The Bellman Equation

Reza Shadmehr

In this document I will provide an explanation of the Bellman equation, which is a method for optimizing a cost function and arriving at a control policy.

1. Example of a game

Suppose that our states x refer to the position on a grid, as shown below. If we are at the goal state, then the state cost per time step is zero. If we are at any other state, the state cost per time step is 5. Let us use the term J_x to refer to this state cost per time step:

$$J_x = \begin{array}{ccc} 5 \| & 5 & 5 \\ 5 \| & 0 & 5 \\ 5 \| & 5 & 5 \\ 5 & 5 & 5 \end{array} \qquad (1)$$

The goal state is at row 2, col. 2, which means that if we are at this state, we incur no state costs. The double lines refer to a 'wall', preventing one to move from one state to the neighboring state. That is, there is a wall between the top left and top middle states. If we perform some action u (say, move from one box to the neighboring box), there will be a motor cost per time step, which we refer to with symbol J_u. The motor cost is one if we move $J_u = 1$, and zero otherwise. So the total cost per time step is:

$$\alpha^{(n)} = J_x + J_u^{(n)} \qquad (2)$$

The term $\pi\left(x^{(n)}\right)$ refers to the policy that we have. This policy specifies the action $u\left(x^{(n)}\right)$ that we will perform for each state x at time point n. For example, if we pick a random policy, then we might have actions that look like this:

$$\pi\left(x^{(n)}\right) = \begin{array}{ccc} \bullet \| & \rightarrow & \nwarrow \\ \downarrow \| & \searrow & \downarrow \\ \uparrow \| & \bullet & \swarrow \\ \nearrow & \rightarrow & \nwarrow \end{array} \qquad (3)$$

Suppose our final time step is p. If we are now at time point k, our objective is to find the policy that minimizes the total cost to go $\sum_{i=k}^{p} \alpha^{(i)}$. Let us define the goodness of each policy via a value function:

$$V_{\pi}\left(x^{(k)}\right) = \alpha^{(k)} + V_{\pi}\left(x^{(k+1)}\right) \tag{4}$$

If we are at the last time step p, then the value of our policy is simply the cost per step at this last time step:

$$V_{\pi}\left(x^{(p)}\right) = J_x + J_u^{(p)} \tag{5}$$

The optimum policy is one that minimizes Eq. (4), which is simply to do nothing:

$$\pi^*\left(x^{(p)}\right) = \begin{array}{c} \bullet \| \ \bullet \ \bullet \\ \bullet \| \ \bullet \ \bullet \\ \bullet \| \ \bullet \ \bullet \\ \bullet \ \ \bullet \ \ \bullet \end{array} \tag{6}$$

In that case, we have:

$$V_{\pi^*}\left(x^{(p)}\right) = J_x = \begin{array}{c} 5 \| \ 5 \ 5 \\ 5 \| \ 0 \ 5 \\ 5 \| \ 5 \ 5 \\ 5 \ \ 5 \ 5 \end{array} \tag{7}$$

The Bellman optimality principle states that an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. This means that in order to find the optimal policy for time point $p-1$, for each state x we must find the command that minimizes the following:

$$\pi^*\left(x^{(p-1)}\right) = \arg\min_u \left\{\alpha^{(p-1)} + V_{\pi^*}\left(x^{(p)}\right)\right\} \tag{8}$$

Once we have the optimal command for each state, the value of that state is:

$$V_{\pi^*}\left(x^{(p-1)}\right) = \min_u \left\{\alpha^{(p-1)} + V_{\pi^*}\left(x^{(p)}\right)\right\}$$

Consider the top middle state. If we were to produce an action that moves us down, we would have the state cost of 5, and motor cost of 1, and so $\alpha^{(p-1)} = 6$. The value of the state we get to is zero. So the total value of this action is 6. If we were to stay and not move, $\alpha^{(p-1)} = 5$ plus the value of the state that we get to (the current state), which is 5, for a sum total of 10. So the value of the action of moving down is 6, whereas the action of doing nothing has the value of 10. The value of the action of moving to the right is 11. So the best action that we can do for the top middle state is to move down. Similarly, the best action that we can do for the bottom middle state is to stay still. In this way, we can define the action that minimizes Eq. (8) for each state, resulting in our policy for time step $p-1$:

$$\pi^*\left(x^{(p-1)}\right) = \begin{array}{ccc} \bullet\| & \downarrow & \swarrow \\ \bullet\| & \bullet & \leftarrow \\ \bullet\| & \uparrow & \nwarrow \\ \bullet & \bullet & \bullet \end{array} \tag{9}$$

The value of this optimal policy is:

$$V_{\pi^*}\left(x^{(p-1)}\right) = \begin{array}{ccc} 10\| & 6 & 6 \\ 10\| & 0 & 6 \\ 10\| & 6 & 6 \\ 10 & 10 & 10 \end{array} \tag{10}$$

Before we proceed to the next step, it is worthwhile to take another look at Eq. (10). We have a value associated with each state. This value is the total cost that will be incurred if, starting at a given state, we were to perform the best sequence of actions possible. The sequence of actions will produce a sequence of cost per steps that together sum up to be the value assigned to each state. In a sense, when we find ourselves at a given state, the value of that state represents the lowest 'cost to go' that we hope to incur if we were to produce the best actions possible.

Now we repeat this process for time step $p-2$. Let us reconsider the top middle state. The value of staying still is: $J_x + J_u + V_{\pi^*}\left(x^{(p-1)}\right) = 5 + 0 + 6 = 11$. The value of moving down is $5 + 1 + 0 = 6$. The value of moving to the right is 12. The best action remains to move down. Consider the bottom middle state. The value of moving up is 12. The value of staying still is 15. The best action is to move up (or move to right-top neighbor). The optimal policy at time step $p-2$ is:

$$\pi^*\left(x^{(p-2)}\right) = \begin{array}{ccc} \bullet\| & \downarrow & \swarrow \\ \bullet\| & \bullet & \leftarrow \\ \bullet\| & \uparrow & \nwarrow \\ \nearrow & \uparrow & \uparrow \end{array} \tag{11}$$

The value of this optimal policy is:

$$V_{\pi^*}\left(x^{(p-2)}\right) = \begin{array}{ccc} 15\| & 6 & 6 \\ 15\| & 0 & 6 \\ 15\| & 6 & 6 \\ 12 & 12 & 12 \end{array} \tag{12}$$

We compute the optimal policy for time step $p-3$:

$$\pi^*\left(x^{(p-3)}\right) = \begin{matrix} \bullet\| & \downarrow & \swarrow \\ \bullet\| & \bullet & \leftarrow \\ \downarrow\| & \uparrow & \nwarrow \\ \nearrow & \uparrow & \uparrow \end{matrix} \tag{13}$$

The value of this optimal policy is:

$$V_{\pi^*}\left(x^{(p-3)}\right) = \begin{matrix} 20\| & 6 & 6 \\ 20\| & 0 & 6 \\ 18\| & 6 & 6 \\ 12 & 12 & 12 \end{matrix} \tag{14}$$

Similarly, we compute the optimal policy for time step $p-4$:

$$\pi^*\left(x^{(p-4)}\right) = \begin{matrix} \bullet\| & \downarrow & \swarrow \\ \downarrow\| & \bullet & \leftarrow \\ \downarrow\| & \uparrow & \nwarrow \\ \nearrow & \uparrow & \uparrow \end{matrix} \tag{15}$$

The value of this optimal policy is:

$$V_{\pi^*}\left(x^{(p-4)}\right) = \begin{matrix} 25\| & 6 & 6 \\ 24\| & 0 & 6 \\ 18\| & 6 & 6 \\ 12 & 12 & 12 \end{matrix} \tag{16}$$

Now the interesting result comes when we consider time step $p-5$. Consider the top left box. If we were to move down, the cost is 30. If we were to stay, the cost is also 30. So the optimal policy is either to stay still or move down, as both gives us the same cost. The reason for this is that the effort cost of moving is so large for this state (it is so far from the goal state) that the reward of getting to the goal only barely compensates for the cost of moving. Suppose we decide to stay, and so we have:

$$\pi^*\left(x^{(p-5)}\right) = \pi^*\left(x^{(p-4)}\right) \tag{17}$$

The value of this optimal policy is:

$$V_{\pi^*}\left(x^{(p-5)}\right) = \begin{matrix} 30\| & 6 & 6 \\ 24\| & 0 & 6 \\ 18\| & 6 & 6 \\ 12 & 12 & 12 \end{matrix} \tag{18}$$

However, at time step $p-6$, for the top left state our optimal action is to move down, and so we get:

$$\pi^*\left(x^{(p-6)}\right) = \begin{matrix} \downarrow & \downarrow & \swarrow \\ \downarrow & \bullet & \leftarrow \\ \downarrow & \uparrow & \nwarrow \\ \nearrow & \uparrow & \uparrow \end{matrix} \tag{19}$$

For each time step, we have used the Bellman equation (Eq. 8) to find the optimal feedback control policy.

2. Example of a linear system without noise

In conditions for which we are dealing with a linear dynamical system, the value function will turn out to be a quadratic function of state, and the control policy will become a linear function of state. These class of control problems are also called Linear Quadratic Regulators. Let us start with a linear system without noise.

$$\begin{aligned} \mathbf{x}^{(n+1)} &= A\mathbf{x}^{(n)} + B\mathbf{u}^{(n)} \\ \mathbf{y}^{(n)} &= C\mathbf{x}^{(n)} \end{aligned} \tag{20}$$

We have the following cost per step:

$$\begin{aligned} \alpha^{(n)} &= \mathbf{y}^{(n)T} T^{(n)} \mathbf{y}^{(n)} + \mathbf{u}^{(n)T} L \mathbf{u}^{(n)} \\ &= \mathbf{x}^{(n)T} C^T T^{(n)} C \mathbf{x}^{(n)} + \mathbf{u}^{(n)T} L \mathbf{u}^{(n)} \end{aligned} \tag{21}$$

Let us begin at the final time step $n = p$. At this time, the best action $\mathbf{u}^{(p)}$ that we can perform is one that minimizes the cost $\alpha^{(p)}$. That action is: $\mathbf{u}^{(p)} = 0$. That is, regardless of state, the optimal policy of action at the final time point is:

$$\pi^*\left(\mathbf{x}^{(p)}\right) = 0 \tag{22}$$

If we perform this optimal action, the value of the state we are at is:

$$V_{\pi^*}\left(\mathbf{x}^{(p)}\right) = \mathbf{x}^{(p)T} C^T T^{(p)} C \mathbf{x}^{(p)} \tag{23}$$

We see that at the final time point, the value function is a quadratic function of state. Let us define matrix $W^{(p)}$ as follows:

$$W^{(p)} = C^T T^{(p)} C \tag{24}$$

And so the value for the optimal policy can be written as:

$$V_{\pi^*}\left(\mathbf{x}^{(p)}\right) = \mathbf{x}^{(p)T} W^{(p)} \mathbf{x}^{(p)} \tag{25}$$

In order to find the optimal policy for time point $p-1$, for each state $\mathbf{x}$ we must find the command that minimizes the sum of the cost at the current time step, plus the value of the state that we arrive at after we produce the command:

$$\pi^*\left(\mathbf{x}^{(p-1)}\right) = \arg\min_{\mathbf{u}}\left\{\alpha^{(p-1)} + V_{\pi^*}\left(\mathbf{x}^{(p)}\middle|\mathbf{x}^{(p-1)},\mathbf{u}^{(p-1)}\right)\right\} \tag{26}$$

We can write the expression inside the brackets as:

$$\begin{aligned}\alpha^{(p-1)} &= \mathbf{x}^{(p-1)T}C^T T^{(p-1)}C\mathbf{x}^{(p-1)} + \mathbf{u}^{(p-1)T}L\mathbf{u}^{(p-1)}\\ V_{\pi^*}\left(\mathbf{x}^{(p)}\middle|\mathbf{x}^{(p-1)},\mathbf{u}^{(p-1)}\right) &= \left(A\mathbf{x}^{(p-1)} + B\mathbf{u}^{(p-1)}\right)^T W^{(p)}\left(A\mathbf{x}^{(p-1)} + B\mathbf{u}^{(p-1)}\right)\\ &= \mathbf{x}^{(p-1)T}A^T W^{(p)}A\mathbf{x}^{(p-1)} + \mathbf{u}^{(p-1)T}B^T W^{(p)}B\mathbf{u}^{(p-1)}\\ &\quad +2\mathbf{u}^{(p-1)T}B^T W^{(p)}A\mathbf{x}^{(p-1)}\end{aligned} \tag{27}$$

To minimize the sum in Eq. (26), we find its derivative with respect to $\mathbf{u}^{(p-1)}$ and set it equal to zero.

$$2L\mathbf{u}^{(p-1)} + 2B^T W^{(p)}B\mathbf{u}^{(p-1)} + 2B^T W^{(p)}A\mathbf{x}^{(p-1)} = 0 \tag{28}$$

This gives us the optimal commands:

$$\mathbf{u}^{*(p-1)} = -\left(L + B^T W^{(p)}B\right)^{-1} B^T W^{(p)}A\mathbf{x}^{(p-1)} \tag{29}$$

Let us define the following matrix:

$$G^{(p-1)} = \left(L + B^T W^{(p)}B\right)^{-1} B^T W^{(p)}A \tag{31}$$

We now can write the optimal policy as follows:

$$\pi^*\left(\mathbf{x}^{(p-1)}\right) = -G^{(p-1)}\mathbf{x}^{(p-1)} \tag{32}$$

The value of each state under the optimal policy can be written as

$$\begin{aligned}V_{\pi^*}\left(\mathbf{x}^{(p-1)}\right) &= \alpha\left(\mathbf{x}^{(p-1)},\mathbf{u}^{*(p-1)}\right) + V_{\pi^*}\left(\mathbf{x}^{(p)}\middle|\mathbf{x}^{(p-1)},\mathbf{u}^{*(p-1)}\right)\\ &= \mathbf{x}^{(p-1)T}B^T T^{(p-1)}B\mathbf{x}^{(p-1)} + \mathbf{x}^{(p-1)T}G^{(p-1)T}LG^{(p-1)}\mathbf{x}^{(p-1)}\\ &\quad +\mathbf{x}^{(p-1)T}A^T W^{(p)}A\mathbf{x}^{(p-1)} + \mathbf{x}^{(p-1)T}G^{(p-1)T}B^T W^{(p)}BG^{(p-1)}\mathbf{x}^{(p-1)}\\ &\quad -2\mathbf{x}^{(p-1)T}G^{(p-1)T}B^T W^{(p)}A\mathbf{x}^{(p-1)}\end{aligned}$$

Notice that the value function is a quadratic function of state. We can simplify it a little using the definition of $G^{(p-1)}$:

$$\begin{aligned}G^{(p-1)T}LG^{(p-1)} + G^{(p-1)T}B^T W^{(p)}BG^{(p-1)} &= G^{(p-1)T}\left(L + B^T W^{(p)}B\right)G^{(p-1)}\\ &= G^{(p-1)T}B^T W^{(p)}A\end{aligned} \tag{33}$$

Let us define $W^{(p-1)}$ as follows:

$$\begin{aligned}W^{(p-1)} &= B^T T^{(p-1)}B + A^T W^{(p)}A - 2G^{(p-1)T}B^T W^{(p)}A + G^{(p-1)T}B^T W^{(p)}A\\ &= B^T T^{(p-1)}B + A^T W^{(p)}A - G^{(p-1)T}B^T W^{(p)}A\end{aligned} \tag{34}$$

We can write the value function as:

$$V_{\pi^*}\left(\mathbf{x}^{(p-1)}\right) = \mathbf{x}^{(p-1)} W^{(p-1)} \mathbf{x}^{(p-1)} \tag{35}$$

We now have a recipe. For step $p-2$ we have:

$$G^{(p-2)} = \left(L + B^T W^{(p-1)} B\right)^{-1} B^T W^{(p-1)} A \tag{36}$$

And the following value function:

$$\begin{aligned} V_{\pi^*}\left(\mathbf{x}^{(p-2)}\right) &= \mathbf{x}^{(p-2)} W^{(p-2)} \mathbf{x}^{(p-2)} \\ W^{(p-2)} &= B^T T^{(p-2)} B + A^T W^{(p-1)} A - G^{(p-2)T} B^T W^{(p-1)} A \end{aligned} \tag{37}$$

And so the procedure is as follows: starting from the last time point, we compute $G^{(p)}$ (which is zero) and $W^{(p)}$ (Eq. 24). We next move time point $p-1$ and compute $G^{(p-1)}$ (Eq. 31) and $W^{(p-1)}$ (Eq. 34). We then use Eq. (36) to compute $G^{(p-2)}$ and $W^{(p-2)}$ (Eq. 37). And so on, until we reach time point 0. For each time step, we will have a policy that transforms our current state into a motor command.

As an example, let us consider moving a single joint model of the elbow. The state of the system is defined by its position x and velocity $\dot{x}$ (referring to angular position and velocity). The dynamics of the system are described as follows:

$$\begin{aligned} x_1 &\equiv x \\ x_2 &\equiv \dot{x} \\ \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{m} \end{bmatrix} \begin{bmatrix} 0 \\ u \end{bmatrix} \\ \dot{\mathbf{x}} &= A_c \mathbf{x} + C_c \mathbf{u} \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x} \end{aligned}$$

The above equations are written in continuous time. To represent it in discrete time (with a time step of Δt), we can write the discrete equations as follows:

$$\begin{aligned} \mathbf{x}^{(k+1)} &= A\mathbf{x}^{(k)} + B\mathbf{u}^{(k)} \\ \mathbf{y}^{(k)} &= C\mathbf{x}^{(k)} \\ A &= \left(I + A_c \Delta t\right) \\ B &= B_c \Delta t \\ C &= C_c \end{aligned}$$

I wanted the elbow to make a movement that ended at a goal state of $x(t = 300ms) = 0.5$, with zero velocity, and held there for an additional 100ms. I used the following parameter values for the arm:

$$k = 3N.m \;/\; rad$$
$$b = 0.45N.m.s \;/\; rad$$
$$m = 0.3kg.m \;/\; rad$$

I set the state cost matrix T to have the following values as a function of time:

3. Example of a linear system with signal dependent noise

Let us consider a simple scalar system of the form:

$$x^{(t+1)} = ax^{(t)} + b\left(u^{(t)} + \varepsilon_u^{(t)}\right) \qquad \varepsilon_u^{(t)} \sim N\left(0, c^2\left(u^{(t)}\right)^2\right)$$
$$y^{(t)} = x^{(t)} + \varepsilon_y \qquad \varepsilon_y \sim N\left(0, \sigma_y^2\right)$$

In this system, the state x is a scalar, and so is the observation y. However, notice that in this system the noise is signal dependent. That is, the variance of the noise depends on the size of the motor commands. We begin by expressing the random variable $\varepsilon_u^{(t)}$ in terms of random variable $\phi^{(t)} \sim N\left(0,1\right)$ and $u^{(t)}$:

$$\varepsilon_u^{(t)} = cu^{(t)}\phi^{(t)}$$

Let us suppose that the cost per step is: $\alpha^{(t)} = \alpha_1\left(x^{(t)}\right)^2 + \alpha_2\left(u^{(t)}\right)^2$. This implies that at the last time point p the optimal policy $\pi^*\left(x^{(p)}\right) = 0$ and the value of the states achieved under this policy is $V_{\pi*}\left(x^{(p)}\right) = \alpha_1\left(x^{(p)}\right)^2$. We now find the optimal policy for time step $p-1$.

We begin by computing the term $E\left[V_{\pi*}\left(x^{(p)}\right)\middle| x^{(p-1)}, u^{(p-1)}\right]$.

$$E\left[\left(x^{(p)}\right)^2\right] = \mathrm{var}\left[x^{(p)}\right] + E\left[x^{(p)}\right]^2$$
$$E\left[V_{\pi*}\left(x^{(p)}\right)\middle| x^{(p-1)}, u^{(p-1)}\right] = \alpha_1 E\left[\left(x^{(p)}\right)^2\right]$$
$$= \alpha_1\left(\mathrm{var}\left[x^{(p)}\right] + E\left[x^{(p)}\right]^2\right)$$
$$= \alpha_1\left(b^2c^2\left(u^{(p-1)}\right)^2 + \left(ax^{(p-1)} + bu^{(p-1)}\right)^2\right)$$

The cost that we need to minimize at time step p-1 is:

$$V_\pi\left(x^{(p-1)}\right) = \alpha_1\left(x^{(p-1)}\right)^2 + \alpha_2\left(u^{(p-1)}\right)^2 + E\left[V_{\pi*}\left(x^{(p)}\right)\middle| x^{(p-1)}, u^{(p-1)}\right]$$
$$= \alpha_1\left(x^{(p-1)}\right)^2 + \alpha_2\left(u^{(p-1)}\right)^2 + \alpha_1\left(b^2c^2\left(u^{(p-1)}\right)^2 + \left(ax^{(p-1)} + bu^{(p-1)}\right)^2\right)$$

We find $u^{(p-1)}$ that minimizes this cost:

$$\frac{dV}{du^{(p-1)}} = 2\alpha_2 u^{(p-1)} + 2\alpha_1 b^2 c^2 u^{(p-1)} + 2\alpha_1 b u^{(p-1)} + 2\alpha_1 a b x^{(p-1)} = 0$$

$$u^{(p-1)} = -\underbrace{\left(\alpha_2 + \alpha_1 b^2 c^2 + \alpha_1 b\right)^{-1} \alpha_1 a b}_{g} x^{(p-1)}$$

$$\pi^*\left(x^{(p-1)}\right) = -g^{(p-1)} x^{(p-1)}$$

However, because x is a random variable, at any time point we will have an estimate of it, $\hat{x}$. And so our optimal policy at time point p-1 is $\pi^*\left(\hat{x}^{(p-1)}\right) = -g^{(p-1)}\hat{x}^{(p-1)}$.

Using our policy for time step p-1, we can compute the value function $V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right)$ and demonstrate that it is a quadratic function of $x^{(p-1)}$, and the error in estimate of that state $x^{(p-1)} - \hat{x}^{(p-1)}$.

$$\begin{aligned}
V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right) &= \alpha_1\left(x^{(p-1)}\right)^2 + \alpha_2\left(g^{(p-1)}\hat{x}^{(p-1)}\right)^2 \\
&\quad + \alpha_1 b^2 c^2\left(g^{(p-1)}\hat{x}^{(p-1)}\right)^2 + \alpha_1\left(a x^{(p-1)} - b g^{(p-1)}\hat{x}^{(p-1)}\right)^2 \\
&= \left(\alpha_1 + \alpha_1 a^2\right)\left(x^{(p-1)}\right)^2 + \left(\alpha_2 + \alpha_1 b^2 c^2 + \alpha_1 b^2\right)\left(g^{(p-1)}\hat{x}^{(p-1)}\right)^2 \\
&\quad - 2\alpha_1 a b g^{(p-1)} x^{(p-1)}\hat{x}^{(p-1)} \\
&= \left(\alpha_1 + \alpha_1 a^2\right)\left(x^{(p-1)}\right)^2 + \underbrace{\alpha_1 a b g^{(p-1)}}_{z}\left(\hat{x}^{(p-1)}\right)^2 - 2\alpha_1 a b g^{(p-1)} x^{(p-1)}\hat{x}^{(p-1)}
\end{aligned}$$

$$z\hat{x}^2 - 2zx\hat{x} = z\left(x - \hat{x}\right)^2 - zx^2$$

$$\begin{aligned}
V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right) &= \left(\alpha_1 + \alpha_1 a^2 - \alpha_1 a b g^{(p-1)}\right)\left(x^{(p-1)}\right)^2 + \alpha_1 a b g^{(p-1)}\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2 \\
&= w_1^{(p-1)}\left(x^{(p-1)}\right)^2 + w_2^{(p-1)}\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2
\end{aligned}$$

Now let us consider the time step p-2. We can observe $y^{(t)}$ and write the equation for the Kalman gain. As we will see, the Kalman gain will not depend on u.

$$\begin{aligned}
\hat{x}^{(t|t)} &= \hat{x}^{(t|t-1)} + k^{(t)}\left(y^{(t)} - \hat{x}^{(t|t-1)}\right) \\
&= \left(1 - k^{(t)}\right)\hat{x}^{(t|t-1)} + k^{(t)} x^{(t)} + k^{(t)}\varepsilon_y \\
\mathrm{var}\left(\hat{x}^{(t|t)}\right) &\equiv P^{(t|t)} \\
&= \left(1 - k^{(t)}\right)^2 \mathrm{var}\left(\hat{x}^{(t|t-1)}\right) + \left(k^{(t)}\right)^2 \sigma_y^2 \\
&= \left(1 - 2k^{(t)} + \left(k^{(t)}\right)^2\right) P^{(t|t-1)} + \left(k^{(t)}\right)^2 \sigma_y^2 \\
\frac{dP^{(t|t)}}{dk^{(t)}} &= -2P^{(t|t-1)} + 2P^{(t|t-1)} k^{(t)} + 2\sigma_y^2 k^{(t)} \\
k^{(t)} &= \frac{P^{(t|t-1)}}{P^{(t|t-1)} + \sigma_y^2}
\end{aligned}$$

At time point t, our estimate of $\hat{x}^{(t)}$ is simply the prior estimate $\hat{x}^{(t|t-1)}$. Let us write $\hat{x}^{(t)}$ in terms of $\hat{x}^{(t-1)}$, $u^{(t-1)}$ and the Kalman gain.

$$\begin{aligned}
\hat{x}^{(t|t)} &= \hat{x}^{(t|t-1)} + k^{(t)}\left(y^{(t)} - \hat{x}^{(t|t-1)}\right) \\
\hat{x}^{(t+1|t)} &= a\hat{x}^{(t|t)} + bu^{(t)} \\
&= a\hat{x}^{(t|t-1)} + ak^{(t)}\left(y^{(t)} - \hat{x}^{(t|t-1)}\right) + bu^{(t)} \\
\hat{x}^{(t+1)} &= a\hat{x}^{(t)} + ak^{(t)}\left(y^{(t)} - \hat{x}^{(t)}\right) + bu^{(t)}
\end{aligned}$$

At time point $p-1$ we showed that the value function under the optimal policy $V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right)$ is a quadratic function of $x^{(p-1)}$ and the error in estimate of that state $x^{(p-1)} - \hat{x}^{(p-1)}$. Let us write that relationship as $V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right) = w_1^{(p-1)}\left(x^{(p-1)}\right)^2 + w_2^{(p-1)}\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2$ and then find the optimal policy for time point $p-2$.

$$\begin{aligned}
V_{\pi}\left(x^{(p-2)}\right) &= \alpha_1\left(x^{(p-2)}\right)^2 + \alpha_2\left(u^{(p-2)}\right)^2 \\
&\quad + E\left[V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right)\middle| x^{(p-2)}, u^{(p-2)}, \hat{x}^{(p-2)}\right] \\
E\left[V_{\pi^*}\left(x^{(p-1)}, \hat{x}^{(p-1)}\right)\middle|\ldots\right] &= w_1^{(p-1)}\left(ax^{(p-2)} + bu^{(p-2)}\right)^2 + w_1^{(p-1)}b^2c^2\left(u^{(p-2)}\right)^2 \\
&\quad + w_2^{(p-1)}E\left[\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2\middle|\ldots\right] \\
E\left[\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2\middle|\ldots\right] &= E\left[\left(ax^{(p-2)} + bu^{(p-2)} + b\varepsilon_u - a\hat{x}^{(p-2)} - ak^{(p-2)}\left(y^{(p-2)} - \hat{x}^{(p-2)}\right) - bu^{(p-2)}\right)^2\right] \\
&= E\left[\left(a\left(x^{(p-2)} - \hat{x}^{(p-2)}\right) - ak^{(p-2)}\left(x^{(p-2)} + \varepsilon_y - \hat{x}^{(p-2)}\right) + b\varepsilon_u\right)^2\right] \\
&= E\left[\left(\left(a - ak^{(p-2)}\right)\left(x^{(p-2)} - \hat{x}^{(p-2)}\right) - ak^{(p-2)}\varepsilon_y + b\varepsilon_u\right)^2\right] \\
d &\equiv \left(a - ak^{(p-2)}\right)\left(x^{(p-2)} - \hat{x}^{(p-2)}\right) \\
E\left[\left(x^{(p-1)} - \hat{x}^{(p-1)}\right)^2\middle|\ldots\right] &= E\left[d^2 - 2adk^{(p-2)}\varepsilon_y + 2bd\varepsilon_u + a^2\left(k^{(p-2)}\right)^2\varepsilon_y^2 - 2abk^{(p-2)}\varepsilon_y\varepsilon_u + b^2\varepsilon_u^2\right] \\
&= d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2 + b^2c^2\left(u^{(p-2)}\right)^2 \\
V_{\pi}\left(x^{(p-2)}\right) &= \alpha_1\left(x^{(p-2)}\right)^2 + \alpha_2\left(u^{(p-2)}\right)^2 + w_1^{(p-1)}\left(ax^{(p-2)} + bu^{(p-2)}\right)^2 + w_1^{(p-1)}b^2c^2\left(u^{(p-2)}\right)^2 \\
&\quad + w_2^{(p-1)}\left(d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2 + b^2c^2\left(u^{(p-2)}\right)^2\right) \\
\frac{dV_{\pi}\left(x^{(p-2)}\right)}{du^{(p-2)}} &= 2\left(\alpha_2 + w_1^{(p-1)}b^2c^2 + w_2^{(p-1)}b^2c^2 + w_1^{(p-1)}b^2\right)u^{(p-2)} + 2abw_1^{(p-1)}x^{(p-2)} \\
\pi^*\left(x^{(p-2)}\right) &= -\underbrace{\left(\alpha_2 + w_1^{(p-1)}b^2c^2 + w_2^{(p-1)}b^2c^2 + w_1^{(p-1)}b^2\right)^{-1}abw_1^{(p-1)}}_{g}x^{(p-2)} \\
&= -g^{(p-2)}x^{(p-2)}
\end{aligned}$$

The best that we can do is implement the policy $\pi^*\left(x^{(p-2)}\right) = -g^{(p-2)}\hat{x}^{(p-2)}$. Let us show that under this policy, the value function remains quadratic in terms $x^{(p-2)}$ and the error in estimate of that state $x^{(p-2)} - \hat{x}^{(p-2)}$.

$$
\begin{aligned}
V_{\pi*}\left(x^{(p-2)}\right) &= \alpha_1\left(x^{(p-2)}\right)^2 + \alpha_2\left(u^{(p-2)}\right)^2 + w_1^{(p-1)}\left(ax^{(p-2)} + bu^{(p-2)}\right)^2 + w_1^{(p-1)}b^2c^2\left(u^{(p-2)}\right)^2 \\
&\quad + w_2^{(p-1)}\left(d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2 + b^2c^2\left(u^{(p-2)}\right)^2\right) \\
&= \left(\alpha_1 + w_1^{(p-1)}a^2\right)\left(x^{(p-2)}\right)^2 + 2w_1^{(p-1)}abx^{(p-2)}u^{(p-2)} \\
&\quad + \left(\alpha_2 + b^2w_1^{(p-1)} + w_1^{(p-1)}b^2c^2 + w_2^{(p-1)}b^2c^2\right)\left(u^{(p-2)}\right)^2 + w_2^{(p-1)}\left(d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2\right) \\
&= \left(\alpha_1 + w_1^{(p-1)}a^2\right)\left(x^{(p-2)}\right)^2 - 2abw_1^{(p-1)}g^{(p-2)}x^{(p-2)}\hat{x}^{(p-2)} \\
&\quad + abw_1^{(p-1)}g^{(p-2)}\left(\hat{x}^{(p-2)}\right)^2 + w_2^{(p-1)}\left(d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2\right)
\end{aligned}
$$

If we note that $z\hat{x}^2 - 2zx\hat{x} = z\left(x - \hat{x}\right)^2 - zx^2$, then we can write the value function under the optimal policy as:

$$
\begin{aligned}
V_{\pi*}\left(x^{(p-2)}\right) &= \left(\alpha_1 + w_1^{(p-1)}a^2 - abw_1^{(p-1)}g^{(p-2)}\right)\left(x^{(p-2)}\right)^2 + abw_1^{(p-1)}g^{(p-2)}\left(x^{(p-2)} - \hat{x}^{(p-2)}\right)^2 \\
&\quad + w_2^{(p-1)}\left(d^2 + a^2\left(k^{(p-2)}\right)^2\sigma_y^2\right) \\
&= w_1^{(p-2)}\left(x^{(p-2)}\right)^2 + w_2^{(p-2)}\left(x^{(p-2)} - \hat{x}^{(p-2)}\right) + w_3^{(p-2)}
\end{aligned}
$$

Now we can summarize the algorithm. At any time point t, the optimal policy and the value of that policy are:

$$
\begin{aligned}
\pi^*\left(x^{(t)}\right) &= -g^{(t)}\hat{x}^{(t)} \\
V_{\pi*}\left(x^{(t)}, \hat{x}^{(t)}\right) &= w_1^{(t)}\left(x^{(t)}\right)^2 + w_2^{(t)}\left(x^{(t)} - \hat{x}^{(t)}\right)^2 + w_3^{(t)}
\end{aligned}
$$

At the last time point we have:

$$
g^{(p)} = 0 \qquad w_1^{(p)} = \alpha_1 \qquad w_2^{(p)} = 0 \qquad w_3^{(p)} = 0
$$

At any other time point we have:

$$
\begin{aligned}
g^{(p-1)} &= \left(\alpha_2 + w_1^{(p)}b^2c^2 + w_1^{(p)}b + w_2^{(p-1)}b^2c^2\right)^{-1} w_1^{(p)}ab \\
w_1^{(p-1)} &= \alpha_1 + w_1^{(p)}a^2 - w_1^{(p)}abg^{(p-1)} \\
w_2^{(p-1)} &= w_1^{(p)}abg^{(p-1)} \\
w_3^{(p-1)} &= w_2^{(p)}\left(d^2 + a^2\left(k^{(p-1)}\right)^2\sigma_y^2\right)
\end{aligned}
$$