Abstract: In this paper we consider discrete time stochastic optimal control problems over infinite and finite time horizons. We show that for a large class of such problems the Taylor polynomials of the solutions to the associated Dynamic Programming Equations can be computed degree by degree.

Keywords: Discrete Time Stochastic Optimal Control, Dynamic Programming, Linear Quadratic Regulator

1. INTRODUCTION

In this paper we consider discrete time stochastic optimal control problems over infinite and finite time horizons. We show that for a large class of such problems the Taylor polynomials of the solutions to the associated Dynamic Programming Equations can be computed degree by degree generalizing the method of Al’brekht (1). There is a vast literature dealing with such problems, we refer the reader to Bertsekas (2). Stochastic optimal control problems in continuous time are discussed in (3) and (6). Navasca generalized Al’brekht’s method to deterministic, discrete time, infinite horizon optimal control problems (5).

We begin with a relatively simple stochastic, infinite horizon, optimal control problem and then move on to more complicated problems over infinite and finite horizons. Consider a discrete time, infinite horizon, stochastic Linear Quadratic Regulator with Bilinear Noise (DLQGB),

\[
\min_{u(\cdot)} \frac{1}{2} E \left\{ \sum_{t=0}^{\infty} (x^t Q x^t + 2 x^t S u^t + u^t R u^t) \right\}
\]

subject to \( x^{t+1} = f(x^t, u^t) + \sum_{k=1}^{r} w_k \gamma_k(x^t, u^t) \)

where \( x(0) = x^0 \) and \( x^t(t) = x(t+1) \).

The state \( x \) is \( n \) dimensional, the control \( u \) is \( m \) dimensional and \( w(t) = (w_1(t), \ldots, w_r(t))^T \) is \( r \) dimensional sequence of independent Gaussian random vectors of mean zero and covariance \( I_r \). The matrices are sized accordingly, in particular \( C_k \) is an \( n \times n \) matrix and \( D_k \) is an \( n \times m \) matrix for each \( k = 1, \ldots, r \).

To the best of our knowledge discrete time infinite horizon problems with bilinear noise have not been considered before. In (2) we studied the continuous time version of this problem. The finite horizon version of this problem with noise entering linearly is well studied in both discrete (2) and continuous time (3), (6).

We restrict our attention to problems with bilinear noise so that we can use power series techniques to solve the dynamic programming equations of nonlinear optimal control problems. The class of infinite horizon nonlinear optimal control problems that are of interest are of the form

\[
\min_{u(\cdot)} E \left\{ \sum_{t=0}^{\infty} l(x^t, u^t) \right\}
\]

subject to \( x^t+1 = f(x^t, u^t) + \sum_{k=1}^{r} w_k \gamma_k(x^t, u^t) \)

where \( x(0) = x^0, f(x, u) \) and \( \gamma_k(x, u) \) are smooth functions of order \( O(x,u) \) and \( l(x, u) \) is a smooth function of order \( O(x,u)^2 \).

Associated to these problems are Bellman’s Dynamic Programming equations for the optimal cost and optimal feedback. Assuming they exist, let \( \pi(x) \) be the optimal cost starting at \( x \) and \( u = \kappa(x) \) be the optimal feedback at \( x \) for this problem. Then they satisfy the Stochastic Infinite Horizon Dynamic Programming Equations (SIDPE),

\[
\pi(x) = \min_{u(\cdot)} E \left\{ \pi(f(x, u)) + \sum_{k=1}^{r} w_k \gamma_k(x, u) + l(x, u) \right\}
\]

(1)

\[
\kappa(x) = \arg \min_{u(\cdot)} E \left\{ \pi(f(x, u)) + \sum_{k=1}^{r} w_k \gamma_k(x, u) + l(x, u) \right\}
\]

(2)

These equations differ from their deterministic counterparts because of the presence of the noise terms.

The class of finite horizon nonlinear optimal control problems that are of interest are of the form

\[
\min_{u(\cdot)} E \left\{ \sum_{t=0}^{T} l(t, x^t + \pi_T(x(T))) \right\}
\]

subject to \( x^{t+1} = f(t, x^t) + \sum_{k=1}^{r} w_k \gamma_k(t, x^t, u^t) \)

\* Research supported by AFOSR
where \( f(t, x, u) \) and \( \gamma_k(t, x, u) \) are smooth \( n \) vector valued functions with respect to \( x, u \) of order \( O(x, u) \), and \( \pi_T(x) \) is a smooth scalar valued function with respect to \( x \) of order \( O(x) \). The solutions \( P(\tau) \) of the DARE is the kernel of the optimal cost of deterministic LQRs and since

\[
x(t_0) = x^0
\]

the \( \pi(x) \) to be a quadratic function of \( x \) and \( \kappa(x) \) to be a linear function of \( x \).

\[
\pi(x) = \frac{1}{2} x' P x, \quad \kappa(x) = K x
\]

We plug these expressions into SIDPE and they simplify to

\[
P = Q + K^T R K + (F + G K) P (F + G K) \quad (5)
\]

\[
K = - \left( R + G' P G + \sum_{k=1}^r D_k^T P D_k \right)^{-1}
\]

\[
\times \left( G' P F + S + \sum_{k=1}^r D_k^T PC_k \right) \quad (6)
\]

We call these equations (5, 6) the Stochastic Discrete Time Algebraic Riccati Equations (SDARE). They reduce to the deterministic Discrete Time Algebraic Riccati Equations (DARE) if \( C_k = 0 \) and \( D_k = 0 \) for \( k = 1, \ldots, r \).

Here is an iterative method for solving SDARE. Let \( P(0) \) be the solution of the first discrete time deterministic algebraic Riccati equation DARE

\[
0 = P(0) F + F' P(0) + Q - (P(0) G + S) R^{-1} (G' P(0) + S')
\]

and \( K(0) \) be solution of the second deterministic equation DARE

\[
K(0) = - R^{-1} (G' P(0) + S')
\]

Given \( P(\tau-1) \) define

\[
Q(\tau) = Q + \sum_{k=1}^r C_k^T P(\tau-1) C_k
\]

\[
R(\tau) = R + \sum_{k=1}^r D_k^T P(\tau-1) D_k
\]

\[
S(\tau) = S + \sum_{k=1}^r C_k^T P(\tau-1) D_k
\]

Let \( P(\tau) \) be the solution of

\[
0 = P(\tau) F + F' P(\tau) + Q(\tau) - (P(\tau) G + S(\tau)) R^{-1}(G' P(\tau) + S'(\tau))
\]

and

\[
K(\tau) = - R^{-1} \left( G' P(\tau) + S'(\tau) \right)
\]

If the iteration on \( P(\tau) \) converges, that is, for some \( \tau, P(\tau) \equiv P(\tau-1) \) then \( P(\tau) \) and \( K(\tau) \) are approximate solutions to SDARE

The solutions \( P(\tau) \) of the DARE is the kernel of the optimal cost of deterministic LQRs and since

(2) The matrix \( R \) is positive definite.

(3) The pair \( F, G \) is stabilizable.

(4) The pair \( Q^{1/2}, F \) is detectable where \( Q = (Q^{1/2})^2 Q^{1/2} \).

Because of the linear dynamics and quadratic cost, we expect
As expected the noisy system is more difficult to control than the noiseless system and the poles are smaller in norm. It should be noted that the above iteration diverged to infinity when the noise coefficients were increased from 0.1 to 1.

4. NONLINEAR STOCHASTIC INFINITE HORIZON DPE

Suppose the problem is not linear-quadratic, the dynamics is given by a nonlinear stochastic difference equation

\[ x^+ = f(x, u) + \sum_{k=1}^{r} w_k \gamma_k(x, u) \]

and the criterion to be minimized is

\[ \min_{u(t)} \left\{ \sum_{0}^{\infty} l(x, u) \right\} \]

As before the noise \( w(t) = (w_1, \ldots, w_r)' \) is a sequence of independent Gaussian vectors of zero mean and covariance \( F \Sigma F' \).

We assume that \( f(x, u), \gamma_k(x, u), l(x, u) \) are smooth functions that have Taylor polynomial expansions around \( x = 0, u = 0 \). We also assume that \( f(x, u) = O(x, u), \gamma_k(x, u) = O(x, u) \) and \( l(x, u) = O(x, u)^2 \) so their Taylor polynomial expansions are of the forms

\[
\begin{align*}
&f(x, u) = Fx + Gu + f_2(x, u) + \ldots + f^{[d]}(x, u) + O(x, u)^{d+1} \\
&\gamma_k(x, u) = C_kx + D_ku + \gamma_k^2(x, u) + \ldots + \gamma_k^{[d]}(x, u) + O(x, u)^{d+1} \\
&l(x, u) = \frac{1}{2} \left( x'Qx + 2x'Su + u'Ru \right) + l_2(x, u) + \ldots + l^{[d+1]}(x, u) + O(x, u)^{d+2}
\end{align*}
\]

where \([d]\) indicates the polynomial terms of homogeneous degree \( d \).

Then if they exist the optimal cost \( \pi(x) \) and optimal feedback \( u = \kappa(x) \) satisfy SIDPE (1. 2). The quantity to be minimized is a smooth function of \( u \) hence (1. 2) imply

\[
\pi(x) = E \left\{ \pi \left( f(x, \kappa(x)) + \sum_{k=1}^{r} w_k \gamma_k(x, \kappa(x)) \right) + l(x, \kappa(x)) \right\}
\]

\[
0 = E \left\{ \frac{\partial \pi}{\partial x} \left( f(x, \kappa(x)) + \sum_{k=1}^{r} w_k \gamma_k(x, \kappa(x)) \right) \right\} \times \frac{\partial f}{\partial u}(x, \kappa(x)) + \sum_{k} w_k \frac{\partial \gamma_k}{\partial u}(x, \kappa(x)) \]

\[
+ \frac{\partial l}{\partial u}(x, \kappa(x)) \]  

(7)

(8)

We call these the simplified Stochastic, Infinite Horizon Dynamic Programming Equations (sSIDPE). Of course the reverse implication is not necessarily true as the quantity to be minimized could have local minima or stationary points.

We assume that the optimal cost and optimal feedback have similar Taylor polynomial expansions.
\[
\pi(x) = \frac{1}{2} x' P x + \pi^3(x) + \ldots + \pi^{d+1}(x) + O(x)^{d+2}
\]
\[
\kappa(x) = K x + \kappa^2(x) + \ldots + \kappa^{d}(x) + O(x)^{d+1}
\]
We plug all these expansions into equations (7, 8). At lowest degrees, degree two in (7) and degree one in (8) we get the familiar SDARE (5, 6).
If (5, 6) are solvable then we may proceed to the next degrees, degree three in (7) and degree two in (8).
\[
\pi^3(x) = E \left\{ \pi^3 \left( (F + GK)x + \sum_k w_k (C_k + D_k K)x \right) \right\} + \sum_k x' (C_k + D_k K)' P^{-1} (x, K x) + \sum_{i \leq 3} (x, K x)
\]
\[
0 = E \left\{ \frac{\partial \pi^3}{\partial x} \right\} \left( (F + GK)x + \sum_k w_k (C_k + D_k K)x \right) \times \left( G + \sum_k w_k D_k \right) + \pi^3 (x, K x) \sum_k (F + GK)' P^{-1} (x, K x) \right\} + \sum_{i \leq 3} (x, K x)
\]
\[
+ (\kappa^2(x))^\prime \left( R + G' P G + \sum_k D_k P D_k \right) \tag{10}
\]
Notice the first equation (9) is a square linear equation for the unknown \(\pi^3(x)\), the other unknown \(\kappa^2(x)\) does not appear in it. If we can solve it for \(\pi^3(x)\) then we can solve the second equation (9) for \(\kappa^2(x)\) because of the second standard LQR assumption that \(R\) is invertible which implies \(R + G' P G + \sum_k D_k P D_k\) is also invertible. But again if the second LQR assumption does not hold, the matrix \(R + G' P G + \sum_k D_k P D_k\) might still be invertible.
In the deterministic case the eigenvalues of the linear operator
\[
\pi^3(x) \mapsto \pi^3 \left( (F + GK)x \right) \tag{11}
\]
are the products of three eigenvalues of \(F + GK\). Under the standard LQR assumptions all the eigenvalues of \(F + GK\) are in the open unit disc so any product of three eigenvalues of \(F + GK\) has norm less than one. Hence the operator
\[
\pi^3(x) \mapsto \pi^3 \left( (F + GK)x \right) \tag{12}
\]
is invertible. If the noise coefficients \(C_k, D_k\) are small relative to the eigenvalues of (11) then the operator
\[
\pi^3(x) \mapsto \pi^3(x) - E \left\{ \pi^3 \left( (F + GK)x + \sum_k w_k (C_k + D_k K)x \right) \right\} \tag{13}
\]
will also be invertible and so we can solve (9) for \(\pi^3(x)\) and then (10) for \(\kappa^2(x)\).
The first SIDPE equation for \(\pi^{d+1}(x)\) contains previously computed lower degree terms and the linear operator is
\[
\pi^{d+1}(x) \mapsto \pi^{d+1}(x) \tag{14}
\]
The eigenvalues of deterministic part of this operator
\[
\pi^{d+1}(x) \mapsto \pi^{d+1}(x) - \pi^{d+1} \left( (F + GK)x \right) \tag{15}
\]
are of the form \(1 - \lambda_i \ldots \lambda_{d+1}\) where \(\lambda_i\) are eigenvalues of \(F + GK\) which are strictly inside the unit disk. Hence (15) will be invertible and its stochastic perturbation (14) will be also if \(C_k\) and \(D_k\) are small enough.

5. NONLINEAR EXAMPLE

Here is a simple example with \(n = 2, m = 1, r = 2\). Consider a pendulum of length 1 m and mass 1 kg orbiting approximately 400 kilometers above Earth on the International Space Station (ISS). The "gravity constant" at this height is approximately \(g = 8.7 \text{ m/sec}^2\). The pendulum can be controlled by a torque \(u\) that can be applied at the pivot and there is damping at the pivot with linear damping constant \(c_1 = 0.1 \text{ kg/sec}\) and cubic damping constant \(c_3 = 0.05 \text{ kg sec/m}^2\). Let \(x_1\) denote the angle of pendulum measured counter clockwise from the outward pointing ray from the center of the Earth and let \(x_2\) denote the angular velocity. The continuous time deterministic equations of motion are
\[
\dot{x}_1 = x_2 \\
\dot{x}_2 = lg \sin x_1 - c_1 x_2 - c_3 x_2^3 + u
\]
The goal is to find a feedback \(u = \kappa(x)\) that stabilizes the pendulum to straight up in spite of the noises so we take the continuous time criterion to be
\[
\min_u \frac{1}{2} \int_0^\infty \|x\|^2 + u^2 \, dt
\]
We time discretize this problem using Euler’s method with a time step of 0.02 seconds to get the discrete time optimal control problem of minimizing
\[
\min_u 0.01 \sum_{t=0}^\infty \|x\|^2 + u^2
\]
subject to
\[
x_{1+} = x_1 + 0.02 x_2 \\
x_{2+} = x_2 + 0.02 \left( (lg \sin x_1 - c_1 x_2 - c_3 x_2^3 + u) \right)
\]
But the shape of the earth is not a perfect sphere and its density is not uniform so there are fluctuations in the "gravity constant". We model these relative fluctuations in the "gravity constant" by \(0.1 w_1\) although they are probably much smaller. There might also be relative fluctuations in the damping constants modeled by \(0.1 w_2\). We model these stochastically by two white noises,
\[
x_{1+} = x_1 + 0.02 x_2 \\
x_{2+} = x_2 + 0.02 \left( (lg \sin x_1 - c_1 x_2 - c_3 x_2^3 + u) \right) + 0.02 \left( 0.1 w_1 lg \sin x_1 - 0.1 w_2 (c_1 x_2 + c_3 x_2^3) \right)
\]
This is an example about how stochastic models with noise coefficients of order \(O(x, u)\) can arise. If the noise is modeling
an uncertain environment then its coefficients are likely to be $O(1)$. But if it is the model that is uncertain then noise coefficients are likely to be $O(x, u)$.

The linear coefficients of the dynamics are

$$\begin{align*}
F &= \begin{bmatrix} 1 & 0.02 \\ 0.1740 & 0.9980 \end{bmatrix}, & G &= \begin{bmatrix} 0 \\ 0.02 \end{bmatrix}, \\
Q &= \begin{bmatrix} 0.02 & 0 \\ 0 & 0.02 \end{bmatrix}, & R &= 0.02, & S &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
C_1 &= \begin{bmatrix} 0 & 0.0174 \\ 0 & 0 \end{bmatrix}, & C_2 &= \begin{bmatrix} 0 & 0 \\ 0 & -0.0002 \end{bmatrix}, \\
D_1 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, & D_2 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{align*}$$

The above iteration converges in six steps to the solution of SDARE (5, 6).

$$P = \begin{bmatrix} 54.9340 & 17.9795 \\ 17.9795 & 6.0744 \end{bmatrix}, \quad K = [-17.9795, -6.0744]$$

The eigenvalues of $F + GK$ are 0.9483 and 0.9282.

By way of comparison if we delete the noise terms from the problem then the solution to DARE is

$$P = \begin{bmatrix} 54.8930 & 17.9739 \\ 17.9739 & 6.0734 \end{bmatrix}, \quad K = [-16.9694, -5.7253]$$

and the eigenvalues of $F + GK$ are 0.9510 and 0.9325.

The dynamics is an odd function of $x, u$ so its quadratic and quartic terms are zero. The cubic terms are

$$\begin{align*}
f^{[3]}(x, u) &= \begin{bmatrix} 0 \\ -0.029x_1^3 - 0.001x_2^3 \end{bmatrix}, \\
\gamma_1^{[3]}(x, u) &= \begin{bmatrix} 0 \\ -0.0029x_3 \end{bmatrix}, \\
\gamma_2^{[3]}(x, u) &= \begin{bmatrix} 0 \\ -0.0001x_4 \end{bmatrix}
\end{align*}$$

and the quintic terms are

$$\begin{align*}
f^{[5]}(x, u) &= \begin{bmatrix} 0 \\ 0.00145x_5 \end{bmatrix}, \\
\gamma_1^{[5]}(x, u) &= \begin{bmatrix} 0 \\ 0.000145x_1^5 \end{bmatrix}, \\
\gamma_2^{[5]}(x, u) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{align*}$$

Because the Lagrangian is an even function and the dynamics is an odd function of $x, u$ we know that $\pi(x)$ is an even function of $x$ and $\kappa(x)$ is an odd function of $x$.

We have computed the optimal cost $\pi(x)$ to degree 6 and the optimal feedback $\kappa(x)$ to degree 5,

$$\begin{align*}
\pi(x) &= 27.4670x_1^2 + 17.9795x_2^2 + 3.0372x_2^2 \\
&\quad -4.4633x_1^3 - 2.7258x_3^2 + 0.4995x_2^2 \\
&\quad -0.0796x_1x_2^2 - 0.0169x_3^2 \\
&\quad 0.3860x_4^6 + 01976x_5^5x_2 + 0.0266x_1^4x_2^2 + 0.0021x_3^3
\end{align*}$$

Because the Lagrangian is an even function and the dynamics is an odd function of $x, u$ we know that $\pi(x)$ is an even function of $x$ and $\kappa(x)$ is an odd function of $x$.

We have computed the optimal cost $\pi(x)$ to degree 6 and the optimal feedback $\kappa(x)$ to degree 5,

$$\begin{align*}
\pi(x) &= 27.4670x_1^2 + 17.9795x_1x_2 + 3.0372x_2^2 \\
&\quad -4.4633x_1^3 - 2.7258x_3^2 + 0.4995x_2^2 \\
&\quad -0.0796x_1x_2^2 - 0.0169x_3^2 \\
&\quad 0.3860x_4^6 + 01976x_5^5x_2 + 0.0266x_1^4x_2^2 + 0.0021x_3^3
\end{align*}$$

Because the Lagrangian is an even function and the dynamics is an odd function of $x, u$ we know that $\pi(x)$ is an even function of $x$ and $\kappa(x)$ is an odd function of $x$.

We have computed the optimal cost $\pi(x)$ to degree 6 and the optimal feedback $\kappa(x)$ to degree 5,

$$\begin{align*}
\pi(x) &= 27.4670x_1^2 + 17.9795x_1x_2 + 3.0372x_2^2 \\
&\quad -4.4633x_1^3 - 2.7258x_3^2 + 0.4995x_2^2 \\
&\quad -0.0796x_1x_2^2 - 0.0169x_3^2 \\
&\quad 0.3860x_4^6 + 01976x_5^5x_2 + 0.0266x_1^4x_2^2 + 0.0021x_3^3
\end{align*}$$

Because the Lagrangian is an even function and the dynamics is an odd function of $x, u$ we know that $\pi(x)$ is an even function of $x$ and $\kappa(x)$ is an odd function of $x$.
\[ \pi(t, x) = l(t, x, \kappa(t, x)) \]
\[ + E \left\{ \pi \left( t + 1, f(t, x, \kappa(t, x)) + \sum_{k=1}^{r} w_k \gamma_k(t, x, \kappa(t, x)) \right) \right\} \]
\[ 0 = E \left\{ \partial \pi \partial x(t + 1, f(t, x, \kappa(t, x)) + \sum_{k=1}^{r} w_k \partial \gamma_k \partial x(t, x, \kappa(t, x)) \right\} \]
\[ + \partial \pi \partial t(t, x, \kappa(t, x)) \]  
(16) We call these equations the stochastic discrete time Riccati difference equations (SDRDE). These difference equations are solved backward in time from the terminal condition \( P(T) = P_T \).

Then we may proceed to the next degrees,
\[ \pi^{[3]}(t, x) = E \left\{ \pi^{[3]}(t + 1, z(t, x, w)) \right\} \]
\[ + x'(F(t) + G(t)K(t))'P(t + 1)f^{[2]}(t, x, Kx) \]
\[ + \sum_{k=1}^{r} x'(C_k(t) + D_k(t)K(t))'P(t + 1)\gamma_k^{[2]}(t, x, Kx) \]
\[ + l^{[3]}(t, x, Kx) \]
\[ 0 = E \left\{ \partial \pi^{[3]} \partial x(t, z(t, x, w)) \right\} \left\{ G(t) + \sum_{k=1}^{r} w_k D_k(t) \right\} \]
\[ + x'P(t + 1)\partial f^{[2]} \partial u(t, x, K(t)x) + \partial \pi^{[3]} \partial u(t, x, K(t)x) \]
\[ + (\kappa^{[2]}(t, x))'R(t) \]
where
\[ z(t, x, w) = (F(t) + G(t)K(t))x \]
\[ + \sum_{k=1}^{r} w_k (C_k(t) + D_k(t)K(t))x \]

Notice again the unknown \( \kappa^{[2]}(t, x) \) does not appear in the first equation which is linear difference equation for \( \pi^{[3]}(t, x) \) running backward in time from the terminal condition, \( \pi^{[3]}(t, x) = \pi_T^{[3]}(x) \). We can solve it and if \( R(t) \) is invertible then we can solve the second equation for \( \kappa^{[2]}(t, x) \). The higher degree terms can be found in a similar fashion.

7. CONCLUSION

We have shown how the Taylor polynomials of the optimal cost and optimal feedback for some stochastic, discrete time optimal control problems can be computed degree by degree.

REFERENCES