MDPs and (PO)MDPs

Nuances, simplifications, generalizations

Cathy Wu

6.7950 Reinforcement Learning: Foundations and Methods
Readings

1. DPOC vol 1, §1.4, §4.1-4.2
Outline

1. MDPs
2. Partially observed problems
Outline

1. MDPs
   a. Assumptions
   b. Sufficiency of Markov policies
   c. Sufficiency of stationary policies
   d. Sufficiency of deterministic policies

2. Partially observed problems
Learning objective

When using MDPs to model a problem of interest, it is key to understand the underlying assumptions, properties, and generalizations of MDPs.
2015:

Markov Decision Process: the Assumptions

*Stationarity assumption:* the dynamics and reward do not change over time

\[ p(s'|s, a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a) \quad r(s, a, s') \]

*Rule of thumb:* stationary \(\rightarrow\) more repeated “tail subproblems” \(\rightarrow\) easier to solve (i.e., benefits from DP recursion)

*Possible relaxations*

- Identify and add/remove the non-stationary components (e.g., cyclo-stationary dynamics)
- Identify the time-scale of the changes
- Work on finite horizon problems
ATARI Breakout

\[ P \left[ s_{t+1} = \begin{array}{c}
\text{image 1}
\end{array} \right| s_t = \begin{array}{c}
\text{image 2}
\end{array}, \text{no-move} \right] \]
ATARI Breakout

Recall: An MDP satisfies the Markovian property if
\[
P(s_{t+1} = s | \tau_t, a_t) = P(s_{t+1} = s | s_t, a_t, s_{t-1}, a_{t-1}, ..., s_0, a_0) = P(s_{t+1} = s | s_t, a_t)
\]
i.e., the current state \( s_t \) and action \( a_t \) are sufficient for predicting the next state \( s \).
ATARI Breakout

\[ S_t = \{ \} \]
Non-Markovian dynamics may be unavoidable: partial observation, multi-agent settings, nonstationary dynamics

Possible relaxation

- Partially observable Markov decision process (POMDP)
- Two more components
  - $\Omega$, a set of observations
  - $O : S \times \Omega \to \mathbb{R}_{\geq 0}$, the observation probability distribution
Markov Decision Process: the Assumptions

*Time assumption*: time is discrete

\[ t \rightarrow t + 1 \]

*Rule of thumb*: shorter horizon \( \rightarrow \) easier to solve

*Possible relaxations*
- Identify the proper time granularity
- Most of MDP literature extends to continuous time
ATARI Breakout

\[ a_t = \text{left} \]
ATARI Breakout

$a_t = \text{left}$

$t + 1$

Too fine-grained resolution
ATARI Breakout

\[ a_t = \text{left} \]
ATARI Breakout

\[ a_t = \text{left} \]

Too coarse-grained resolution
Markov Decision Process: the Assumptions

*Reward assumption*: the reward is uniquely defined by a transition (or part of it)

\[ r(s, a, s') \]

*Rule of thumb*: denser reward \( \Rightarrow \) extent to which each state-value updates towards the optimal solution with each update \( \Rightarrow \) easier to solve

*Possible relaxations*

- Distinguish between global goal and reward function
- Move to inverse reinforcement learning (IRL) to induce the reward function from desired behaviors
ATARI Breakout

Reward: extent to which paddle is moving in the optimal direction

Reward: score

vs

Reward: score > human baseline

Reward: win/lose
Outline

1. MDPs
   a. Assumptions
   b. Sufficiency of Markov policies
   c. Sufficiency of stationary policies
   d. Sufficiency of deterministic policies

2. Partially observed problems
Question

What is an appropriate class of policies when solving MDPs?
A decision rule $d$ can be
- Deterministic: $d: S \to A$,
- Stochastic: $d: S \to \Delta(A)$,
- History-dependent: $d: H_t \to A$,
- Markov: $d: S \to \Delta(A)$,

A policy (strategy, plan) can be
- Stationary: $\pi = (d, d, d, ... )$,
- (More generally) Non-stationary: $\pi = (d_0, d_1, d_2, ... )$

For simplicity, we will typically write $\pi$ instead of $d$ for stationary policies, and $\pi_t$ instead of $d_t$ for non-stationary policies. Except here!
The (General) Optimization Problem

\[
\begin{align*}
\max V^{\pi}(s_0) &= \max \mathbb{E}[r(s_0, d_0(a_0|s_0)) + \gamma r(s_1, d_1(a_1|s_0, s_1)) + \gamma^2 r(s_2, d_2(a_2|s_0, s_1, s_2)) + \ldots]
\end{align*}
\]
Plan to Simplify the Optimization Problem

1. Reduce the search space
   i. History-based $\Rightarrow$ Markov decision rules
   ii. Non-stationary $\Rightarrow$ Stationary policies
      $\Rightarrow$ Focus on stationary policies with Markov decision rules

2. Leverage Markov property of the MDP to “simplify” the value function

3. Stochastic $\Rightarrow$ Deterministic decision rules
   $\Rightarrow$ Focus on stationary policies with deterministic Markov decision rules
Theorem (Bertsekas (2007))

Consider an MDP with $|A| < \infty$ and an initial distribution $\beta$ over states such that $|\{s \in S : \beta(s) > 0\}| < \infty$. For any policy $\pi$, let

$$p_t^\pi(s,a) = \mathbb{P}[S_t = s, A_t = a]; \quad p_t^\pi(s) = \mathbb{P}[S_t = s]$$

Then for any history-based policy $\pi$ there exists a Markov policy $\bar{\pi}$ such that

$$p_t^\bar{\pi}(s,a) = p_t^\pi(s,a); \quad p_t^\bar{\pi}(s) = p_t^\pi(s)$$

For any $s \in S, a \in A$, and $t \in \mathbb{N}^+$. 

$\implies$ Markov policies are as “expressive” as history-based policies.

Intuition: Recall that the MDP is Markovian! No need for memory in the policy if there is no memory in the system.
For any $\pi = (d_0, d_1, \ldots)$ with $d_t$ a randomized history-dependent decision rule, let $\bar{\pi} = \bar{d}_0, \bar{d}_1, \ldots$ be a randomized Markov policy such that

$$
\bar{d}_t(a|s) = \frac{p_t^{\pi}(s, a)}{p_t^{\pi}(s)}
$$

Base case. For any $s$, $p^\bar{\pi}_0(s) = p^\pi_0(s)$ by definition. And

$$
p^\bar{\pi}_0(s, a) = p^\bar{\pi}_0(s)\bar{d}_0(a|s) = p^\bar{\pi}_0(s)\frac{p^\pi_0(s, a)}{p^\pi_0(s)} = p^\bar{\pi}_0(s)\frac{p^\pi_0(s, a)}{p^\pi_0(s)} = p^\pi_0(s, a)
$$
Proof: From History-Based to Markov Policies

*Induction.* For any $s$ and some $t > 0$, $p_t^\pi (s) = p_t^\pi (s)$ and $p_t^\pi (s, a) = p_t^\pi (s, a)$ by inductive assumption. Then:

$$p_{t+1}^\pi (s_{t+1}) = \sum_{s_t, a_t} p_t^\pi (s_t, a_t) p (s_{t+1} | s_t, a_t)$$

$$= \sum_{s_t, a_t} p_t^\pi (s_t) \tilde{d}_t (a_t | s_t) p (s_{t+1} | s_t, a_t)$$

$$= \sum_{s_t, a_t} p_t^\pi (s_t) \frac{p_t^\pi (s_t, a_t)}{p_t^\pi (s_t)} p (s_{t+1} | s_t, a_t)$$

$$= \sum_{s_t, a_t} p_t^\pi (s_t, a_t) p (s_{t+1} | s_t, a_t)$$

$$= p_{t+1}^\pi (s_{t+1})$$

The essence of why this works: the MDP is Markovian!

If non-Markovian, this final step would not hold.

Similar for $p_{t+1}^\pi (s_{t+1}, a_{t+1}) = p_{t+1}^\pi (s_{t+1}, a_{t+1})$
From Non-Stationary to Stationary Policies

Theorem (Bertsekas (2007))

Consider an MDP with $|A| < \infty$ and an initial distribution $\beta$ over states such that $|\{s \in S : \beta(s) > 0 \}| < \infty$.

Then for any non-stationary policy $\pi$ there exists a stationary policy $\bar{\pi}$ such that

$$\rho^\pi_Y(s, a) = \rho^\pi_Y(s, a); \quad \rho^\pi_Y(s) = \rho^\pi_Y(s)$$

For any $s \in S$, $a \in A$, and $t \in \mathbb{N}^+$. 

• Intuition: Again, Markovian!
• $\rho$ is the discounted occupancy measure.

$\Rightarrow$ Stationary policies are as “expressive” as non-stationary policies.
$\Rightarrow$ Stationary policies can “generate” any value function.
The Discounted Occupancy Measure $\rho$

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, d_t(s_t))\right]$$

$$= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[r(s_t, d_t(s_t))]$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{s,a} \mathbb{P}[S_t = s, A_t = a] r(s, a)$$

$$= \frac{1}{1 - \gamma} \sum_{s,a} (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}[S_t = s, A_t = a] r(s, a)$$

$$= \frac{1}{1 - \gamma} \sum_{s,a} \rho^\pi_{\gamma}(s, a) r(s, a)$$
Proof: From Non-Stationary to Stationary Policies

tl;dr (strategy):

1. Define $\tilde{\pi}(a|s') = \frac{\rho_{\pi}(s',a)}{\rho_{\pi}(s')}$. 
2. Show that $\rho_{\pi}(s)$ and $\rho_{\pi}(s)$ end up being the same.
Proof: From Non-Stationary to Stationary Policies

State discounted occupancy measure for stationary policy $\bar{\pi}$ (with Markov decision rules)

\[
\rho^\bar{\pi}_\gamma(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}[S_t = s] \\
= (1 - \gamma) \beta(s) + (1 - \gamma) \sum_{t=1}^{\infty} \gamma^t \mathbb{P}[S_t = s] \\
= (1 - \gamma) \beta(s) + (1 - \gamma) \gamma \sum_{t=1}^{\infty} \sum_{s'} \sum_a \mathbb{P}[S_{t-1} = s', A_{t-1} = a] p(s'|s, a) \\
= (1 - \gamma) \beta(s) + \gamma \sum_{s'} (1 - \gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}[S_{t-1} = s'] \sum_a \bar{\pi}(a|s') p(s'|s, a) \\
= (1 - \gamma) \beta(s) + \gamma \sum_{s'} (1 - \gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}[S_{t-1} = s'] p\bar{\pi}(s'|s') \\
= (1 - \gamma) \beta(s) + \gamma \sum_{s'} \rho^\bar{\pi}_\gamma(s') p\bar{\pi}(s'|s')
Proof: From Non-Stationary to Stationary Policies

For any non-stationary policy $\pi$ define a stationary policy $\pi$ as

$$\bar{\pi}(a|s') = \frac{\rho^\pi_y(s', a)}{\rho^\pi_y(s')}$$

$$\rho^\pi_y(s) = (1 - \gamma)\beta(s) + \gamma \sum_{s'} \sum_a (1 - \gamma) \sum_{t=1}^{\infty} \gamma^{t-1} P[S_{t-1} = s', A_{t-1} = a) p(s|s', a)$$

$$= (1 - \gamma)\beta(s) + \gamma \sum_{s'} \sum_a \rho^\pi_y(s', a) p(s|s', a)$$

$$= (1 - \gamma)\beta(s) + \gamma \sum_{s'} \sum_a \bar{\pi}(a|s') \rho^\pi_y(s') p(s|s', a)$$

$$= (1 - \gamma)\beta(s) + \gamma \sum_{s'} \rho^\pi_y(s') \sum_a \bar{\pi}(a|s') p(s|s', a)$$

$$= (1 - \gamma)\beta(s) + \gamma \sum_{s'} \rho^\pi_y(s') p^\pi(s|s')$$
Proof: From Non-Stationary to Stationary Policies

Moving to \textit{matrix} formulation

\[
\begin{align*}
[rho_{\bar{\pi}}]_s &= \rho_{\bar{\pi}}(s) \\
[P_{\bar{\pi}}]_{s,s'} &= p_{\bar{\pi}}(s'|s)
\end{align*}
\]

\[
\rho_{\bar{\pi}}(s) = (1 - \gamma)\beta(s) + \gamma \sum_{s'} \rho_{\bar{\pi}}(s') p_{\bar{\pi}}(s|s')
\]

\[\Rightarrow \rho_{\bar{\pi}} = (1 - \gamma)\beta + \gamma \rho_{\bar{\pi}} P_{\bar{\pi}}\]

\[\Rightarrow \rho_{\bar{\pi}} = (1 - \gamma)\beta (I - \gamma P_{\bar{\pi}})^{-1}\]
Proof: From Non-Stationary to Stationary Policies

Moving to matrix formulation

\[ \rho_\gamma^\pi (s) = (1 - \gamma) \beta (s) + \gamma \sum_{s'} \rho_\gamma^\pi (s') p^\pi (s|s') \]

\[ \Rightarrow \rho_\gamma^\pi = (1 - \gamma) \beta (I - \gamma P^\pi)^{-1} \]

\[ \Rightarrow \rho_\gamma^\pi = \rho_\gamma^\pi \]
The Optimization Problem

$$\max_{\pi} V^\pi(s_0)$$

$$= \max_{\pi} \mathbb{E} [r(s_0, d_0(a_0|s_0)) + \gamma r(s_1, d_1(a_1|s_0, s_1)) + \gamma^2 r(s_2, d_2(a_2|s_0, s_1, s_2)) + \ldots]$$

$$= \max_{\pi \in \Pi^{MRS}} \mathbb{E} [r(s_0, \pi(a_0|s_0)) + \gamma r(s_1, \pi(a_1|s_1)) + \gamma^2 r(s_2, \pi(a_2|s_2)) + \ldots]$$

_even if we restrict to deterministic policies, we still have $|A|^{|S|}$ policies to check._

_even Better than $\sum_t |A|^{|S|^t}$_
Recap

- Although quite general, Markov Decision Processes (MDPs) bake in numerous assumptions. Care should be taken when modeling a problem as an MDP.
- Similarly, care should be taken to select an appropriate type of policy and value function, depending on the use case.
- For well-conditioned infinite-horizon MDPs, stationary policies are as expressive as non-stationary history-dependent policies.
- Moreover, for discounted bounded-cost problems, there always exists an optimal deterministic policy.
Outline

1. MDPs

2. Partially observed problems
   a. State augmentation
   b. Imperfect state information
   c. State estimation, LQR and the separation principle
Partially observed problems

Strategies:

- **State augmentation**: add the missing information
- **Belief state**: Bayesian approach
- **Estimate** the missing information (e.g. Kalman filtering)
State Augmentation

- When assumptions of the basic problem (MDP) are violated, reformulate or **augment the state**.
  - e.g. disturbances are correlated, cost is nonadditive, etc.
- DP algorithm still applies, but the problem gets **bigger**.
Example: Time lags

- Consider:
  \[ s_{t+1} = f_t(s_t, s_{t-1}, a_t, \epsilon_t) \]

- Introduce additional state variable \( y_t = s_{t-1} \). New system takes the form:
  \[
  \begin{pmatrix}
  s_{t+1} \\
  y_{t+1}
  \end{pmatrix}
  =
  \begin{pmatrix}
  f_t(s_t, y_t, a_t, \epsilon_t) \\
  s_t
  \end{pmatrix}
  
- View \( \tilde{s}_t = (s_t, y_t) \) as the new state.

- DP algorithm for the reformulated problem:
  \[
  V_t(s_t, y_t) = \max_{a_t \in \mathcal{A}_t(s_t, y_t)} \mathbb{E}\{r_t(s_t, y_t, a_t, \epsilon_t) + V_{t+1}(f_t(s_t, y_{t-1}, a_t, \epsilon_t), s_{t+1})\}
  \]
Motivation: Diabetes Management

- What if the requisite state information is not accessible?
- Assume that a patient’s blood glucose level evolves each day as the following dynamic system
  \[ s_{t+1} = f(s_t, a_t, w_t) \]
- The action set may include: physical activity, measuring glucose, taking insulin etc.
- We never see the true blood glucose level \( s_t \) but instead a noisy measurement of it in case the patient does measure their level at time \( t \).

\[
o_t = \begin{cases} 
  s_t + \sigma(s_t)\xi_t & \text{if } \{\text{measure}\} \subset a_t \\
  \emptyset & \text{o.w.}
\end{cases}
\]
Problems With Imperfect State Information

- Consider a dynamic system that evolves according to
  \[ s_{t+1} = f(s_t, a_t, w_t) \]
  where the disturbances \( \{w_t\} \) are independent.
- At time \( t \), instead of the state \( s_t \), we observe
  \[ o_t = O_t(s_t, a_{t-1}, \xi_t) \]
  where \( \{\xi_t\} \) is an independent sequence.
- As before, the objective is to maximize the cumulative expected reward
  \[
  \max_{\pi \in \Pi} \mathbb{E}_{\{w_t\},\{\xi_t\}} \left[ \sum_{t=0}^{T-1} r_t(s_t, a_t, w_t) + r_T(s_T) \right]
  \]
  but now \( \pi_t \) is a map \( H_t \triangleq (o_0, \pi_0, ..., o_{t-1}, \pi_{t-1}, o_t) \mapsto \pi_t(H_t) \in \overline{\Pi}_t \).
Also called a partially observed Markov decision process (POMDP)

Typically restricting to discrete states and actions.

PSPACE-complete -- even harder than NP-Complete problems!

*PSPACE is the class of all decision problems solvable by a Turing machine in polynomial space with respect to the input size.
Approach

Leverage state augmentation to reduce imperfect information problem to perfect information problem.
History as State

- **Key insight:** The conditional probability of $o_t$ (given the history) is fully observed.
- Given transition function $f$, observation functions $O_t$, distributions of disturbances, there are known probability distributions $q_{s,a}$ such that

$$o_t \mid s_t, a_{t-1} \sim q_{s_t,a_t}(\cdot)$$

- Consider **history as state:** The state at time $t + 1$ is $H_t$ augmented with $a_t$ and $o_{t+1}$,

$$H_{t+1} = (H_t, a_t, o_{t+1})$$
History as State

- **Marginalize over states** for conditional probability of observation $o_{t+1}$:

$$\mathbb{P}[H_{t+1} = (H_t, a_t, o) | a_t = a, H_t] = \mathbb{P}(o_{t+1} = o | a_t = a, H_t) = \sum_s p_t(s)q_{s,a}(o)$$

where $p_t(s) = \mathbb{P}(s_t = s | H_t)$.

- And conditional probability of reward:

$$\tilde{r}_t(H_t, a) = \mathbb{E}[r_t(s_t, a, w_t) | H_t] = \sum_s p_t(s) \mathbb{E}[r_t(s, a, w_t)]$$

- Revised problem objective: (for simplicity, assume $r_T = 0$)

$$\max_{\pi} \mathbb{E}^{\pi} \left[ \sum_{t=0}^{T-1} r_t(s_t, a_t, w_t) \right] = \max_{\pi} \mathbb{E}^{\pi} \left[ \sum_{t=0}^{T-1} \mathbb{E}[r_t(s_t, a_t, w_t) | H_t] \right] = \max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[ \sum_{t=0}^{T-1} \tilde{r}_t(H_t, a_t) \right]$$

where $a_t = \pi_t(H_t) \in \Pi_t$.

- **Discuss**: Issues with this approach?
Posterior ("belief") as State

- **History is sufficient, but is it necessary?** We are ultimately interested in \( s_t \), not \( o_t \).
- **Key idea:** Maintain a sufficient summary of the history \( H_t \) to inform the probability of the next state \( s_{t+1} \).
- We define state and \( \tilde{r}(\cdot) \) as a function of our belief about the state \( s_t \) denoted as \( p_t(s) \).

\[
\tilde{r}_t(p_t, a) = \sum_s p_t(s) \mathbb{E}[r_t(s, a, w)]
\]

- The corresponding objective:

\[
\max \pi \mathbb{E}^\pi \left[ \sum_{t=0}^{T-1} r_t(s_t, a_t, w_t) \right] = \max \pi \mathbb{E}^\pi \left[ \sum_{t=0}^{T-1} \tilde{r}_t(p_t, a_t) \right]
\]

which is optimized over policies \( \pi = (a_0, ..., a_{T-1}) \) where \( a_t = \pi_t(p_t) \).

- Here, \( p_t \) is a posterior distribution and it evolves according to sequential Bayesian updating:

\[
p_{t+1}(s') = \mathbb{P}(s_{t+1} = s'|o_{t+1}, a_t, H_t) = \sum_s p_t(s) \mathbb{P}(s_{t+1} = s'|o_{t+1}, a_t, s_t = s)
\]

- **Issue:** the vector of beliefs can generally take on any value in the probability simplex \( \{p|p \geq 0, \sum_s p(s) = 1\} \). In general, computing the optimal policy for problems with continuous state vectors of moderate dimension is intractable.
Recall (L3): Linear quadratic control (stochastic)

Assumptions: deterministic, finite horizon, discrete time

Gaussian noise $\rightarrow$ Linear quadratic Gaussian (LQG) problem

$$s_{t+1} = f(s_t, a_t, \epsilon_t) = As_t + Ba_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \Sigma)$$

Revised optimization problem:

$$a = \min_{a_0, \ldots, a_{T-1}} V(s_0; a) = \mathbb{E} \left[ \sum_{t=0}^{T-1} s_t^T Q s_t + a_t R a_t + s_T^T Q_T s_T \right]$$

subject to $s_{t+1} = As_t + Ba_t + \epsilon_t$

Theorem (LQG)

The optimal cost-to-go and optimal control at time $t$ are given by:

$$V^*(s_t) = s_t^T P_t s_t + \Sigma_t$$

$$a^*_t = -K_t s_t$$

where

$$P_t = Q + K_t^T R K_t + (A - B K_t)^T P_{t+1} (A - B K_t), \quad P_T = Q_f$$

$$K_t = (R + B^T P_{t+1} B)^{-1} B^T P_{t+1} A, \quad \Sigma_{t-1} = Tr (\Sigma P_t) + \Sigma_t, \quad \Sigma_T = 0$$

$t \in \{0, \ldots, T - 1\}$

- Intuition (certainty equivalence): noise terms are independent of actions $\rightarrow$ optimal actions don’t change.
Consider the LQG problem (like before), where the system state evolves as
\[ x_{t+1} = Ax_t + Bu_t + w_t, \quad \forall t = \{0, \ldots, T - 1\} \]
Instead of the state \( x_t \), we observe a noisy measurement of it,
\[ y_t = Cx_t + \xi_t \]
where we assume \( \{w_t\}, \{\xi_t\} \) to be independent sequences (and also independent of \( x_0 \)).
As before, the objective is to minimize the total cost
\[
\min_{\pi} \mathbb{E}_\pi \left[ \sum_{t=0}^{T-1} (x_t^T Q x_t + u_t^T R u_t + x_T^T Q x_T) \right]
\]
over policies \( \pi = (\pi_0, \ldots, \pi_{T-1}) \) where \( u_t = \pi_t(H_t) \). (For simplicity, we let \( Q_T := Q \).)
Proposition (Separation principle)

The optimal policy of the LQ control with imperfect state information is
\[ \pi^* = (\pi_0^*, \ldots, \pi_{T-1}^*) \] where
\[ \pi_t^* (H_t) = -K_t \cdot \mathbb{E}[x_t | H_t] \]

The matrices $K, P$ can be computed recursively using the same formulas as before.

- The optimal policy for LQ control with imperfect state information is very similar to that of the perfect state case. The only difference being that instead of acting on the state $x_t$, we now plug in our best estimate of the state $\mathbb{E}[x_t | H_t]$.
- Due to this remarkable fact, one can separate the problem of designing an optimal feedback controller (designing $K_t$) and the optimal state estimation procedure.
- In the important special case where the disturbances $\{w_t\}, \{\xi_t\}$ and the initial state $x_0$ are independent Gaussian vectors, a convenient implementation of computing the conditional mean is possible by means of the Kalman filtering algorithm, which is developed in DPOC Appendix E.
Warmup (1-step)

- Why might the conditional mean be good in LQ control?
- Optimization problem: quadratic estimation loss and a quadratic penalty
  \[ \min_u \mathbb{E}_x [(x - u)^T Q (x - u) + u^T Ru] \]

  where \( Q, R > 0 \).
- Minimizer is a linear function of the mean: \( u^* = (Q + R)^{-1} Q \mathbb{E}[x] \).
- When \( R = 0 \), the optimal objective value penalizes the variance of estimation error

  \[ \mathbb{E}[(x - \mathbb{E}[x])^T Q (x - \mathbb{E}[x])] \]

  Otherwise, the objective value separates into the sum of two terms: one of which depends on the variance of \( x \) and one which depends on the mean, which influences the energy cost \( u^T Ru \).
State estimation error is independent of control

Lemma

For every $t$, the estimation error, $x_t - \mathbb{E}[x_t | H_t]$, does not depend on $u_1, \ldots, u_{t-1}$

- To prove Proposition, we first show the Lemma, which states that the state estimation error, $x_t - \mathbb{E}[x_t | H_t]$ is independent of the control choice.
- This is due to the linearity of both the system and the measurement equation. In particular, $x_t$ and $\mathbb{E}[x_t | H_t]$ contain the same linear terms in $(u_0, \ldots, u_{t-1})$, which cancel each other out.
Proof: Lemma

- Since there is no control when $t = 0$, the claim is obviously true.
- For $t > 0$, we can write $x_t$ recursively as follows
  $$x_t = Ax_{t-1} + Bu_{t-1} + w_{t-1}$$
  $$= A(Ax_{t-2} + Bu_{t-2} + w_{t-2}) + Bu_{t-1} + w_{t-1}$$
  $$= \ldots$$
  $$= A^tx_0 + \sum_{i=0}^{t-1} A^i Bu_i + \sum_{i=0}^{t-1} A^{t-1-i} w_i$$

- Then
  $$x_t - \mathbb{E}[x_t|H_t] = A^t(x_0 - \mathbb{E}[x_0|H_t]) - \sum_{i=0}^{t-1} A^{t-1-i}(w_i - \mathbb{E}[w_i|H_t])$$
  which is independent of the control sequence $\{u_1, \ldots, u_{t-1}\}$. 
Proof: Separation Principle

- For $P_T = Q$ and $\bar{K}_T = 0$, write the cost-to-go function as the mean cost plus the estimation variance (which does not depend on the controls)
  $$V_T(H_T) = \mathbb{E}[x_T^T P_T x_T | H_T] + \mathbb{E}[e_T^T K_T e_T | H_T]$$
  where $e_T := x_T - \mathbb{E}[x_T | H_T]$.

- For time $T - 1$:
  $$V_{T-1}(H_{T-1}) = \min_u \{ l(H_{T-1}, u) \}$$
  where
  $$l(H_{T-1}, u) = u^T R u + \mathbb{E}[x_{T-1}^T Q x_{T-1} | H_{T-1}] + V_T((H_{T-1}, u))$$
  $$= u^T R u + \mathbb{E}[x_{T-1}^T Q x_{T-1} | H_{T-1}]$$
  $$+ \mathbb{E}[(A x_{T-1} + B u_{T-1} + w_{T-1})^T P_T (A x_{T-1} + B u_{T-1} + w_{T-1}) | H_{T-1}, u_{T-1} = u]$$
  $$+ \mathbb{E}[e_{T}^T K_{T} e_{T} | H_{T-1}]$$

- The cost-to-go at the previous stage is the instantaneous cost + cost-to-go, where the next state is given by linear dynamics.

- Differentiating with respect to $u$ we get
  $$\pi^*(H_{T-1}) = -K_{T-1} \mathbb{E}[x_{T-1} | H_{T-1}]$$
  where
  $$K_{T-1} = (R + B^T P_T B)^{-1} B^T P_T A$$
Proof: Separation Principle

- Plugging the linear policy back into the quadratic function leads to
  \[ V_{T-1}(H_{T-1}) = l(H_{T-1}, -K_{T-1}\mathbb{E}[x_{T-1}|H_{T-1}]) \]
  \[ = \mathbb{E}[w_{T-1}^TQw_{T-1}] + \mathbb{E}[x_{T-1}^T(Q + A^TP_TA)x_{T-1}|H_{T-1}] \]
  \[ -\mathbb{E}[x_{T-1}|H_{T-1}]^T\bar{K}_{T-1}\mathbb{E}[x_{T-1}|H_{T-1}] + \mathbb{E}[e_{T}\bar{K}Te_T|H_{T-1}] \]
  where \( \bar{K}_{T-1} := A^TP_TB K_{T-1} = A^TP_TB (R + B^TP_TB)^{-1}BP_TA \).

- Notice that we can write the last term as
  \[ \mathbb{E}[x_{T-1}|H_{T-1}]^T\bar{K}_{T-1}\mathbb{E}[x_{T-1}|H_{T-1}] = \mathbb{E}[x_{T-1}^T\bar{K}_{T-1}x_{T-1}|H_{T-1}] - \mathbb{E}[e_{T-1}^T\bar{K}_{T-1}e_{T-1}|H_{T-1}] \]
  - This is a generalization of \( Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \)

- Plugging this back, we have
  \[ V_{T-1}(H_{T-1}) = \mathbb{E}[x_{T-1}^TP_{T-1}x_{T-1}|H_{T-1}] \]
  \[ +\mathbb{E}[e_{T-1}^T\bar{K}_{T-1}e_{T-1}|H_{T-1}] + \mathbb{E}[e_{T}\bar{K}Te_T|H_{T-1}] + C_{T-1} \]
  where \( P_{T-1} := Q + P_TA - \bar{K}_{T-1} = Q + A^TP_A - A^TP_TB (R + B^TP_TB)^{-1}BP_TA \)

- Thus, the cost-to-go function is a quadratic function of state plus terms that are not affected by
  the control decision (via lemma).

- Recurse, and we get the desired result.
References

1. DPOC vol 1, §1.4, §4.1-4.2

2. DPOC vol 2, §1.1.4


4. Some material adapted from:
   - Alessandro Lazaric (FAIR/INRIA)
   - Daniel Russo (Columbia)
   - Dimitrios Katselis, R. Srikant (UIUC)