u_t, d_t) = x_t - \min(u_t, d_t), \qquad t = 1, \ldots, T-1,$$

with $x_1 = B$. The cost function in period t is the negative revenue,

$$l_t(x_t, u_t, d_t) = \begin{cases} -p_t \min(u_t, d_t) & \text{if } u_t \leq x_t \\ +\infty & \text{otherwise.} \end{cases}$$

(The value $+\infty$ here encodes the constraint that $u_t \leq x_t$.) With these identifications, the RM problem is exactly the general stochastic control problem from §2.

Conditionally independent demands. If the demands $d_1, \ldots, d_T$ satisfy (6), the RM problem is easily solved by DP. With discrete variables, the value function V_t reduces to a vector in $\mathbf{R}^B$. The Bellman recursion can be carried out by direct evaluation of the expectiation (truncating the distribution of d_t at some reasonable large value); the minimization can be carried out by exhaustive search over $v = 0, 1, \ldots, x_t$. With continuous variables, V_t is a function defined on the real interval $[0, B]$; we describe it by its values at, say, $M = 100$ values in the interval, and use piecewise-linear interpolation to evaluate it between these sample points. The Bellman recursion can be evaluated as in the discrete case.

Known demands. To compute the prescient bound, and to carry out CE-MPC, we must be able to solve optimization problems of the form (5) or (10), *i.e.*, solve RM problems when the future demand is known. We now show how this can be done analytically, for both the discrete and continuous cases. We argue informally here, but optimality of the control input we describe is easily proved. The strategy is to first satisfy (if possible) all the demand in period T (which corresponds to highest price p_T); then we satisfy (again, if possible) all demand in period $T-1$ (which has second highest price), with any left over asset. We continue working backward this way until we run out of asset to allocate, or all demand is satisfied.

To describe this formally, let $\bar{t}$ be the largest integer for which $\sum_{t=\bar{t}}^{T} d_t > B$. Then we have

$$u_t^\star = \begin{cases} d_t & t > \bar{t} \\ B - \sum_{t=\bar{t}+1}^{T} d_t & t - \bar{t} \\ 0 & t < \bar{t}. \end{cases}$$

The associated optimal revenue is

$$p_{\bar{t}} \left(B - \sum_{t=\bar{t}+1}^{T} d_t \right) + \sum_{t=\bar{t}}^{T} p_t d_t.$$

5.1 Discrete RM example

Here we assume that all asset quantities are discrete, *i.e.*, integers. We model the demand d with an auto-regressive Poisson process, *i.e.*,

$$d_t \sim \text{Poisson}(\alpha d_{t-1} + \beta d_{t-2} + \gamma), \quad t = 1, \ldots, T,$$

where α, β, γ are scalars, and d_0, d_{-1} are given integers. The mean demand follows the linear auto-regressive process

$$\bar{d}_t = \alpha \bar{d}_{t-1} + \beta \bar{d}_{t-2} + \gamma, \quad t = 1, \ldots, T.$$

The conditional distribution $\mathcal{D}_t$ of $d_t, \ldots, d_T$ given $d_1, \ldots, d_t - 1$, is also auto-regressive Poisson:

$$d_\tau \sim \text{Poisson}(\alpha d_{\tau-1} + \beta d_{\tau-2} + \gamma), \quad \tau = t, \ldots, T.$$

We can easily generate samples from this conditional distribution by simulation, which can be used to evaluate expectations over the marginals.

Note that, given the demand process above, the optimal policy can be computed exactly by DP with state augmentation as described in §3.1. The state at time t is augmented to be (x_t, d_{t-1}, d_{t-2}). Under this formulation, the conditional independence (6) holds and DP can be applied.

Numerical instance. We consider a particular problem instance with $T = 10$ periods, a total initial asset level $B = 200$ units, and linearly increasing prices $p_t = (4t + 5)/9$, $t = 1, \ldots, 10$, that vary from $p_1 = 1$ to $p_{10} = 5$. We choose $\alpha = 0.3$, $\beta = 0.2$, $\gamma = 8$, $d_0 = d_{-1} = 70$. Figure 1 shows the mean and standard deviation of the demands versus time. It also shows the evolution of the price versus time. The mean total demand is 235 which is slightly more than the initial asset level.

We generate 1000 realizations of d, and for each one, we work out the revenue obtained using CE-OLC, CE-MPC, SHDP, and the prescient control policy, which gives us an upper bound on the revenue. For the predictions of future demand, we use the mean (conditional mean in CE-MPC), rounded to the closest integer. We also computed the optimal policy by brute force, using DP with state augmentation. (This required approximately 10 CPU hours on a quadcore 3GHz machine, as compared to 2.75 CPU seconds for each run of SHDP.) The results are shown below. We can see that SHDP substantially outperforms CE-OLC and CE-MPC, achieving less than half the suboptimality of these policies.

Policy	Revenue mean $\pm$ std. dev.	Suboptimality
CE-OLC	517.43 ± 40.61	7.7%
CE-MPC	511.95 ± 49.26	8.7%
SHDP	539.96 ± 53.89	3.6%
Optimal	560.39	0%
Prescient	(568.72 ± 53.05)	

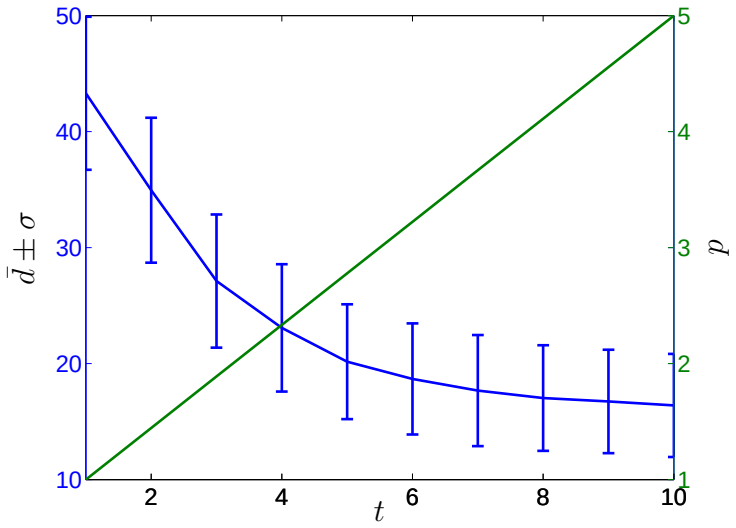

Figure 1: Mean and standard deviation of demand d versus t (blue), and price p versus t (green).

5.2 Continuous RM example

Here we assume that all asset quantities are continuous. We will model the demand d with a log-normal distribution, *i.e.*, we assume that $\log d \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbf{R}^T$ and $\Sigma \in \mathbf{S}^T$ are known. The mean demand is given by

$$\bar{d}_t = \exp(\mu_t + \Sigma_{tt}/2), \quad t = 1, \ldots, T.$$

The demand covariance matrix is

$$\Sigma_d = \mathbf{E}(d - \bar{d})(d - \bar{d})^T = (\bar{d}\bar{d}^T) \circ (\exp(\Sigma) - 1),$$

where $\exp(\cdot)$ is entrywise, and $\circ$ is the Hadamard product, *i.e.*, entrywise multiplication.

The conditional distribution $\mathcal{D}_t$ of $d_{t:T}$ given $d_{1:t-1}$ is also log-normal:

$$\log(d_{t:T})|d_{1:t-1} \sim \mathcal{N}(\nu_t, \Lambda_t), \tag{13}$$

where $\log(\cdot)$ is entrywise, and

$$\begin{aligned} \nu_t &= \mu_{t:T} + \Sigma_{t:T,1:t-1}\Sigma^{-1}_{1:t-1,1:t-1}(\log((d_1, \ldots, d_{t-1})) - \mu_{1:t-1}), \\ \Lambda_t &= \Sigma_{t:T,t:T} - \Sigma_{t:T,1:t-1}\Sigma^{-1}_{1:t-1,1:t-1}\Sigma^T_{t:T,1:t-1}. \end{aligned}$$

(The subscripts denote subvectors or submatrices of μ and Σ, with the given index ranges.) The marginal distribution of $d_{t:T}$ under $\mathcal{D}_t$ is log-normal too: the product measure $\tilde{\mathcal{D}}_t$ is log-normal with parameters ν_t and $\mathbf{diag}(\Lambda_t)$. In particular, we can easily determine the marginal distributions of $d_t, \ldots, d_T$, conditioned on $d_{1:t-1}$.

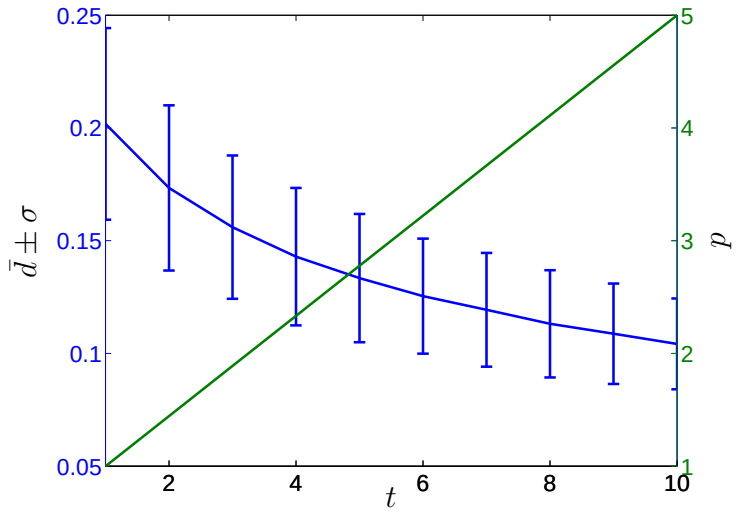

Figure 2: Mean and standard deviation of demand d versus t (blue), and price p versus t (green).

Numerical instance. We consider a particular problem instance with $T = 10$ periods, a total initial asset level $B = 1$, and linearly increasing prices $p_t = (4t + 5)/9$, $t = 1, \ldots, 10$, that vary from $p_1 = 1$ to $p_{10} = 5$.

The parameters μ and Σ, which determine the demand distribution, come from a constant-elasticity model, and a model for inter-period demand dependence. We choose the mean demands to follow a constant-elasticity price-response function, *i.e.*,

$$\bar{d}_t = cp_t^{-\epsilon}, \qquad t = 1, \ldots, T$$

where $c = 0.2$ and $\epsilon = 0.4$. We describe Σ via its diagonal elements and correlations, $\Sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$, where σ_t is the standard deviation of $\log d_t$, and ρ_{ij} is the correlation of $\log d_i$ and $\log d_j$. We take $\sigma_t = 0.2$ for all t, which means each demand often varies $\pm 20\%$, and sometimes $\pm 40\%$. We use a simple model of decaying correlation: for $i \neq j$,

$$\rho_{ij} = \alpha \exp(-\beta |i - j|),$$

with $\alpha = 0.7$, $\beta = 0.1$. Thus the correlation between $\log d_t$ and $\log d_{t+1}$ is 63%; the smallest correlation is between $\log d_1$ and $\log d_{10}$, around 28.5%. Figure 2 shows the mean and standard deviation of the demands versus time. It also shows the evolution of the price versus time. The mean total demand is 1.38 which is slightly more than the initial asset level.

To evaluate each $V_\tau(z)$ we discretize x and u over an evenly-spaced grid of 100 points over the interval $[0, B]$. We approximate the expectations in (11) and (12) by replacing them by their empirical mean over 100 samples of d_t generated from the marginal of d_t under $\mathcal{D}_t$.

We generate 1000 realizations of d, and for each one, we work out the revenue obtained using CE-OLC, CE-MPC, SHDP, and the prescient control policy. The results are shown below.

Policy	Revenue mean ± std. dev.
CE-OLC	3.05 ± 0.26
CE-MPC	3.02 ± 0.29
SHDP	3.11 ± 0.25
Prescient	(3.28 ± 0.27)

We can see that SHDP substantially outperforms CE-OLC and CE-MPC. The mean revenue obtained with SHDP is only 5% under that achieved with full knowledge of future demand. In particular, the SHDP control policy is at most 5% suboptimal. (It is likely to be substantially less suboptimal.)

Acknowledgments

We thank Trevor Hastie for suggesting the AR-Poisson model for discrete demand, and we thank Michael Harrison for introducing the RM problems to us.

References

[1] K. Åström. *Introduction to Stochastic Control Theory*. Dover Publications, 2006.

[2] R. Bellman. *Dynamic Programming*. Courier Dover Publications, 1957.

[3] R. Bellman and S. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, 1962.

[4] A. Bemporad. Model predictive control design: New trends and tools. In *Proceedings of 45th IEEE Conference on Decision and Control*, pages 6678–6683, 2006.

[5] D. Bertsekas. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[6] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2005.

[7] D. Bertsimas and I. Popescu. Revenue management in a dynamic network environment. *Transportation Science*, 37(3):257–277, 2003.

[8] J. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 1997.

[9] E. Cho, K. Thoney, T. Hodgson, and R. King. Supply chain planning: Rolling horizon scheduling of multi-factory supply chains. In *Proceedings of the 35th conference on Winter simulation: Driving innovation*, pages 1409–1416, 2003.

[10] C. Cutler. *Dynamic Matrix Control: An Optimal Multivariable Control Algorithm with Constraints*. PhD thesis, University of Houston, 1983.

[11] E. Denardo. *Dynamic Programming: Models and Applications*. Prentice-Hall, 1982.

[12] V. Farias. *Revenue Management Beyond "Estimate, then Optimize"*. PhD thesis, Stanford University, 2007.

[13] P. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification and adaptive control*. Prentice-Hall, 1986.

[14] W. Kwon and S. Han. *Receding Horizon Control*. Springer-Verlag, 2005.

[15] R. Levi, M. Pál, R. Roundy, and D. Shmoys. Approximation algorithms for stochastic inventory control models. *Mathematics of Operations Research*, 32(2):284–302, 2007.

[16] J. McGill and G. Van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 33:233–256, 1999.

[17] S. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2006.

[18] W. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, Inc., 2007.

[19] A. Prekopa. *Stochastic Programming*. Kluwer Academic Publishers, 1995.

[20] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[21] S. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Academic Press, Inc. Orlando, FL, USA, 1983.

[22] K. Talluri and G. Van Ryzin. *The Theory and Practice of Revenue Management*. Springer, 2005.

[23] A. Thiele. *A Robust Optimization Approach to Supply Chains and Revenue Management*. PhD thesis, Massachusetts Institute of Technology, 2004.

[24] P. Whittle. *Optimization Over Time: Dynamic Programming and Stochastic Control*. Wiley, 1982.