# LINMA2725 Stochastic Optimal Control and Reinforcement Learning Part III

## Course 1: Stochastic Systems and Stochastic Optimal Control

Guillaume Berger

November 4, 2024

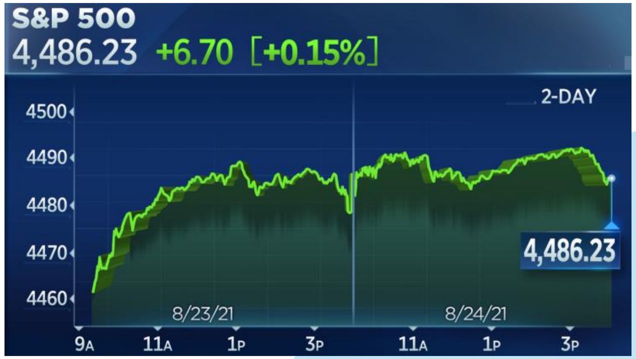
Reference: [1], Chapters 6 and 7.

Any questions or feedback are welcome.

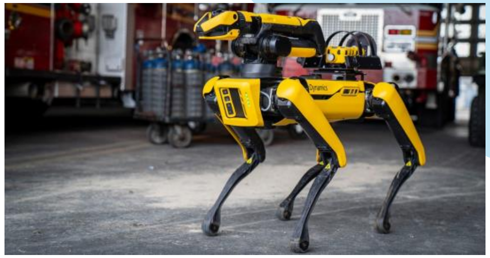
# Stochastic Optimal Control and Reinforcement Learning

## Part III: Stochastic Systems

Guillaume Berger



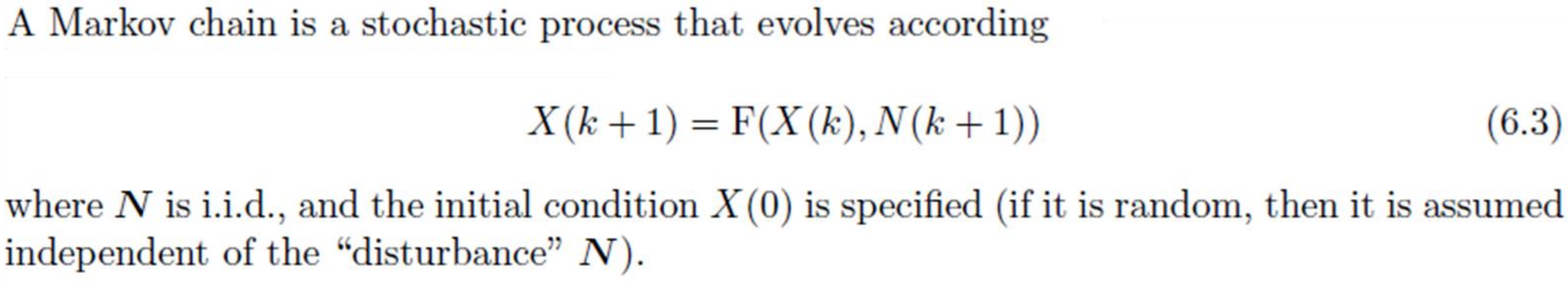
# Stochasticity
# in
# Systems and Control



# Table of content

- Stochastic systems and stochastic control (1 course)
  - Systems definition, ergodicity, …
  - Optimal cost, Poisson and Bellman equations, …
  - Value iterations, Policy improvement, LP
  - Fluid model
  - Sensitivity and parameterized policies
- Learning techniques for stochastic control (1-2 courses)
- Online learning techniques for stochastic control (1-2 courses)

# Stochastic system



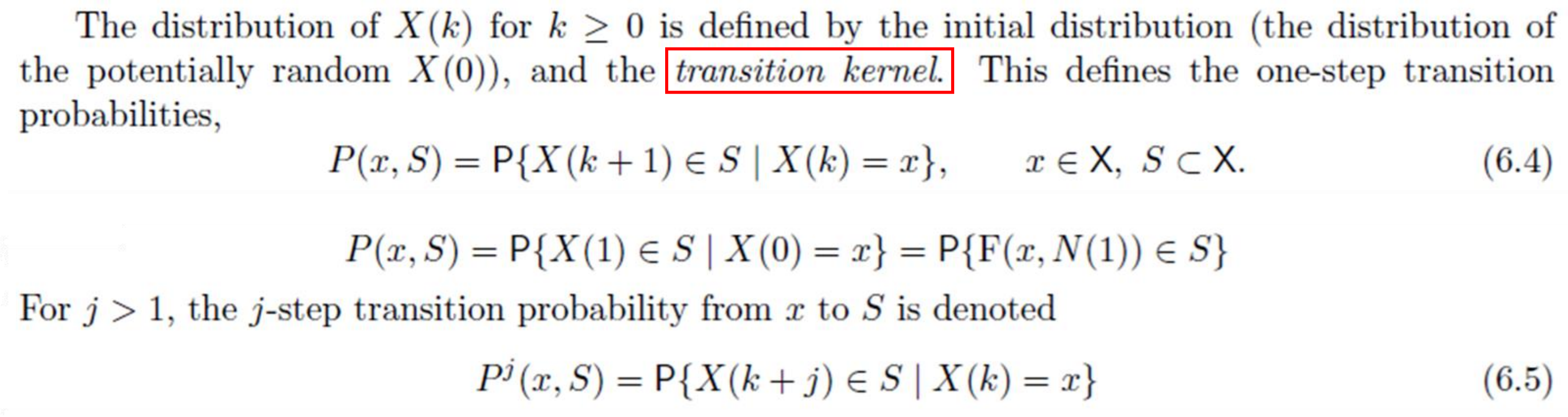
A Markov chain is a stochastic process that evolves according

$$X(k+1) = \mathrm{F}(X(k), N(k+1)) \tag{6.3}$$

where $\boldsymbol{N}$ is i.i.d., and the initial condition $X(0)$ is specified (if it is random, then it is assumed independent of the "disturbance" $\boldsymbol{N}$).

Andreï Markov
1856-1922

# Transition kernel



The distribution of $X(k)$ for $k \geq 0$ is defined by the initial distribution (the distribution of the potentially random $X(0)$), and the *transition kernel.* This defines the one-step transition probabilities,

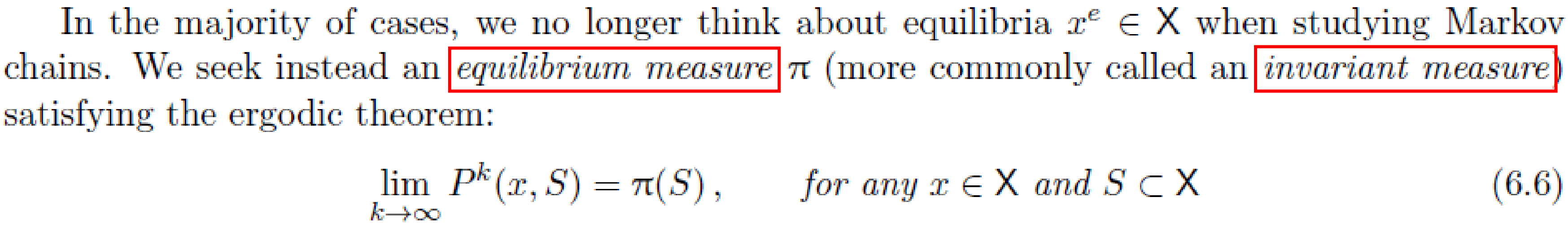
$$P(x,S) = \mathsf{P}\{X(k+1) \in S \mid X(k) = x\}, \qquad x \in \mathsf{X},\ S \subset \mathsf{X}. \tag{6.4}$$

$$P(x,S) = \mathsf{P}\{X(1) \in S \mid X(0) = x\} = \mathsf{P}\{\mathrm{F}(x, N(1)) \in S\}$$

For $j > 1$, the $j$-step transition probability from $x$ to $S$ is denoted

$$P^j(x,S) = \mathsf{P}\{X(k+j) \in S \mid X(k) = x\} \tag{6.5}$$

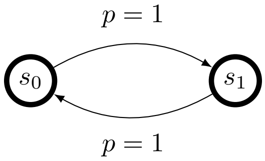
# Invariant measure



In the majority of cases, we no longer think about equilibria $x^e \in \mathsf{X}$ when studying Markov chains. We seek instead an *equilibrium measure* $\pi$ (more commonly called an *invariant measure*) satisfying the ergodic theorem:

$$\lim_{k\to\infty} P^k(x,S) = \pi(S)\,, \qquad \textit{for any } x \in \mathsf{X} \textit{ and } S \subset \mathsf{X} \tag{6.6}$$

An equilibrium measure may not always exist. Example: the system



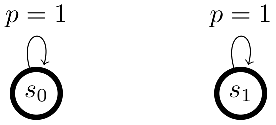
has no invariant measure.

**Definition 1.** A *stationary measure* for the kernel $P$ is a measure $\pi$ satisfying $\int P(x,S)\,\pi(\mathrm{d}x) = \pi(S)$.

An invariant measure is stationary, but the converse is not always true. Example: the system above has stationary measure $\pi(s_0) = \pi(s_1) = \frac{1}{2}$ but no invariant measure.

A stationary measure may not always exist. Example: the random walk (aka. Brownian motion) has no stationary measure.

A stationary measure, if it exists, may not be unique. Example: the system

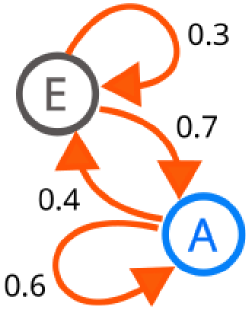


admits any probability distribution on $X = \{s_0, s_1\}$ as stationary measure.

# Finite systems

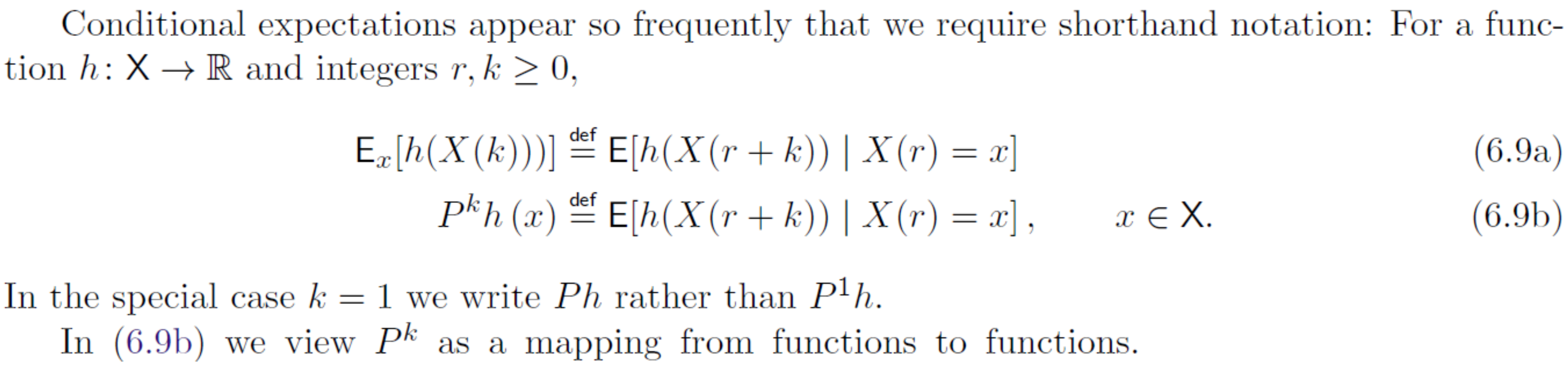
If X is finite, of size $m$, then $P$ is interpreted as an $m \times m$ matrix. In this case $P(x_0, x_1)$ is the probability of moving from $x_0$ to $x_1$ in one time-step. The conditional expectation is expressed as a sum,

$$\mathsf{E}[h(X(k+1)) \mid X(k) = x] = \sum_{x_1 \in \mathsf{X}} P(x, x_1) h(x_1)$$



$$\begin{array}{ccc} \curvearrowright & A & E \\ A & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 0{,}6 & 0{,}4 \\ 0{,}7 & 0{,}3 \end{bmatrix}$}} \\ E & & \end{array}$$

# Conditional expectation



Conditional expectations appear so frequently that we require shorthand notation: For a function $h\colon \mathsf{X} \to \mathbb{R}$ and integers $r, k \ge 0$,

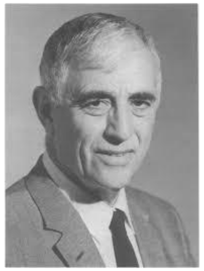
$$\mathsf{E}_x[h(X(k))] \overset{\text{def}}{=} \mathsf{E}[h(X(r+k)) \mid X(r) = x] \tag{6.9a}$$

$$P^k h\,(x) \overset{\text{def}}{=} \mathsf{E}[h(X(r+k)) \mid X(r) = x]\,, \qquad x \in \mathsf{X}. \tag{6.9b}$$

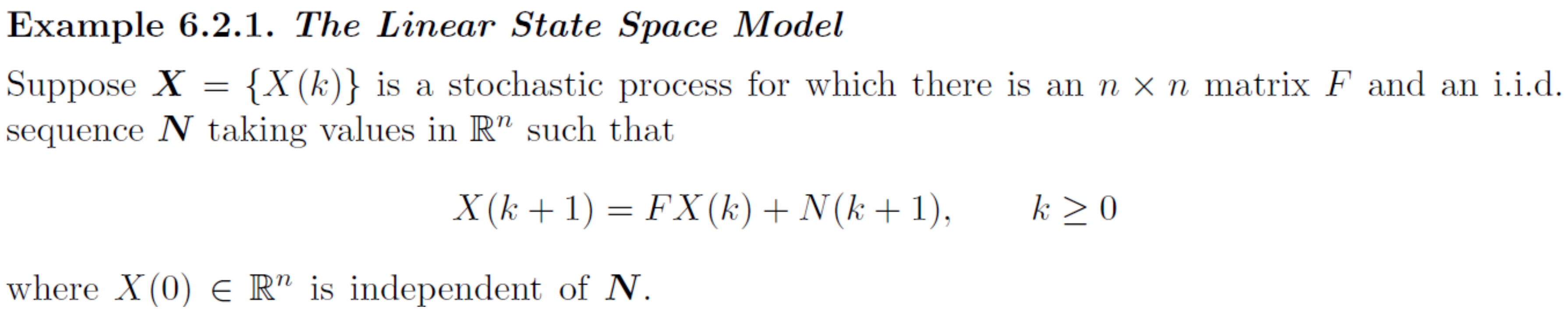
In the special case $k = 1$ we write $Ph$ rather than $P^1 h$.

In (6.9b) we view $P^k$ as a mapping from functions to functions.

Bernard O. Koopman
1900-1981



# Example 1: Linear model



**Example 6.2.1.** ***The Linear State Space Model***

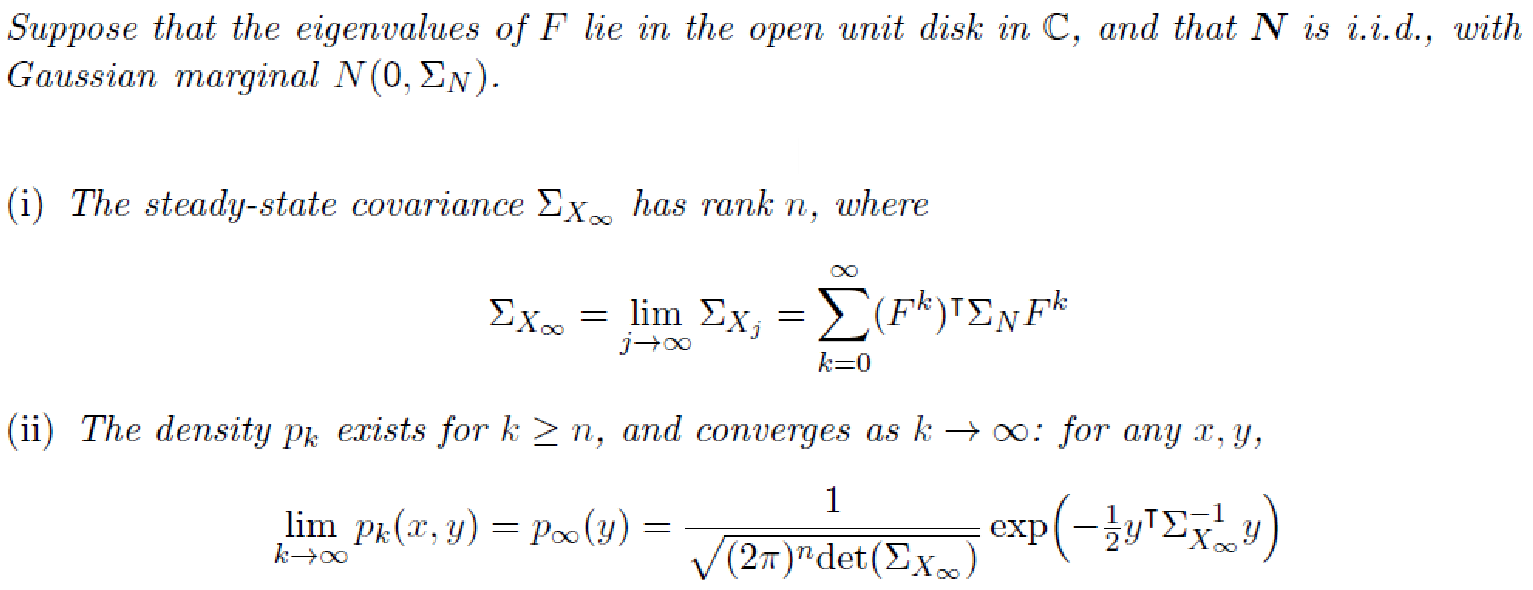
Suppose $\boldsymbol{X} = \{X(k)\}$ is a stochastic process for which there is an $n \times n$ matrix $F$ and an i.i.d. sequence $\boldsymbol{N}$ taking values in $\mathbb{R}^n$ such that

$$X(k+1) = FX(k) + N(k+1), \qquad k \geq 0$$

where $X(0) \in \mathbb{R}^n$ is independent of $\boldsymbol{N}$.

# Example 1: Linear model

(continued)



*Suppose that the eigenvalues of $F$ lie in the open unit disk in $\mathbb{C}$, and that $\mathbf{N}$ is i.i.d., with Gaussian marginal $N(0, \Sigma_N)$.*

(i) *The steady-state covariance $\Sigma_{X_\infty}$ has rank $n$, where*

$$\Sigma_{X_\infty} = \lim_{j\to\infty} \Sigma_{X_j} = \sum_{k=0}^{\infty} (F^k)^\intercal \Sigma_N F^k$$

(ii) *The density $p_k$ exists for $k \geq n$, and converges as $k \to \infty$: for any $x, y$,*

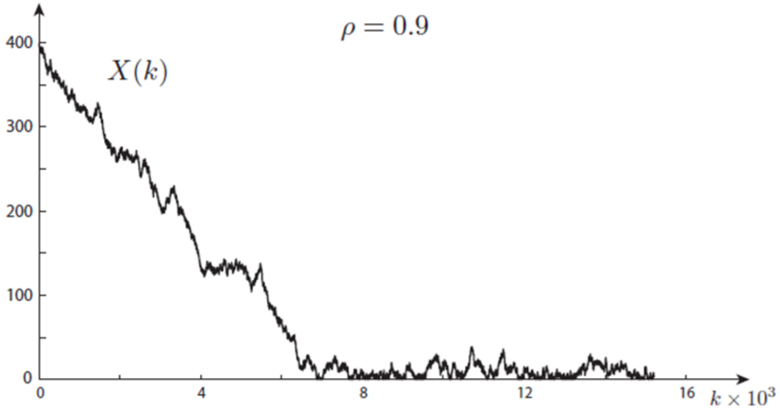
$$\lim_{k\to\infty} p_k(x,y) = p_\infty(y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_{X_\infty})}} \exp\Big(-\tfrac{1}{2} y^\intercal \Sigma_{X_\infty}^{-1} y\Big)$$

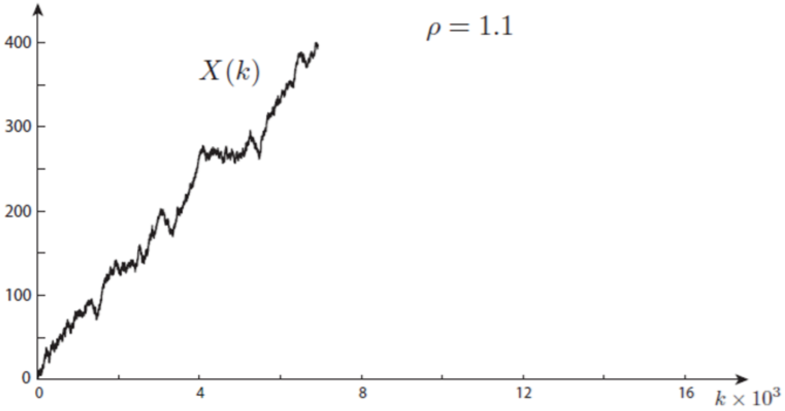
# Example 2: Queue

The transition function for the *M/M/1 queue* is defined as

$$\mathsf{P}(X(k+1) = y \mid X(k) = x) = P(x,y) = \begin{cases} \alpha & \text{if } y = x+1 \\ \mu & \text{if } y = (x-1)_+, \end{cases} \tag{6.14}$$

where $\alpha$ denotes the arrival rate to the queue, $\mu$ is the service rate, and these parameters are normalized so that $\alpha + \mu = 1$.



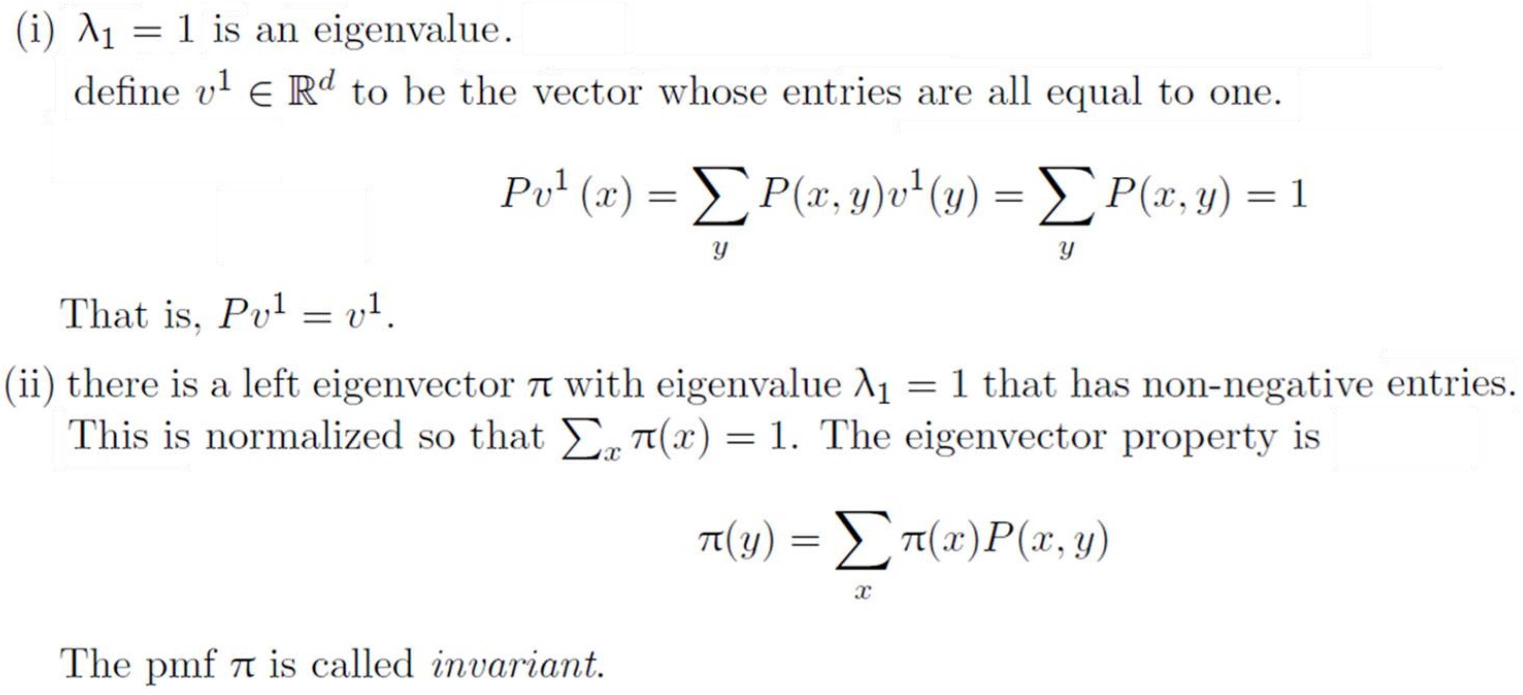


The parameter $\rho \overset{\mathsf{def}}{=} \alpha/\mu$ is known as the *load* for the queue. If $\rho < 1$ then the arrival rate is strictly less than the service rate. In this case the process is ergodic: there is a pmf $\pi$ on the non-negative integers such that for any initial queue length $X(0) = x$, and any integer $m \geq 0$,

$$\lim_{k\to\infty} \mathsf{P}_x\{X(k) = m\} = \pi(m)$$

The invariant pmf is geometric with parameter $\rho$, so that $\pi(m) = (1-\rho)\rho^m$.

# Finite systems: spectrum and ergodicity



(i) $\lambda_1 = 1$ is an eigenvalue.

define $v^1 \in \mathbb{R}^d$ to be the vector whose entries are all equal to one.

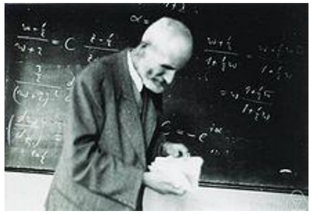
$$Pv^1\,(x) = \sum_y P(x,y)v^1(y) = \sum_y P(x,y) = 1$$

That is, $Pv^1 = v^1$.

(ii) there is a left eigenvector $\pi$ with eigenvalue $\lambda_1 = 1$ that has non-negative entries. This is normalized so that $\sum_x \pi(x) = 1$. The eigenvector property is

$$\pi(y) = \sum_x \pi(x)P(x,y)$$

The pmf $\pi$ is called *invariant.*



Oskar Perron
1880-1975

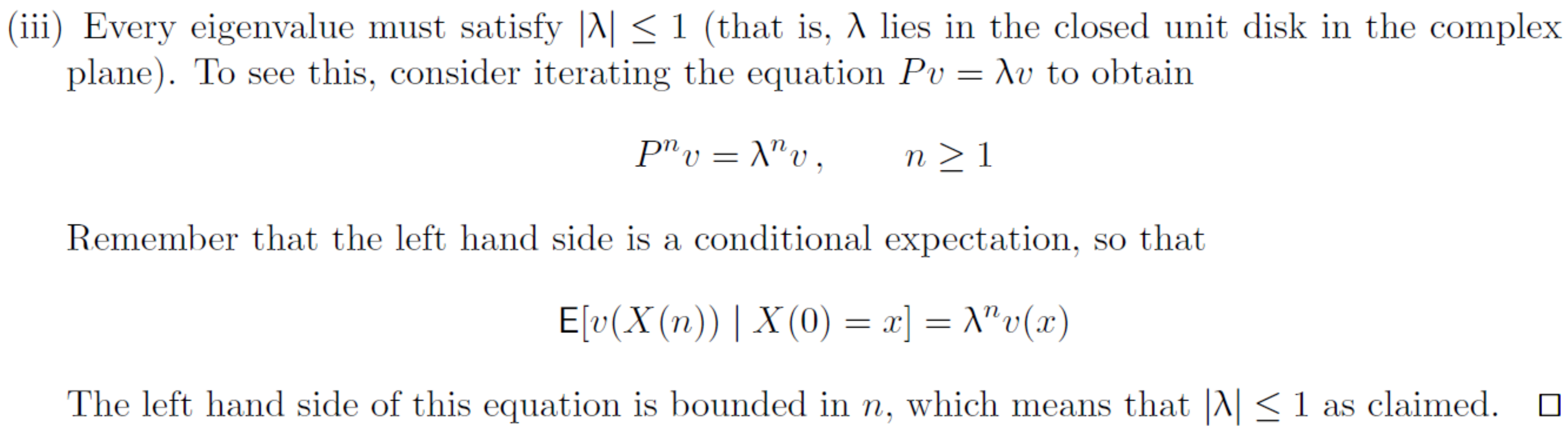
Georg F. Frobenius
1849-1917

Point (ii) says that a finite stochastic system always admits a stationary measure $\pi$.

*Proof of (ii).* This can be obtained, e.g., by Brouwer fixed-point theorem since $P^\top \Delta \subseteq \Delta$, where $\Delta = \{\pi \in \mathbb{R}^m : \pi \geq 0,\ \mathbf{1}^\top \pi = 1\}$ is the set of probability distributions on $\mathsf{X}$. Indeed, if $\pi \in \Delta$, then $P^\top \pi \geq 0$ and $\mathbf{1}^\top P^\top \pi = \mathbf{1}^\top \pi = 1$, so that $P^\top \pi \in \Delta$. □

# Finite systems: spectrum and ergodicity

(continued)



(iii) Every eigenvalue must satisfy $|\lambda| \leq 1$ (that is, $\lambda$ lies in the closed unit disk in the complex plane). To see this, consider iterating the equation $Pv = \lambda v$ to obtain

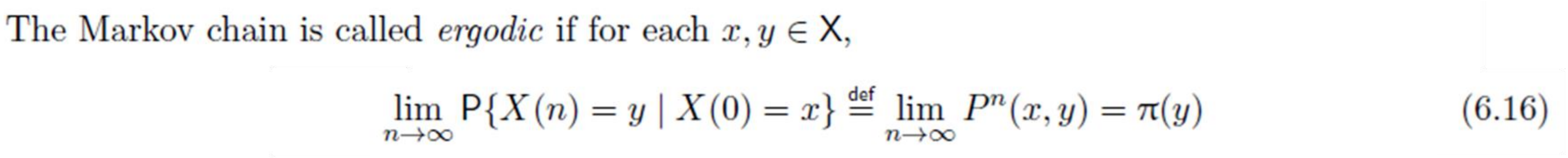
$$P^n v = \lambda^n v\,, \qquad n \geq 1$$

Remember that the left hand side is a conditional expectation, so that

$$\mathsf{E}[v(X(n)) \mid X(0) = x] = \lambda^n v(x)$$

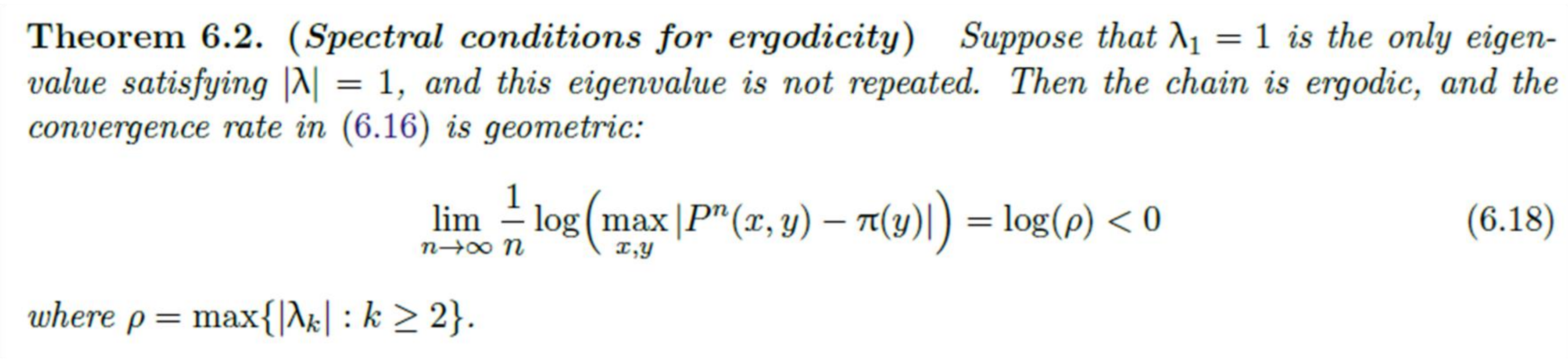
The left hand side of this equation is bounded in $n$, which means that $|\lambda| \leq 1$ as claimed. □

# Finite systems: spectrum and ergodicity



The Markov chain is called *ergodic* if for each $x, y \in \mathsf{X}$,

$$\lim_{n\to\infty} \mathsf{P}\{X(n) = y \mid X(0) = x\} \overset{\text{def}}{=} \lim_{n\to\infty} P^n(x,y) = \pi(y) \tag{6.16}$$



**Theorem 6.2.** **(*Spectral conditions for ergodicity*)** *Suppose that $\lambda_1 = 1$ is the only eigenvalue satisfying $|\lambda| = 1$, and this eigenvalue is not repeated. Then the chain is ergodic, and the convergence rate in (6.16) is geometric:*

$$\lim_{n\to\infty} \frac{1}{n} \log\Bigl(\max_{x,y} |P^n(x,y) - \pi(y)|\Bigr) = \log(\rho) < 0 \tag{6.18}$$

*where $\rho = \max\{|\lambda_k| : k \geq 2\}$.*

*Proof.* The first step is to consider a modified matrix $\widetilde{P}$ defined by

$$\widetilde{P}(x,y) = P(x,y) - \pi(y)\,, \qquad x,y \in \mathsf{X}. \tag{6.19}$$

This can be expressed $\widetilde{P} = P - \mathbf{1} \otimes \pi$, where $\mathbf{1} = v^1$ is a column vector of ones, $\pi$ is the invariant pmf, and "$\otimes$" is an outer product. It can be shown by induction that

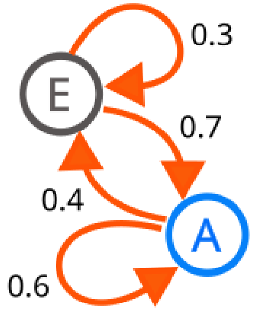
$$\widetilde{P}^n = P^n - \mathbf{1} \otimes \pi \tag{6.20}$$

That is, for each $x, y$,

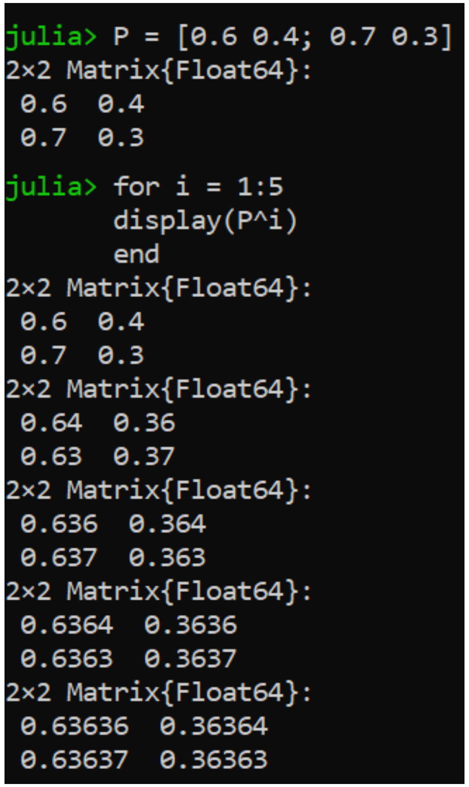
$$\widetilde{P}^n(x,y) = P^n(x,y) - \pi(y)$$

With a bit more effort it can be shown that $\lambda(\widetilde{P}) = \{0, \lambda_2, \dots, \lambda_m\}$. That is, all of the eigenvalues of $\widetilde{P}$ coincide with those of $P$, except the first eigenvalue which is moved to the origin. A bit of linear algebra completes the proof of (6.18). □

# Example: Ergodic Markov chain



$$\begin{array}{c|cc} \curvearrowright & A & E \\ \hline A & 0{,}6 & 0{,}4 \\ E & 0{,}7 & 0{,}3 \end{array}$$



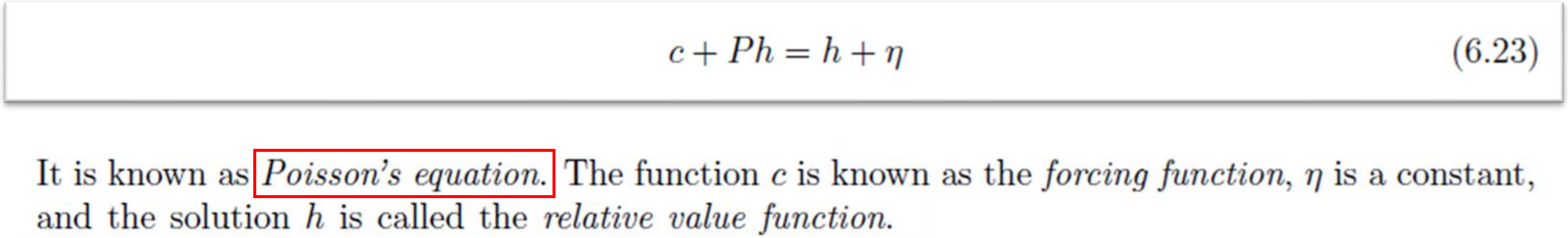
```
julia> P = [0.6 0.4; 0.7 0.3]
2×2 Matrix{Float64}:
 0.6  0.4
 0.7  0.3

julia> for i = 1:5
       display(P^i)
       end
2×2 Matrix{Float64}:
 0.6  0.4
 0.7  0.3
2×2 Matrix{Float64}:
 0.64  0.36
 0.63  0.37
2×2 Matrix{Float64}:
 0.636  0.364
 0.637  0.363
2×2 Matrix{Float64}:
 0.6364  0.3636
 0.6363  0.3637
2×2 Matrix{Float64}:
 0.63636  0.36364
 0.63637  0.36363
```

$$\pi = \begin{array}{c} \begin{array}{cc} A & E \end{array} \\ \left[\frac{63}{99}, \frac{36}{99}\right] \end{array}$$

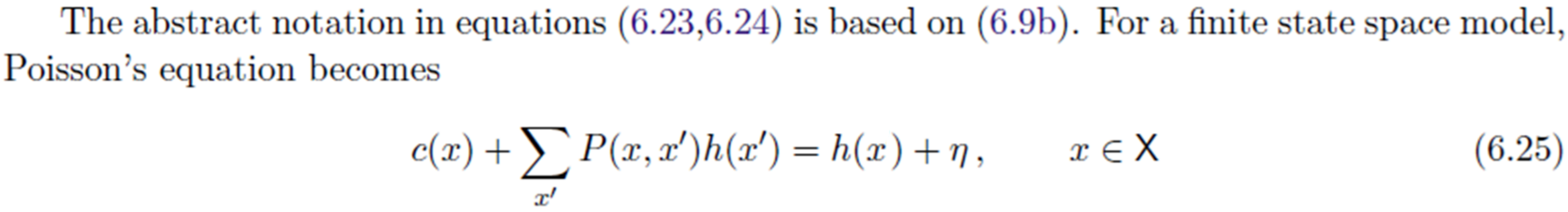
# Poisson equation

Siméon Denis Poisson
1781-1840



$$c + Ph = h + \eta \tag{6.23}$$

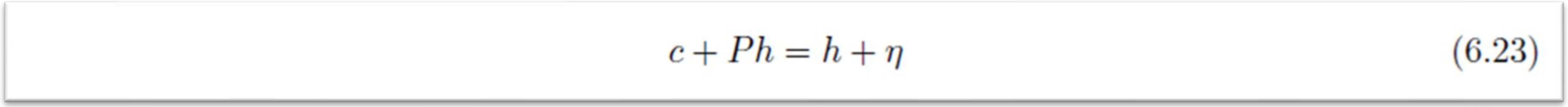
It is known as *Poisson's equation.* The function $c$ is known as the *forcing function*, $\eta$ is a constant, and the solution $h$ is called the *relative value function.*



The abstract notation in equations (6.23,6.24) is based on (6.9b). For a finite state space model, Poisson's equation becomes

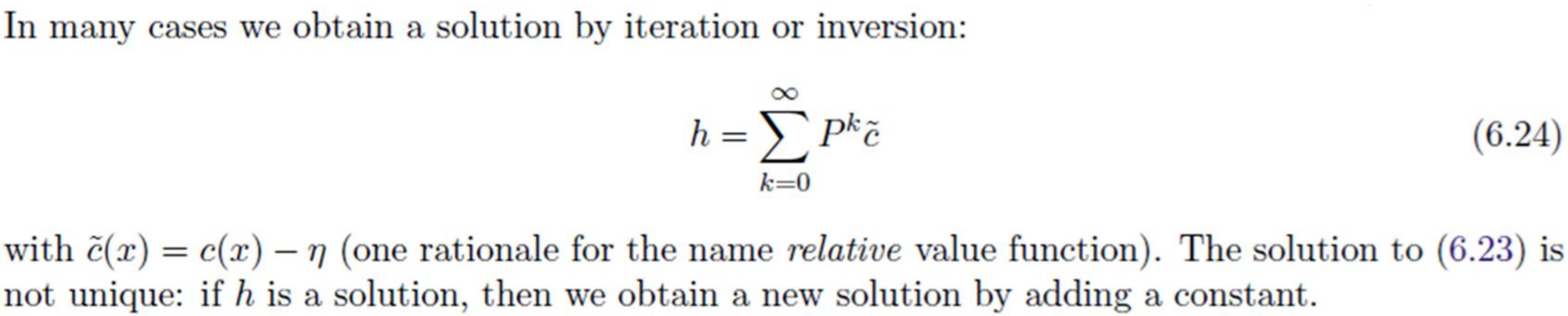
$$c(x) + \sum_{x'} P(x,x')h(x') = h(x) + \eta\,, \qquad x \in \mathsf{X} \tag{6.25}$$

# Poisson equation: properties of solutions



$$c + Ph = h + \eta \tag{6.23}$$

If $\pi$ is an invariant measure, then $\eta = \pi(c) \overset{\text{def}}{=} \int c(x)\pi(dx)$

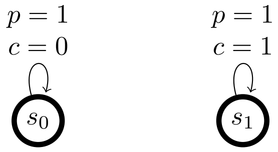


In many cases we obtain a solution by iteration or inversion:

$$h = \sum_{k=0}^{\infty} P^k \tilde{c} \tag{6.24}$$

with $\tilde{c}(x) = c(x) - \eta$ (one rationale for the name *relative* value function). The solution to (6.23) is not unique: if $h$ is a solution, then we obtain a new solution by adding a constant.

A solution to Poisson equation may not always exist. Example: the system

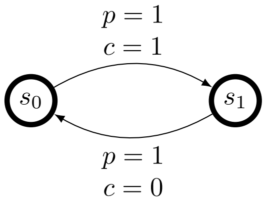


admits no solution to Poisson equation.

If a solution exists and the system admits an invariant measure (or more generally a stationary measure) $\pi$, then $\eta = \int c(x)\,\pi(\mathrm{d}x)$. Indeed,

$$\int c(x) + Ph(x)\,\pi(\mathrm{d}x) = \int h(x) + \eta\,\pi(\mathrm{d}x) \qquad \Longrightarrow \qquad \int c(x)\,\pi(\mathrm{d}x) = \int \eta\,\pi(\mathrm{d}x) = \eta.$$

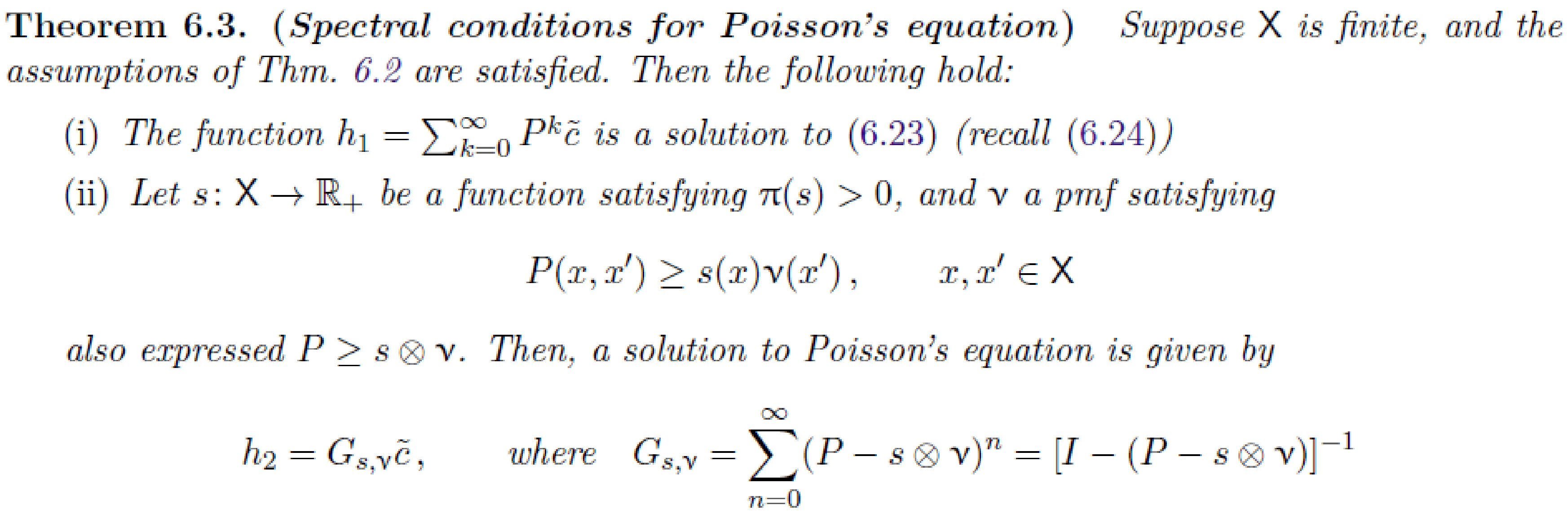
If the sum in (6.24) converges, then it is a solution to Poisson equation. However, it may not always converge. Example: for the system



,

it does not converge.

# Poisson equation: finite systems



**Theorem 6.3.** (*Spectral conditions for Poisson's equation*) *Suppose* $\mathsf{X}$ *is finite, and the assumptions of Thm. 6.2 are satisfied. Then the following hold:*

(i) *The function* $h_1 = \sum_{k=0}^{\infty} P^k \tilde{c}$ *is a solution to* (6.23) *(recall* (6.24)*)*

(ii) *Let* $s\colon \mathsf{X} \to \mathbb{R}_+$ *be a function satisfying* $\pi(s) > 0$, *and* $\nu$ *a pmf satisfying*

$$P(x, x') \geq s(x)\nu(x'), \qquad x, x' \in \mathsf{X}$$

*also expressed* $P \geq s \otimes \nu$. *Then, a solution to Poisson's equation is given by*

$$h_2 = G_{s,\nu}\tilde{c}, \qquad \textit{where} \quad G_{s,\nu} = \sum_{n=0}^{\infty} (P - s \otimes \nu)^n = [I - (P - s \otimes \nu)]^{-1}$$

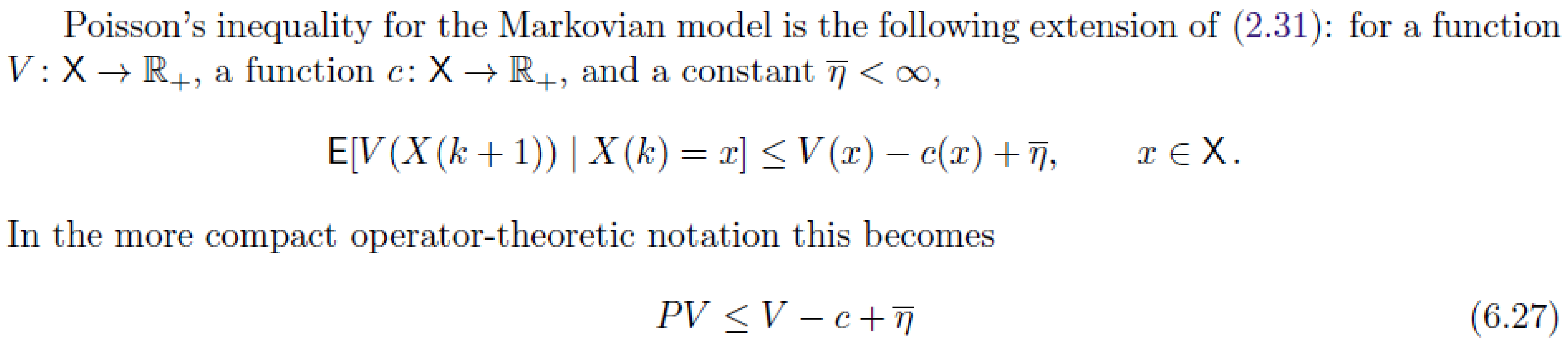
*Proof of (i).* Let $\tilde{P} = P - \mathbf{1}^\top \pi$, where $\pi$ is the invariant measure of $P$. It holds that $\tilde{P}^k \tilde{c} = (P^k - \mathbf{1}^\top \pi)\tilde{c}$ and $P^k \tilde{c}$. Since the eigenvalues of $\tilde{P}$ lie in the interior of the unit disk (show this), the sum converges. □

*Proof of (ii).* Let $\tilde{P} = P - s\nu^\top$. First, we show that the eigenvalues of $\tilde{P}$ lie in the interior of the unit disk. Let $\rho$ be the spectral radius of $\tilde{P}$. Since, $\tilde{P} \geq 0$, $\rho$ is an eigenvalue and there is a nonnegative associated eigenvector $v$. Hence, $Pv - s\nu^\top v = \rho v$. This implies that $\pi^\top Pv - (\pi^\top s)(\nu^\top v) = \rho \pi^\top v$. Hence, $-(\pi^\top s)(\nu^\top v) = (\rho - 1)\pi^\top v$. Assume that $\rho \geq 1$. This implies that $\nu^\top v = 0$, since $\pi^\top s > 0$ by assumption on $s$. It follows that $Gv = Pv = \rho v$. Hence, $\rho$ is an eigenvalue of $P$. Hence, $\rho = 1$ and $v$ is an eigenvector of $P$ with eigenvalue 1. This implies that $v$ is a positive multiple of $\mathbf{1}$. This is a contradiction with $\nu^\top v = 0$. Hence, $\rho < 1$.

The above shows that the sum of $\tilde{P}^k$ converges to $[I - \tilde{P}]^{-1}$. Finally, we show that $h = G_{s,\nu}\tilde{c}$ provides a solution to Poisson equation. Indeed, it holds that $[I - \tilde{P}]h = h - Ph + s\nu^\top h = c - \eta\mathbf{1}$. Hence, by left-multiplying by $\pi^\top$ on both sides, we get $(\pi^\top s)(\nu^\top h) = 0$. This implies $\nu^\top h = 0$, so that $h - Ph = c - \eta\mathbf{1}$, concluding the proof. □

*Remark* 1. The proof of (ii) only uses the existence of a *unique* stationary measure, not the existence of an invariant measure.

# Lyapunov functions

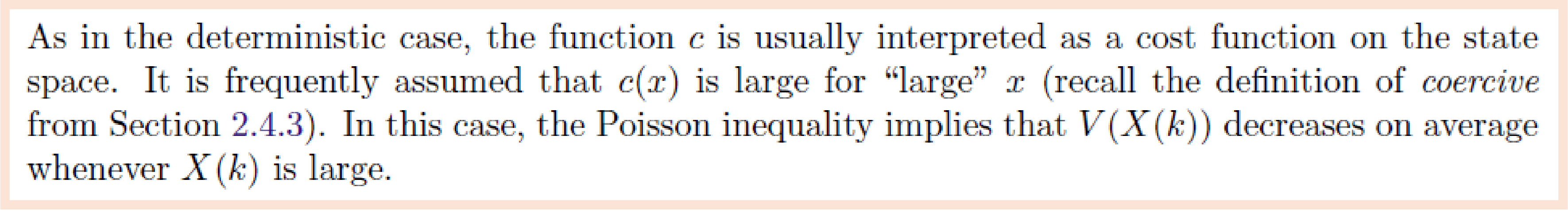


Poisson's inequality for the Markovian model is the following extension of (2.31): for a function $V : \mathsf{X} \to \mathbb{R}_+$, a function $c : \mathsf{X} \to \mathbb{R}_+$, and a constant $\overline{\eta} < \infty$,

$$\mathsf{E}[V(X(k+1)) \mid X(k) = x] \le V(x) - c(x) + \overline{\eta}, \qquad x \in \mathsf{X}.$$

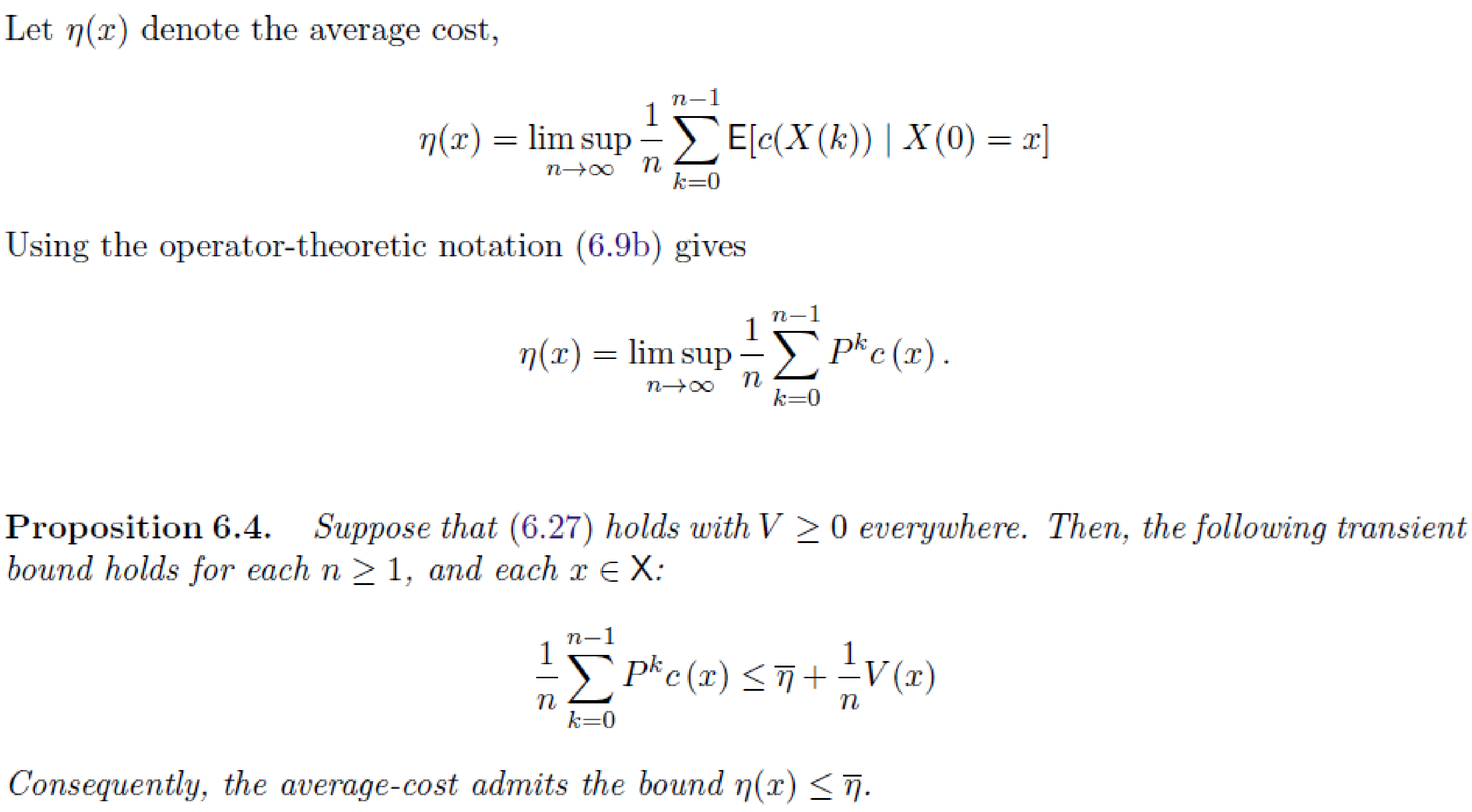
In the more compact operator-theoretic notation this becomes

$$PV \le V - c + \overline{\eta} \tag{6.27}$$



As in the deterministic case, the function $c$ is usually interpreted as a cost function on the state space. It is frequently assumed that $c(x)$ is large for "large" $x$ (recall the definition of *coercive* from Section 2.4.3). In this case, the Poisson inequality implies that $V(X(k))$ decreases on average whenever $X(k)$ is large.

# Average cost



Let $\eta(x)$ denote the average cost,

$$\eta(x) = \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathsf{E}[c(X(k)) \mid X(0) = x]$$

Using the operator-theoretic notation (6.9b) gives

$$\eta(x) = \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k c\,(x)\,.$$

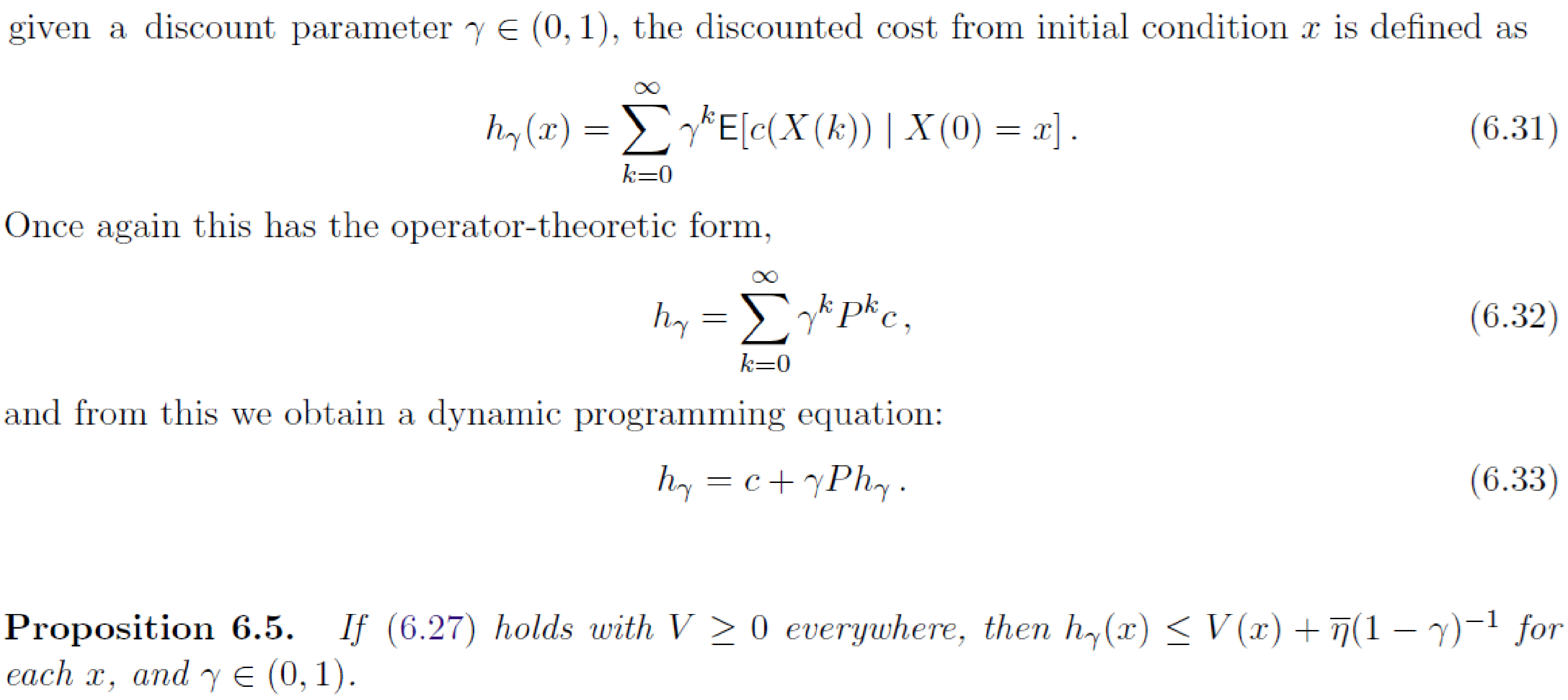
**Proposition 6.4.** *Suppose that (6.27) holds with $V \geq 0$ everywhere. Then, the following transient bound holds for each $n \geq 1$, and each $x \in \mathsf{X}$:*

$$\frac{1}{n} \sum_{k=0}^{n-1} P^k c\,(x) \leq \overline{\eta} + \frac{1}{n} V(x)$$

*Consequently, the average-cost admits the bound $\eta(x) \leq \overline{\eta}$.*

*Proof.* Note that $c \leq V - PV + \bar{\eta}\mathbf{1}$. Hence, for all $k \geq 0$, $P^k \leq P^k V - P^{k+1} V + \bar{\eta}\mathbf{1}$. Summing for $k = 0, \ldots, n-1$, we get that $\sum_{k=0}^{n-1} P^k c \leq V - P^n V + n\bar{\eta}\mathbf{1} \leq V + n\bar{\eta}\mathbf{1}$, where we used that $P^n V \geq 0$ to obtain the last inequality. □

# Discounted cost



given a discount parameter $\gamma \in (0,1)$, the discounted cost from initial condition $x$ is defined as

$$h_\gamma(x) = \sum_{k=0}^{\infty} \gamma^k \mathsf{E}[c(X(k)) \mid X(0) = x]\,. \tag{6.31}$$

Once again this has the operator-theoretic form,

$$h_\gamma = \sum_{k=0}^{\infty} \gamma^k P^k c\,, \tag{6.32}$$

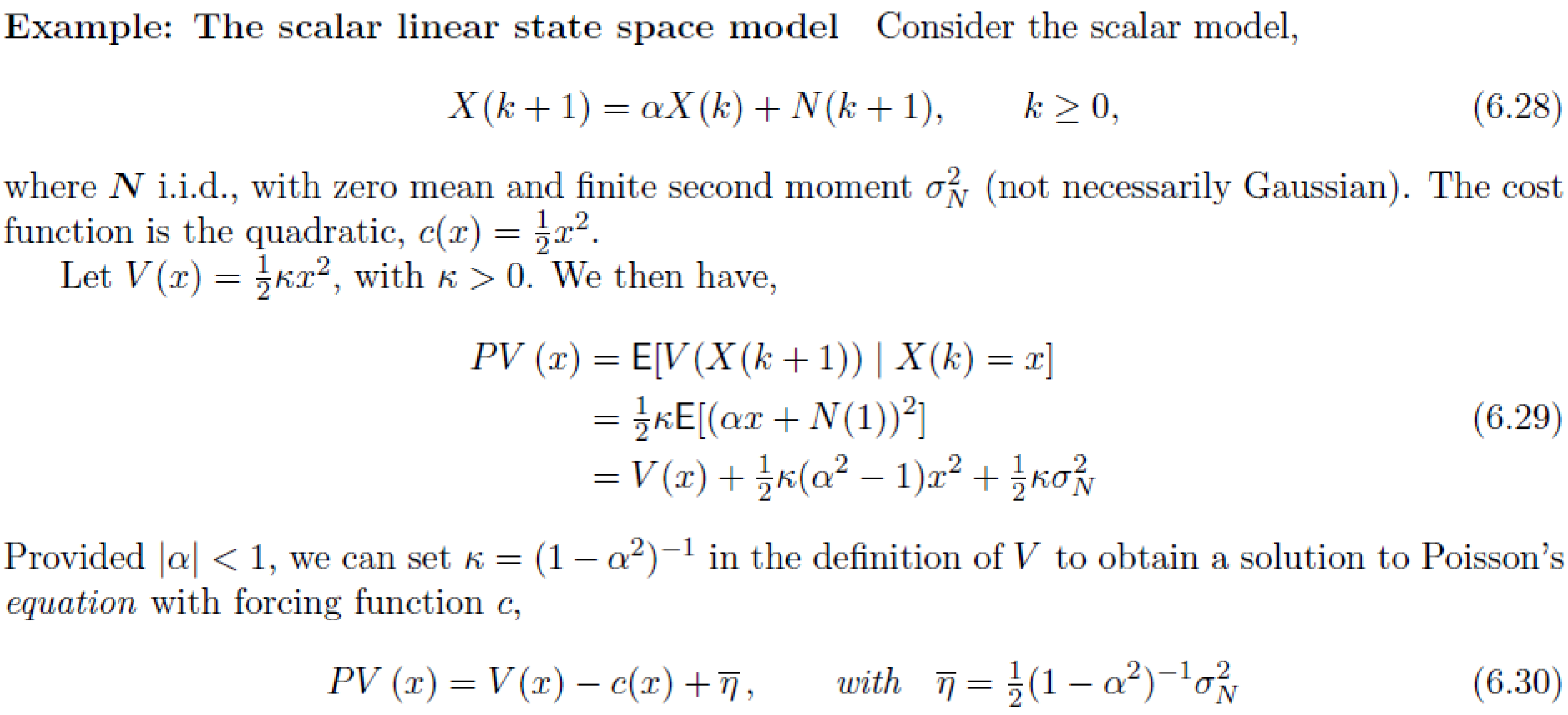
and from this we obtain a dynamic programming equation:

$$h_\gamma = c + \gamma P h_\gamma\,. \tag{6.33}$$

**Proposition 6.5.** *If (6.27) holds with $V \ge 0$ everywhere, then $h_\gamma(x) \le V(x) + \overline{\eta}(1-\gamma)^{-1}$ for each $x$, and $\gamma \in (0,1)$.*

Proof is similar.

# Example: Linear system



**Example: The scalar linear state space model** Consider the scalar model,

$$X(k+1) = \alpha X(k) + N(k+1), \qquad k \geq 0, \tag{6.28}$$

where $\boldsymbol{N}$ i.i.d., with zero mean and finite second moment $\sigma_N^2$ (not necessarily Gaussian). The cost function is the quadratic, $c(x) = \frac{1}{2}x^2$.

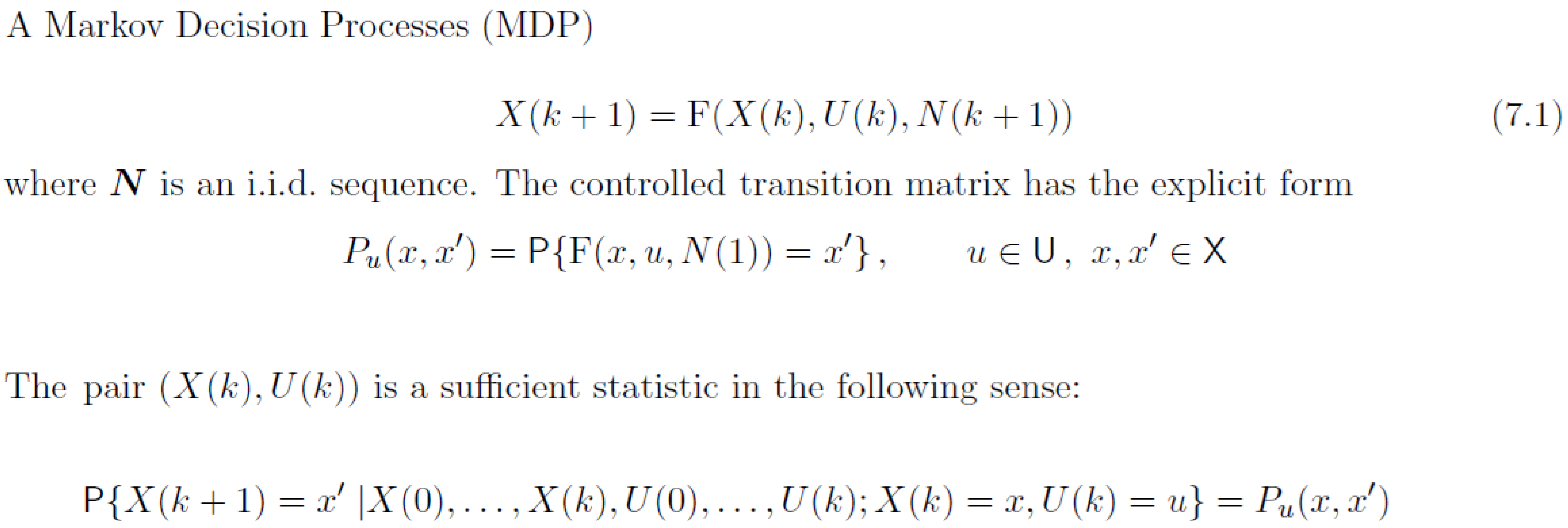
Let $V(x) = \frac{1}{2}\kappa x^2$, with $\kappa > 0$. We then have,

$$\begin{aligned} PV(x) &= \mathsf{E}[V(X(k+1)) \mid X(k) = x] \\ &= \tfrac{1}{2}\kappa \mathsf{E}[(\alpha x + N(1))^2] \\ &= V(x) + \tfrac{1}{2}\kappa(\alpha^2 - 1)x^2 + \tfrac{1}{2}\kappa\sigma_N^2 \end{aligned} \tag{6.29}$$

Provided $|\alpha| < 1$, we can set $\kappa = (1-\alpha^2)^{-1}$ in the definition of $V$ to obtain a solution to Poisson's *equation* with forcing function $c$,

$$PV(x) = V(x) - c(x) + \overline{\eta}, \qquad \textit{with} \quad \overline{\eta} = \tfrac{1}{2}(1-\alpha^2)^{-1}\sigma_N^2 \tag{6.30}$$

# Stochastic control systems



A Markov Decision Processes (MDP)

$$X(k+1) = \mathrm{F}(X(k), U(k), N(k+1)) \tag{7.1}$$

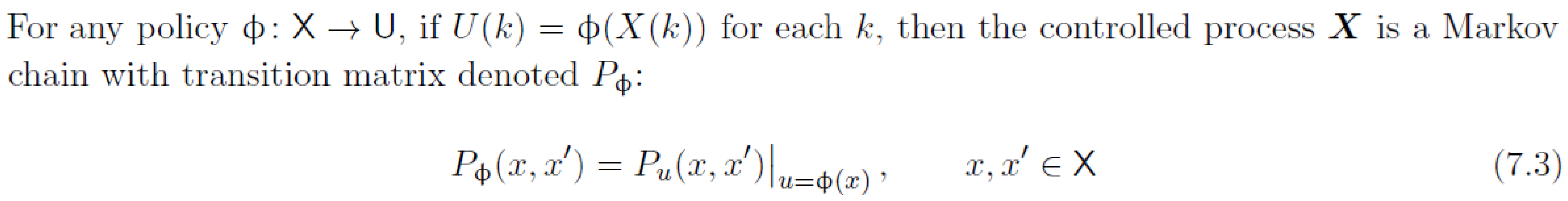
where $\boldsymbol{N}$ is an i.i.d. sequence. The controlled transition matrix has the explicit form

$$P_u(x, x') = \mathsf{P}\{\mathrm{F}(x, u, N(1)) = x'\}\,, \qquad u \in \mathsf{U}\,,\ x, x' \in \mathsf{X}$$

The pair $(X(k), U(k))$ is a sufficient statistic in the following sense:

$$\mathsf{P}\{X(k+1) = x' \mid X(0), \dots, X(k), U(0), \dots, U(k); X(k) = x, U(k) = u\} = P_u(x, x')$$

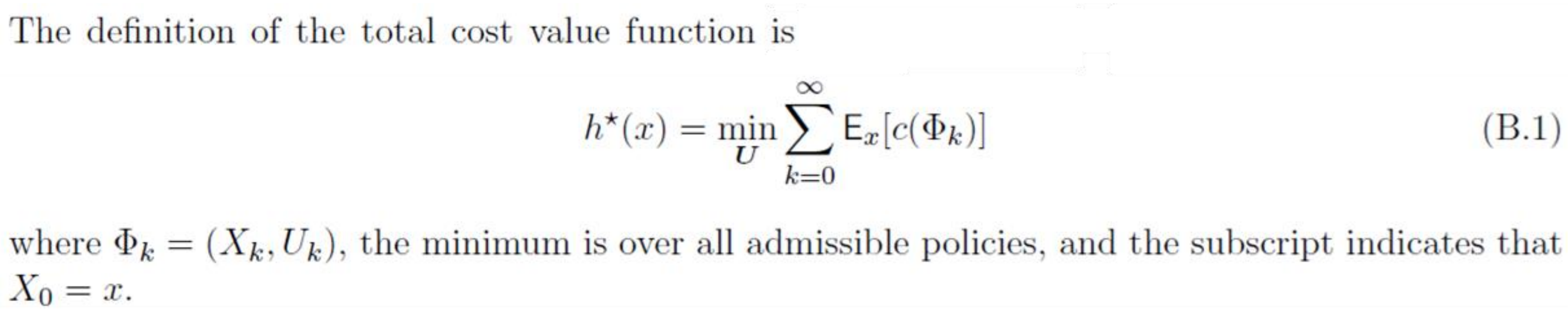
# Policy and closed-loop system



For any policy $\phi\colon \mathsf{X} \to \mathsf{U}$, if $U(k) = \phi(X(k))$ for each $k$, then the controlled process $\boldsymbol{X}$ is a Markov chain with transition matrix denoted $P_\phi$:

$$P_\phi(x, x') = P_u(x, x')\big|_{u=\phi(x)}, \qquad x, x' \in \mathsf{X} \tag{7.3}$$

# Optimal total cost

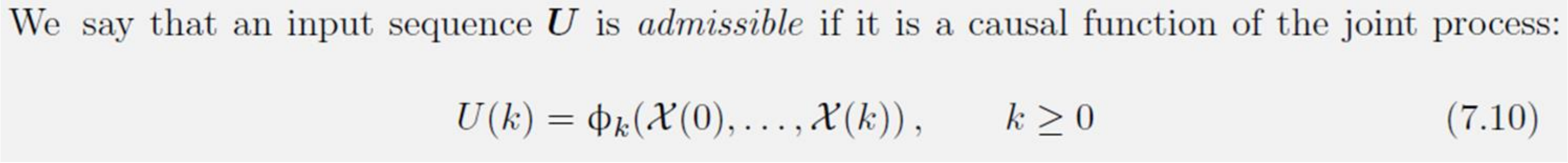


The definition of the total cost value function is

$$h^\star(x) = \min_{\boldsymbol{U}} \sum_{k=0}^{\infty} \mathsf{E}_x[c(\Phi_k)] \tag{B.1}$$

where $\Phi_k = (X_k, U_k)$, the minimum is over all admissible policies, and the subscript indicates that $X_0 = x$.

Note 1:



We say that an input sequence $\boldsymbol{U}$ is *admissible* if it is a causal function of the joint process:

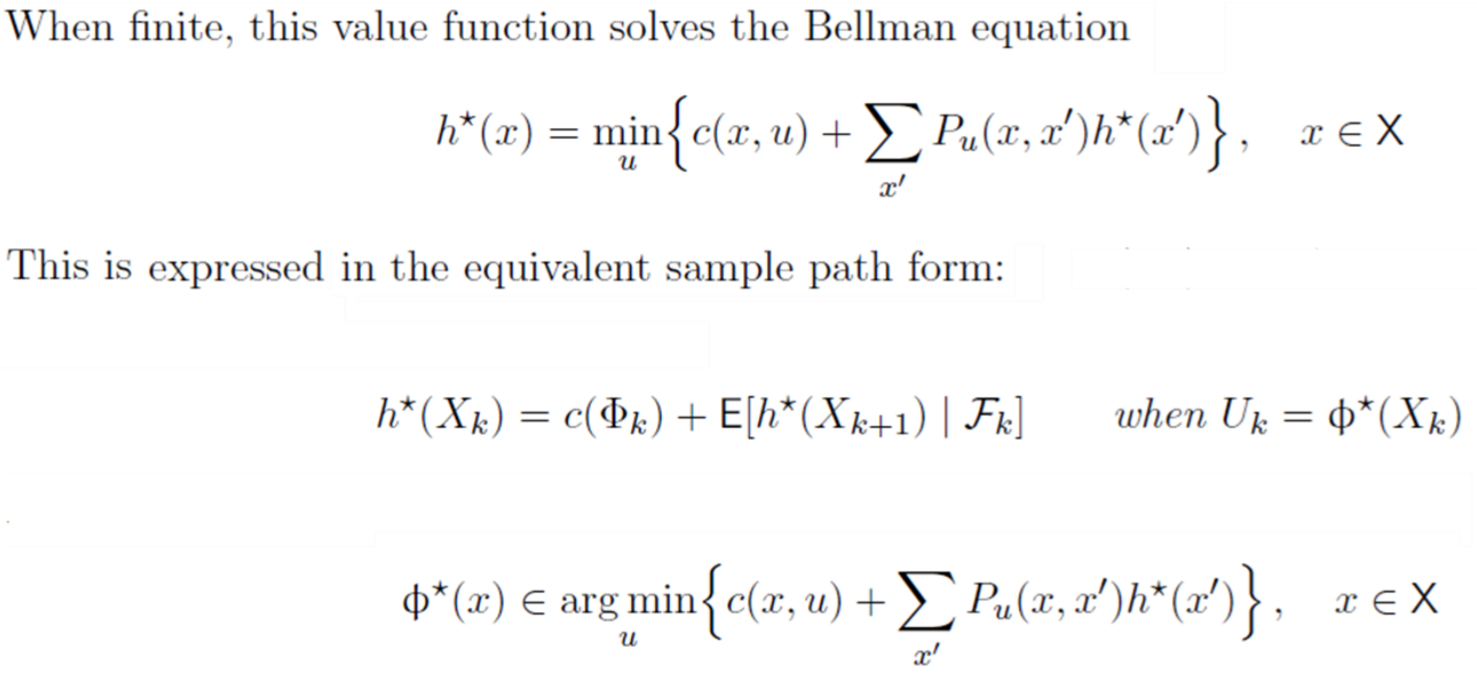
$$U(k) = \phi_k(\mathcal{X}(0), \dots, \mathcal{X}(k)), \qquad k \geq 0 \tag{7.10}$$

Note 2: For stochastic systems, in many cases the optimal total cost is **not** finite

The intuitive idea of admissible input is that it depends only on the current and past states of the system.

A more formal definition of admissible input is that $U$ must be adapted to the filtration $(\mathcal{F}_k)_{k=0}^{\infty}$, where $\mathcal{F}_k = \sigma(N(1), \ldots, N(k))$ is the $\sigma$-algebra generated by the noise process $N$ up to time $k$. A stochastic process $U$ is adapted to a filtration $(\mathcal{F}_k)_{k=0}^{\infty}$ if for all $k \geq 0$, $U(k)$ is measurable with respect to $\mathcal{F}_k$.

# Bellman equation for the optimal total cost



When finite, this value function solves the Bellman equation

$$h^\star(x) = \min_u \Big\{ c(x,u) + \sum_{x'} P_u(x,x') h^\star(x') \Big\}, \quad x \in \mathsf{X}$$

This is expressed in the equivalent sample path form:

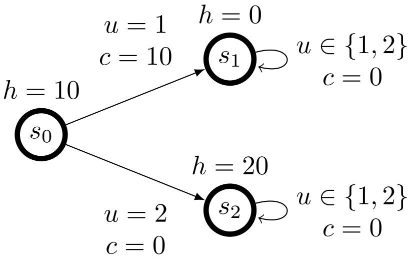
$$h^\star(X_k) = c(\Phi_k) + \mathsf{E}[h^\star(X_{k+1}) \mid \mathcal{F}_k] \qquad \textit{when } U_k = \phi^\star(X_k)$$

$$\phi^\star(x) \in \arg\min_u \Big\{ c(x,u) + \sum_{x'} P_u(x,x') h^\star(x') \Big\}, \quad x \in \mathsf{X}$$

Richard E. Bellman
1920-1984

If the optimal total cost $h^\star$ is finite for every $x$, then it satisfies the Bellman equation.

In many cases, our approach to stochastic optimal control is to solve (approximately) the Bellman equation (or similar) and hope that it provides the optimal cost and policy. This raises the question of whether any solution to the Bellman equation provides the optimal cost. The answer in general is no (even for finite systems). Example: consider the system



.

We observe that $h$ satisfies the Bellman equation but $\phi^h$ is not optimal.

A sufficient condition for a solution $h$ of the Bellman equation to provide an optimal policy is that $X$ and $U$ are finite, $c$ is nonnegative and vanishes only at $x = x_e$.

*Proof.*

$$\mathsf{E}\left[\sum_{k=0}^{n-1} c(X(k), U(k))\right] \geq h(X(0)) - \mathsf{E}[h(X(n))].$$

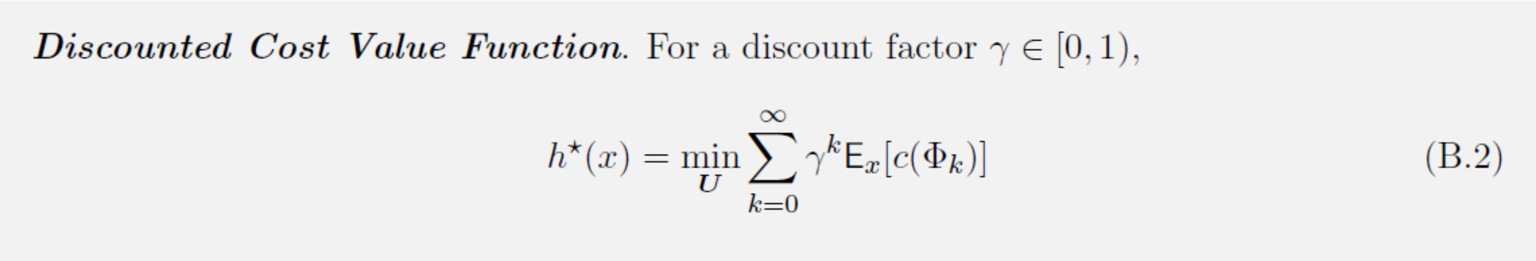
When the optimal total cost is finite, the assumptions on $c$ imply that $\mathsf{P}[X(n) = x_e] \to 1$ as $n \to \infty$. Hence, it holds that $\mathsf{E}\left[\sum_{k=0}^{\infty} c(X(k), U(k))\right] \geq h(X(0)) - h(x_e)$. The inequality is an equality when $U(k) = \phi(X(k))$ where

$$\phi(x) \in \operatorname*{arg\,min}_{u \in \mathsf{U}} \left\{c(x, u) + P_u h(x)\right\},$$

concluding the proof. □

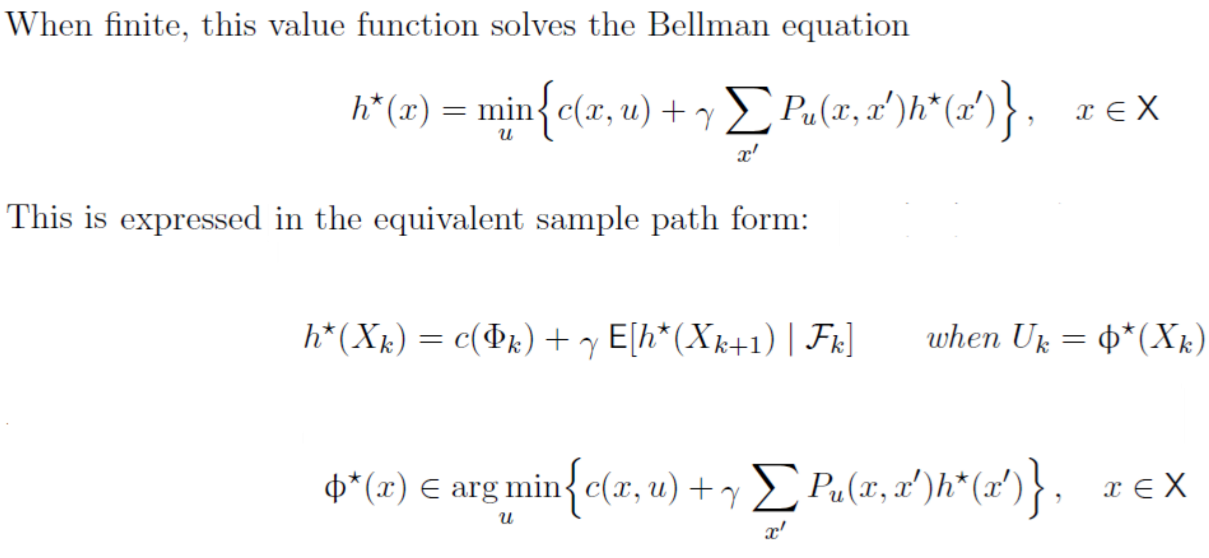
In practice, we will ignore these issues and assume that the (approximate) solutions to Bellman equation provide near-optimal policies.

# Optimal total discounted cost



***Discounted Cost Value Function.*** For a discount factor $\gamma \in [0,1)$,

$$h^\star(x) = \min_{\boldsymbol{U}} \sum_{k=0}^{\infty} \gamma^k \mathsf{E}_x[c(\Phi_k)] \tag{B.2}$$



When finite, this value function solves the Bellman equation

$$h^\star(x) = \min_u \Big\{ c(x,u) + \gamma \sum_{x'} P_u(x,x') h^\star(x') \Big\}, \quad x \in \mathsf{X}$$

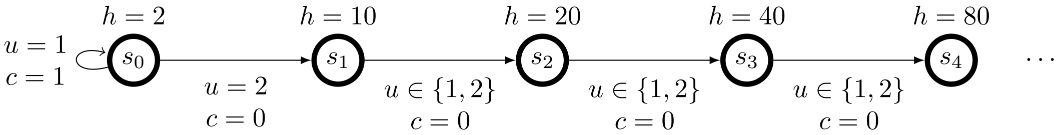
This is expressed in the equivalent sample path form:

$$h^\star(X_k) = c(\Phi_k) + \gamma\, \mathsf{E}[h^\star(X_{k+1}) \mid \mathcal{F}_k] \qquad \textit{when } U_k = \phi^\star(X_k)$$

$$\phi^\star(x) \in \arg\min_u \Big\{ c(x,u) + \gamma \sum_{x'} P_u(x,x') h^\star(x') \Big\}, \quad x \in \mathsf{X}$$

If the optimal discounted cost $h^\star$ is finite for every $x$, then it satisfies the Bellman equation.

The converse is not always true: a solution to Bellman equation may not always provide an optimal policy for the discounted cost. Example: consider the system



with $\gamma = \frac{1}{2}$. We observe that $h$ satisfies the Bellman equation but $\phi^h$ is not optimal.

A sufficient condition for a solution $h$ of the Bellman equation to provide an optimal policy is that $h$ is bounded from below and above.

*Proof.*

$$\mathsf{E}\left[\sum_{k=0}^{n-1} \gamma^k c(X(k), U(k))\right] \geq h(X(0)) - \gamma^n \mathsf{E}[h(X(n))].$$

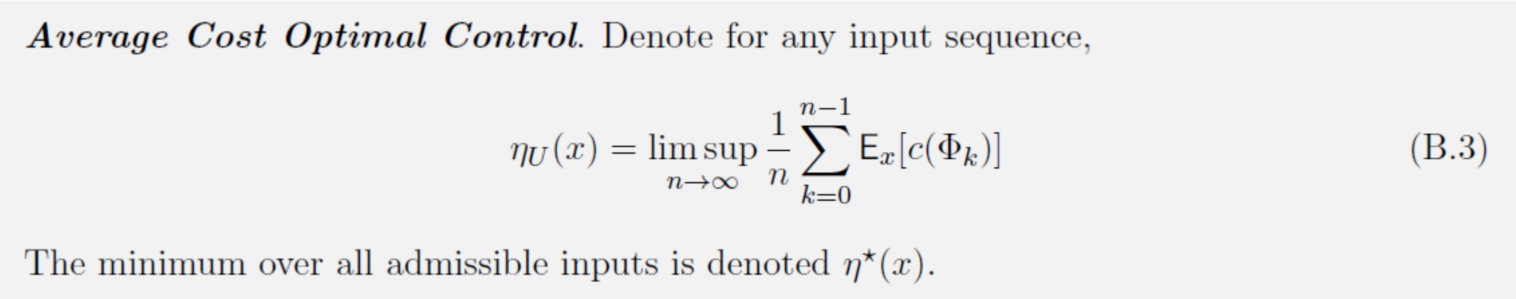
Hence, $\mathsf{E}\left[\sum_{k=0}^{\infty} \gamma^k c(X(k), U(k))\right] \geq h(X(0))$. The inequality is an equality when $U(k) = \phi(X(k))$ where

$$\phi(x) \in \operatorname*{arg\,min}_{u \in \mathsf{U}} \left\{c(x, u) + \gamma P_u h(x)\right\},$$

concluding the proof. □

In practice, we will ignore these issues and assume that the (approximate) solutions to Bellman equation provide near-optimal policies.

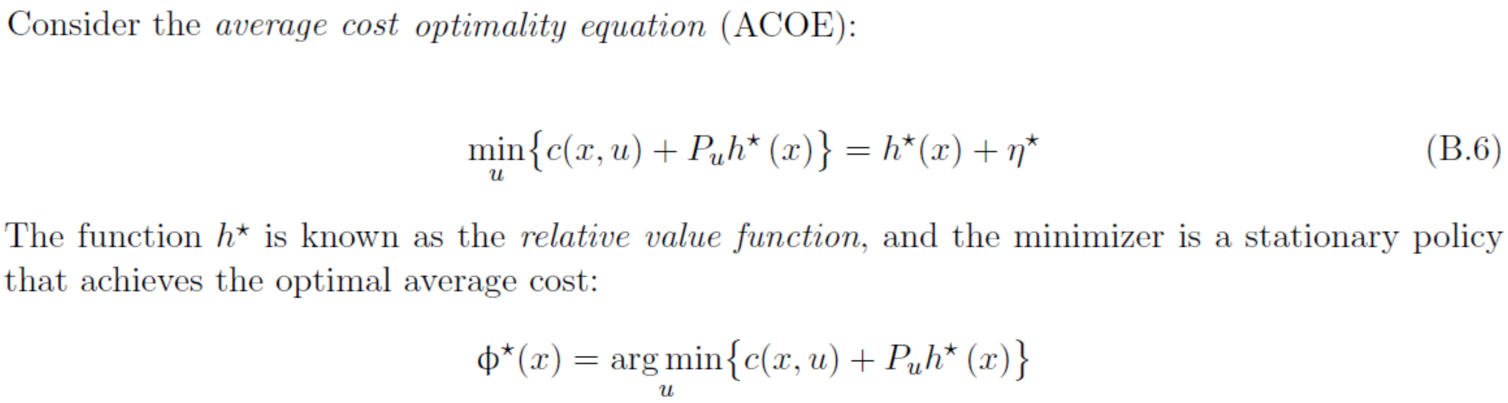
# Optimal average cost



***Average Cost Optimal Control.*** Denote for any input sequence,

$$\eta_U(x) = \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathsf{E}_x[c(\Phi_k)] \tag{B.3}$$

The minimum over all admissible inputs is denoted $\eta^\star(x)$.



Consider the *average cost optimality equation* (ACOE):

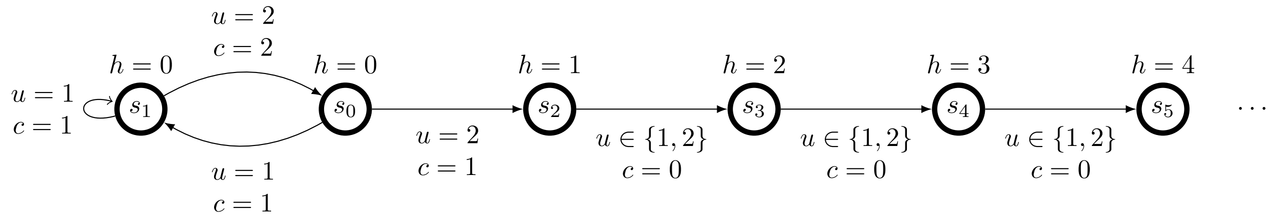
$$\min_u \bigl\{c(x,u) + P_u h^\star(x)\bigr\} = h^\star(x) + \eta^\star \tag{B.6}$$

The function $h^\star$ is known as the *relative value function*, and the minimizer is a stationary policy that achieves the optimal average cost:

$$\phi^\star(x) = \arg\min_u \bigl\{c(x,u) + P_u h^\star(x)\bigr\}$$

The optimal average cost, even if finite, can depend on $x$. In this case, no solution to the ACOE can be obtained. Even if $\eta^\star(x)$ is independent of $x$, it is not clear how to build from it a solution to the ACOE.

What can we say about the converse: does any solution $(h, \eta)$ to the ACOE provide an optimal policy for the average cost? The answer in general is no. Example: consider the system



.

We observe that $(h, \eta)$, with $h$ given above and $\eta = 1$, satisfies the ACOE but $\phi^h$ is not optimal.

A sufficient condition for a solution $(h, \eta)$ of the ACOE to provide an optimal policy is that $h$ is bounded from below and above.

*Proof.*

$$\frac{1}{n} \mathsf{E}\left[\sum_{k=0}^{n-1} c(X(k), U(k))\right] \geq \frac{1}{n} h(X(0)) - \frac{1}{n} \mathsf{E}[h(X(n))] + \eta.$$

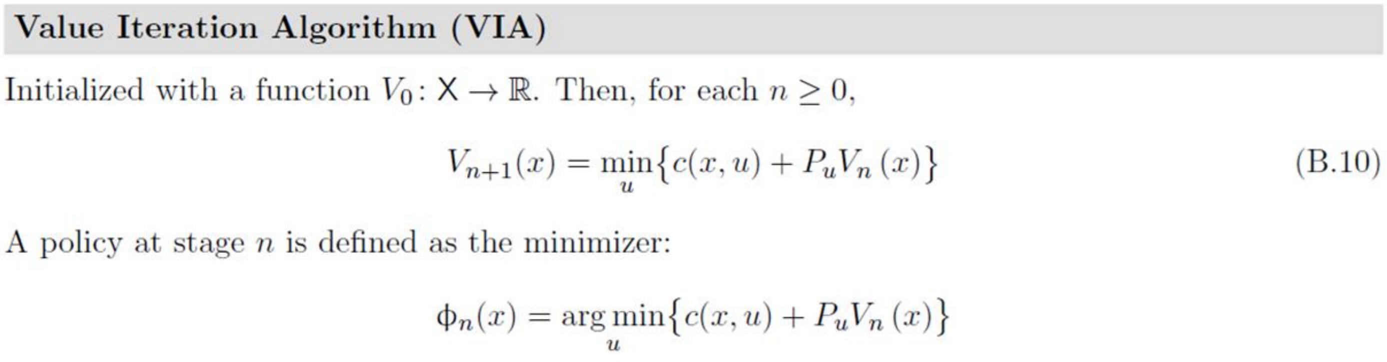
Hence, $\limsup_{n\to\infty} \frac{1}{n} \mathsf{E}\left[\sum_{k=0}^{n-1} c(X(k), U(k))\right] \geq \eta$. The inequality is an equality when $U(k) = \phi(X(k))$ where

$$\phi(x) \in \operatorname*{arg\,min}_{u \in \mathsf{U}} \left\{c(x, u) + P_u h(x)\right\},$$

concluding the proof. □

In practice, we will ignore these issues and assume that the (approximate) solutions to the ACOE provide near-optimal policies.

# Value iteration



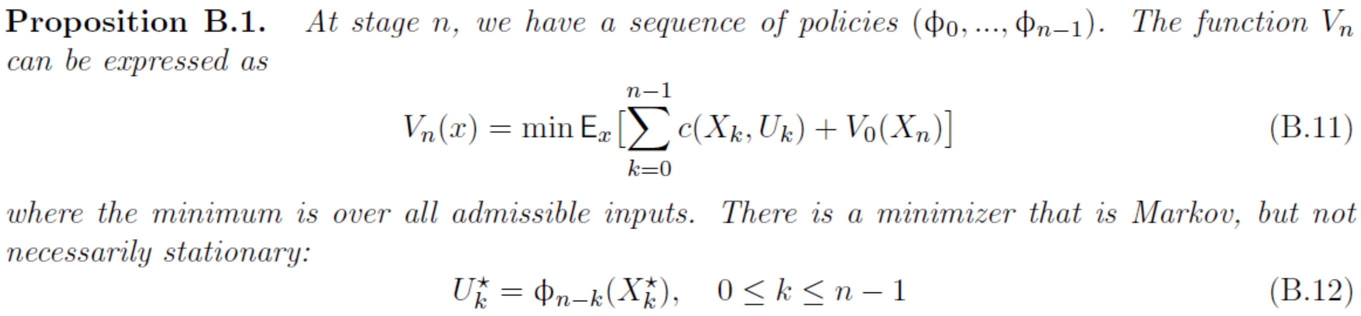
**Value Iteration Algorithm (VIA)**

Initialized with a function $V_0 \colon \mathsf{X} \to \mathbb{R}$. Then, for each $n \geq 0$,

$$V_{n+1}(x) = \min_u \{ c(x,u) + P_u V_n\,(x) \} \tag{B.10}$$

A policy at stage $n$ is defined as the minimizer:

$$\phi_n(x) = \arg\min_u \{ c(x,u) + P_u V_n\,(x) \}$$

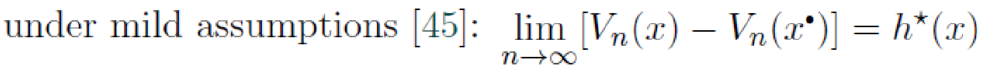


**Proposition B.1.** *At stage $n$, we have a sequence of policies $(\phi_0, \ldots, \phi_{n-1})$. The function $V_n$ can be expressed as*

$$V_n(x) = \min \mathsf{E}_x \Big[ \sum_{k=0}^{n-1} c(X_k, U_k) + V_0(X_n) \Big] \tag{B.11}$$

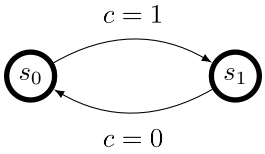
*where the minimum is over all admissible inputs. There is a minimizer that is Markov, but not necessarily stationary:*

$$U_k^\star = \phi_{n-k}(X_k^\star), \quad 0 \leq k \leq n-1 \tag{B.12}$$



under mild assumptions [45]: $\lim_{n \to \infty} [V_n(x) - V_n(x^\bullet)] = h^\star(x)$

VIA may not always converge (even if a solution to the ACOE exists). Example: for the system



with $V_0(s_0) = V_0(s_1) = 0$, it does not converge.

A sufficient condition for the VIA to converge to a solution of the ACOE is the following: (i) $X$ and $U$ are finite, (ii) the ACOE has a solution $(h^\star, \eta^\star)$, (iii) there is a unique policy $\phi^\star$ corresponding to $(h^\star, \eta^\star)$, (iv) $P \triangleq P_{\phi^\star}$ admits an invariant measure $\pi$, and (v) there is $\epsilon > 0$ and $T \geq 0$ such that for all trajectory $(X, U)$ of the system, all $x \in \mathsf{X}$ and all $k \geq T$, $\mathsf{P}[X(k) = x] \geq \epsilon$.

*Proof.* Let $n \geq 0$ (which will be assumed very large) and let $(X^\star, U^\star)$ be as in (B.12) for this $n$. The assumptions (iii) and (v) imply that there is $K \geq 0$ (independent of $n$, $X^\star$ and $U^\star$) such that $U^\star(k) \neq \phi^\star(X^\star(k))$ for at most $K$ values of $k$ in $\{0, \dots, n-1\}$. Let $1 \leq k_1 \leq k_2 \leq n$. Assume that $U^\star(k) = \phi^\star(X^\star(k))$ for all $k_1 \leq k \leq k_2 - 1$. Then,

$$\begin{aligned}
V_{n-k_1}(X^\star(k_1)) &= \mathsf{E}\left[\sum_{k=k_1}^{k_2-1} c(X^\star(k), U^\star(k)) + V_{n-k_2}(X^\star(k_2)) \;\middle|\; X^\star(k_1)\right] \\
&= \mathsf{E}\left[h^\star(X^\star(k_1)) + (k_2 - k_1)\eta^\star - h^\star(X^\star(k_2)) + V_{n-k_2}(X^\star(k_2)) \mid X^\star(k_1)\right] \\
&= h^\star(X^\star(k_1)) + (k_2 - k_1)\eta^\star - (P^{k_2-k_1} h^\star)(X^\star(k_1)) + (P^{k_2-k_1} V_{n-k_2})(X^\star(k_1)) \\
&= h^\star(X^\star(k_1)) + (k_2 - k_1)\eta^\star - \pi(h^\star) + \pi(V_{n-k_2}) + O((n - k_2)\rho^{k_2-k_1}),
\end{aligned}$$

where $\rho \in [0, 1)$ is the spectral radius of $P - \mathbf{1}^\top \pi$. Assume that $k_2 - k_1$ is large enough so that

$$|V_{n-k_1}(X^\star(k_1)) - \{h^\star(X^\star(k_1)) + (k_2 - k_1)\eta^\star - \pi(h^\star) + \pi(V_{n-k_2})\}| \ll 1.$$

Since $U^\star(k_1 - 1)$ minimizes

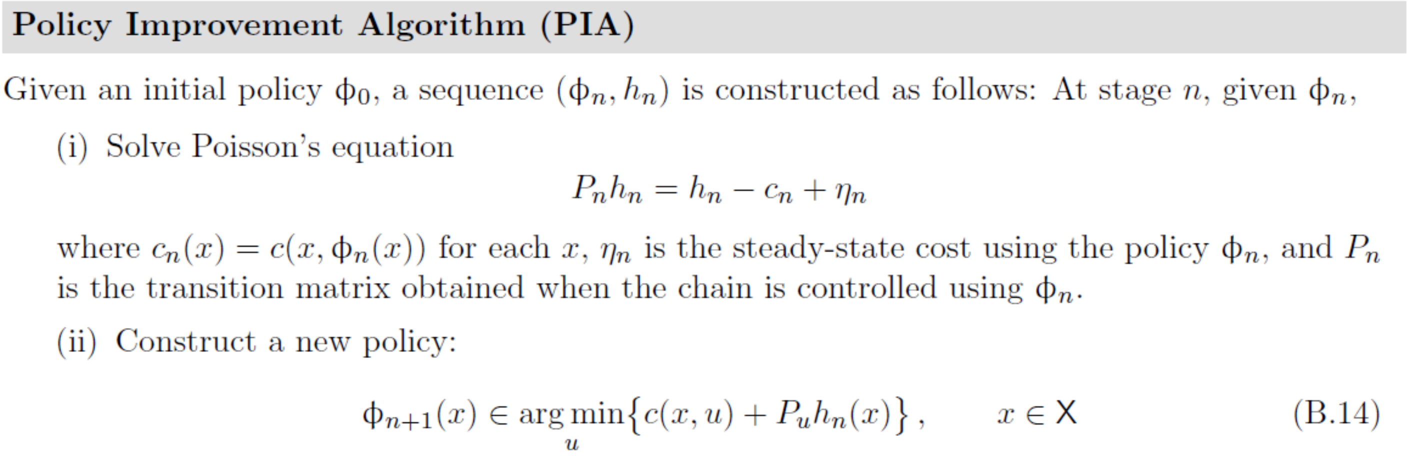
$$c(X^\star(k_1 - 1), U^\star(k_1 - 1)) + \mathsf{E}\left[V_{n-k_1}(X^\star(k_1)) \mid X^\star(k_1 - 1), U^\star(k_1 - 1)\right],$$

this implies by assumption (iii) that $U^\star(k_1 - 1) = \phi^\star(X^\star(k_1 - 1))$. Hence, the above all together implies that there is $m \geq 0$ (independent of $n$, $X^\star$ and $U^\star$) such that for all $0 \leq k \leq n - m$, $U^\star(k) = \phi^\star(X^\star(k))$. It follows that

$$V_n(x) = h^\star(x) + (n - m)\eta^\star - \pi(h^\star) + \pi(V_m) + O(m\rho^{n-m}).$$

Hence, $V_n(x) - V_n(x^\bullet) = h^\star(x) - h^\star(x^\bullet) + O(m\rho^{n-m})$. This concludes the proof. □

# Policy improvement



**Policy Improvement Algorithm (PIA)**

Given an initial policy $\phi_0$, a sequence $(\phi_n, h_n)$ is constructed as follows: At stage $n$, given $\phi_n$,

(i) Solve Poisson's equation

$$P_n h_n = h_n - c_n + \eta_n$$

where $c_n(x) = c(x, \phi_n(x))$ for each $x$, $\eta_n$ is the steady-state cost using the policy $\phi_n$, and $P_n$ is the transition matrix obtained when the chain is controlled using $\phi_n$.

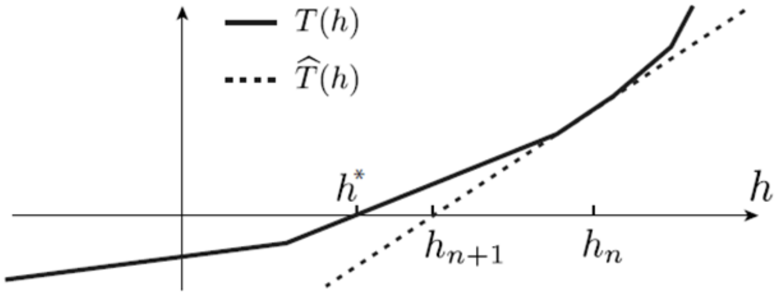
(ii) Construct a new policy:

$$\phi_{n+1}(x) \in \arg\min_u \{c(x,u) + P_u h_n(x)\}\,, \qquad x \in \mathsf{X} \qquad \text{(B.14)}$$

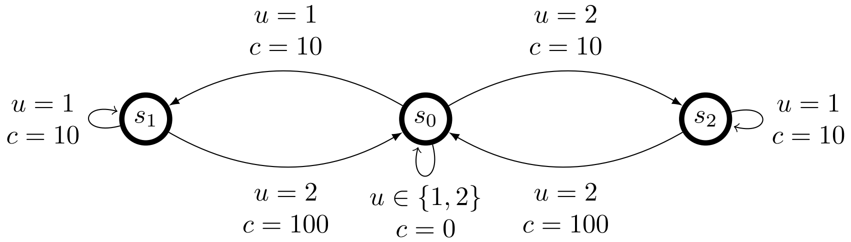
PIA is Newton-Raphson applied to the convex piecewise linear operator $T : (h, \eta) \mapsto T(h, \eta)$ defined by

$$T(h,\eta)(x) = \; h(x) + \eta - \min_u \; \{c(x,u) + P_u h(x)\}$$

Under mild assumptions, PIA converges in a finite number of steps



PIA may not always converge (even if a solution to the ACOE exists). Example: Consider the system



.

We start with $\phi(s_0) = \phi(s_1) = \phi(s_2) = 1$. A valid $h$ is given by $h(s_0) = 20$, $h(s_1) = 20$ and $h(s_2) = 0$, with $\eta = 10$. Hence, a new policy is $\phi(s_0) = 2$, $\phi(s_1) = \phi(s_2) = 1$. A valid $h$ is given by $h(s_0) = 20$, $h(s_1) = 0$ and $h(s_2) = 20$, with $\eta = 10$. Hence, a new policy is $\phi(s_0) = \phi(s_1) = \phi(s_2) = 1$, and we have looped back. However, the best average cost is $\eta^\star = 0$ with associated policy, e.g., $\phi(s_0) = \phi(s_1) = \phi(s_2) = 2$.

A sufficient condition for the PIA to converge toward an optimal policy in finite time is that $X$ and $U$ are finite and for every policy, the closed-loop system admits a stationary measure with full support.

*Proof.*

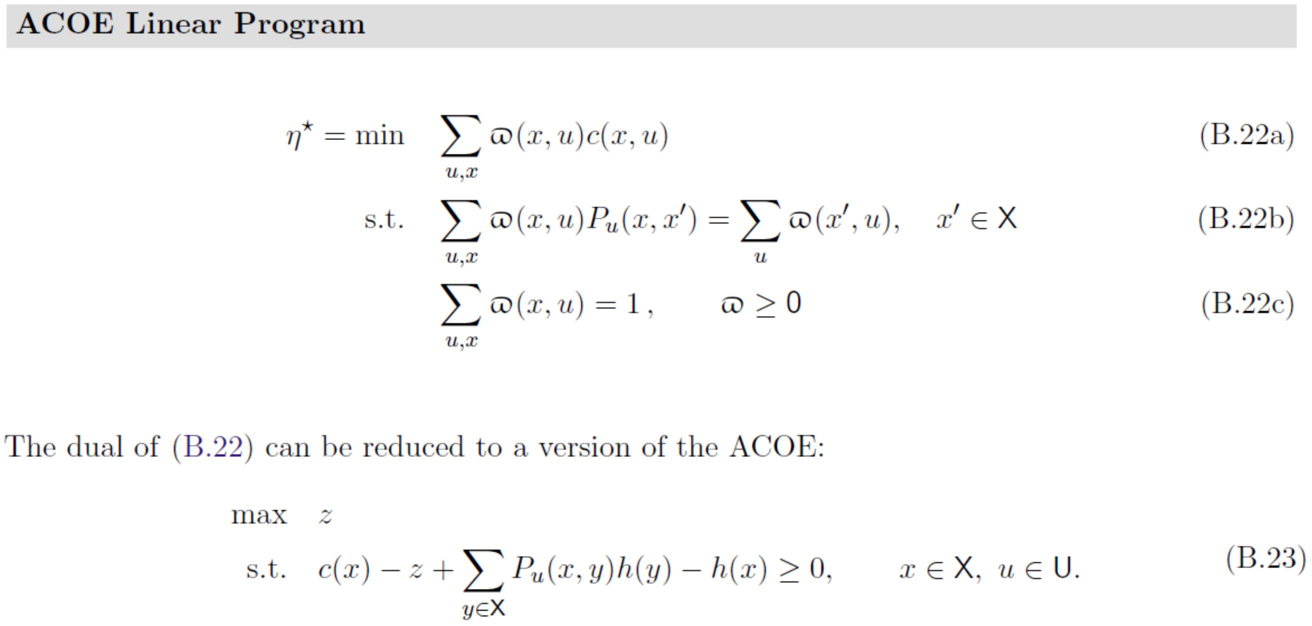
$$c_{n+1} + P_{n+1}h_n \leq c_n + P_n h_n = \eta_n \mathbf{1} + h_n,$$

where the inequality comes from the definition of $\phi_{n+1}$ and the equality comes from the definition of $h_n$. Left-multiplying both sides by $\pi_{n+1}^\top$ gives

$$\pi_{n+1}^\top c_{n+1} + \pi_{n+1}^\top P_{n+1} h_n \leq \eta_n \pi_{n+1}^\top \mathbf{1} + \pi_{n+1}^\top h_n \quad \implies \quad \eta_{n+1} + \pi_{n+1}^\top h_n \leq \eta_n + \pi_{n+1} n^\top h_n \quad \implies \quad \eta_{n+1} \leq \eta_n.$$

Hence, $\eta_n$ decreases with $n$ and strictly decreases if $\phi_n$ is not optimal for $h_n$ (we used the full support assumption here). Since there is a finite number of different policies, there is $n \geq 0$ such that $\phi_n$ is optimal for $h_n$. □

# ACOE Linear Programming



**ACOE Linear Program**

$$\eta^\star = \min \quad \sum_{u,x} \varpi(x,u)c(x,u) \tag{B.22a}$$

$$\text{s.t.} \quad \sum_{u,x} \varpi(x,u)P_u(x,x') = \sum_{u} \varpi(x',u), \quad x' \in \mathsf{X} \tag{B.22b}$$

$$\sum_{u,x} \varpi(x,u) = 1\,, \qquad \varpi \geq 0 \tag{B.22c}$$

The dual of (B.22) can be reduced to a version of the ACOE:

$$\begin{aligned} \max \quad & z \\ \text{s.t.} \quad & c(x) - z + \sum_{y\in\mathsf{X}} P_u(x,y)h(y) - h(x) \geq 0, \qquad x \in \mathsf{X},\ u \in \mathsf{U}. \end{aligned} \tag{B.23}$$

If $(h^*, \eta^*)$ solves the ACOE, then $(h, z) = (h^*, z^*)$ is an optimal solution of (B.23)

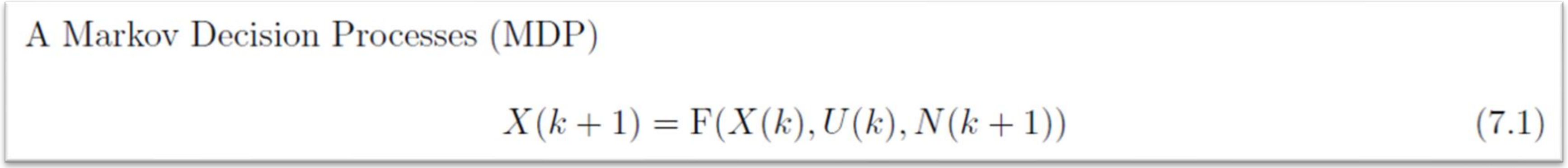
$\varpi(x, u)$ in (B.22) is interpreted as the invariant probability of being in state $x$ and taking input $u$ for an optimal closed-loop system

*Proof.* It is clear that $(h, z) = (h^\star, \eta^\star)$ is a feasible solution of (B.23). Let $(h, z)$ be a feasible solution of (B.23). Then, $z + h \leq \min_{u \in U} c + P_u h$. It follows that

$$\frac{1}{n}\mathsf{E}\left[\sum_{k=0}^{n-1} c(X(k), U(k))\right] \geq z + \frac{1}{n}h(X(0)) + \frac{1}{n}\mathsf{E}[h(X(n))],$$

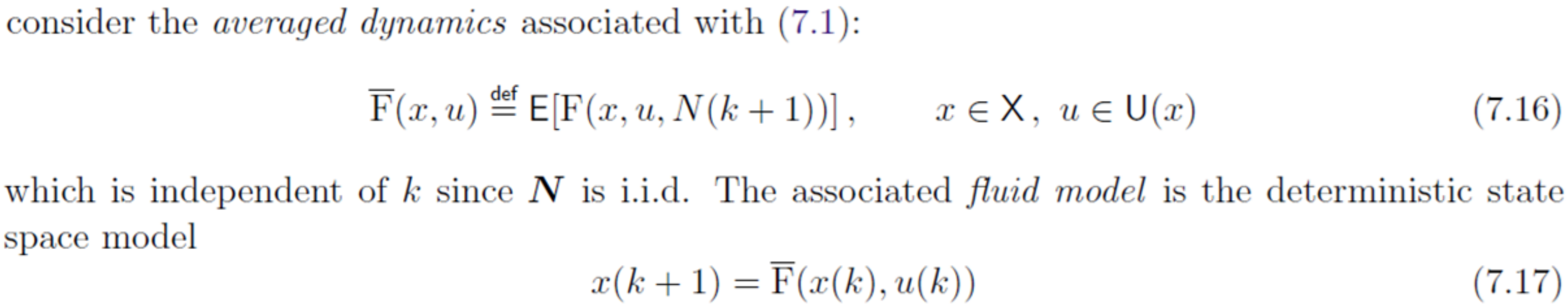
This implies that $z \leq \eta^\star$. Hence, $(h, z) = (h^\star, \eta^\star)$ is optimal. □

# Fluid model approximation



A Markov Decision Processes (MDP)

$$X(k+1) = \mathrm{F}(X(k), U(k), N(k+1)) \tag{7.1}$$



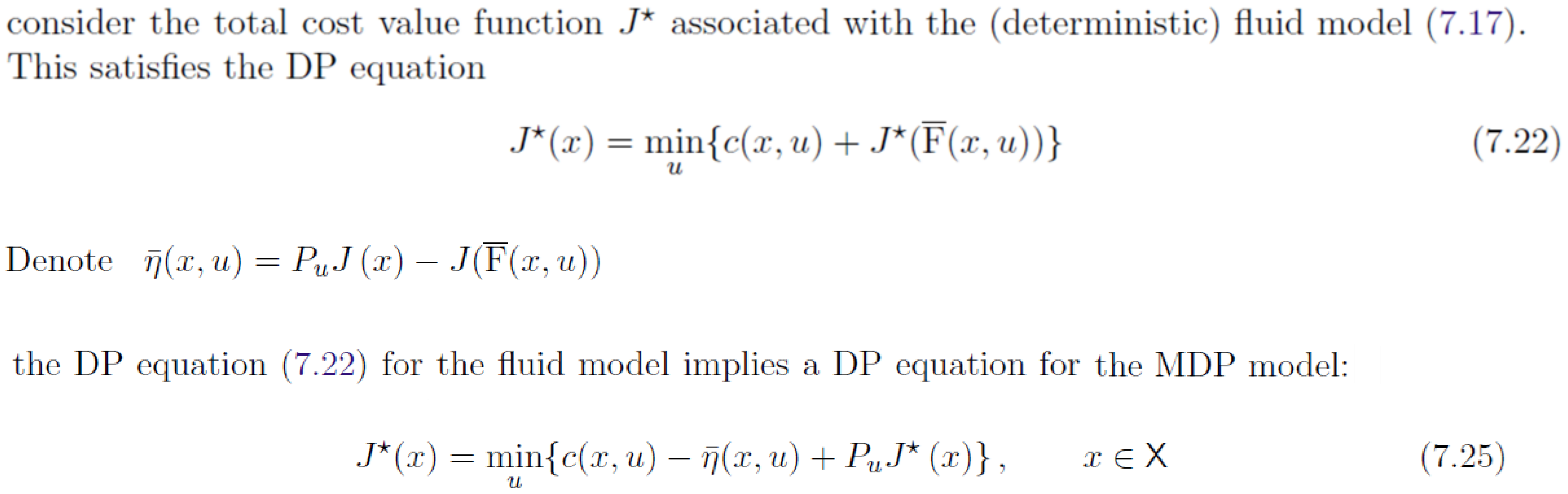
consider the *averaged dynamics* associated with (7.1):

$$\overline{\mathrm{F}}(x,u) \overset{\text{def}}{=} \mathsf{E}[\mathrm{F}(x,u,N(k+1))]\,, \qquad x \in \mathsf{X}\,,\ \ u \in \mathsf{U}(x) \tag{7.16}$$

which is independent of $k$ since $\boldsymbol{N}$ is i.i.d. The associated *fluid model* is the deterministic state space model

$$x(k+1) = \overline{\mathrm{F}}(x(k), u(k)) \tag{7.17}$$

# Fluid model and optimality



consider the total cost value function $J^\star$ associated with the (deterministic) fluid model (7.17). This satisfies the DP equation

$$J^\star(x) = \min_u \{c(x,u) + J^\star(\overline{\mathrm{F}}(x,u))\} \tag{7.22}$$

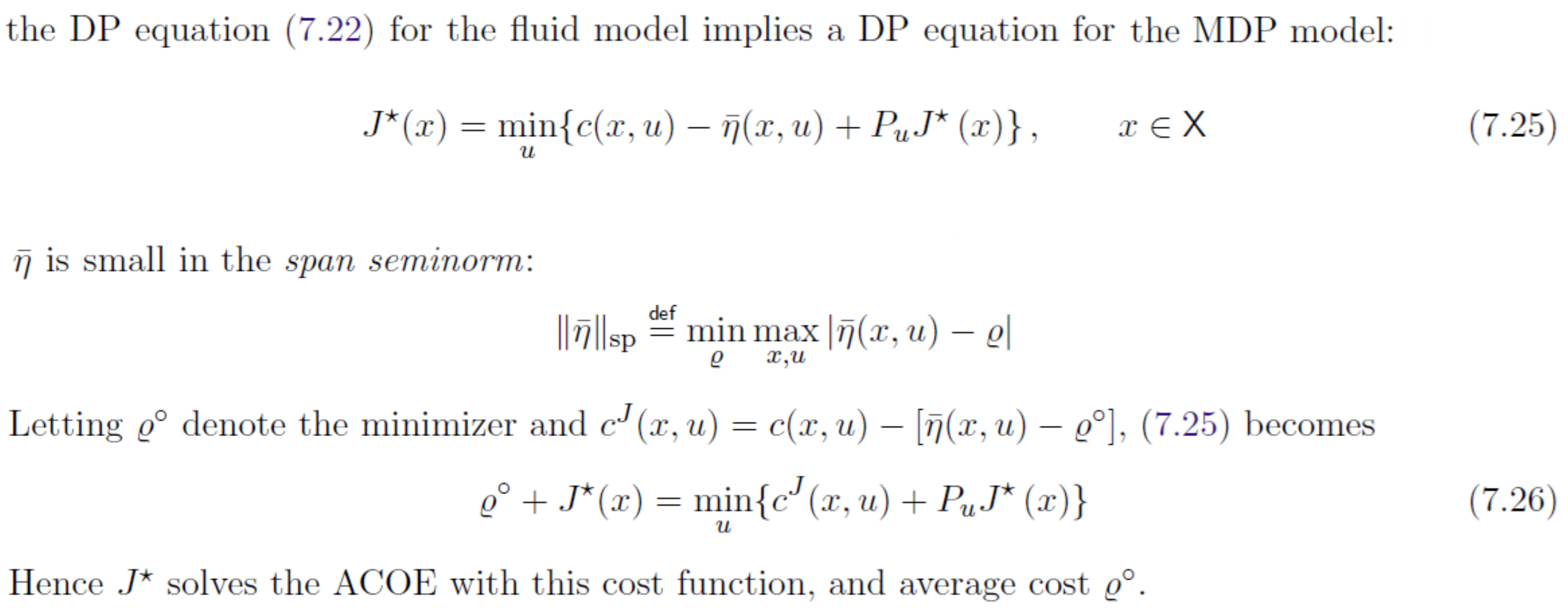
Denote $\bar{\eta}(x,u) = P_u J\,(x) - J(\overline{\mathrm{F}}(x,u))$

the DP equation (7.22) for the fluid model implies a DP equation for the MDP model:

$$J^\star(x) = \min_u \{c(x,u) - \bar{\eta}(x,u) + P_u J^\star\,(x)\}\,, \qquad x \in \mathsf{X} \tag{7.25}$$

# Fluid model and optimality

(continued)



the DP equation (7.22) for the fluid model implies a DP equation for the MDP model:

$$J^\star(x) = \min_u \{c(x,u) - \bar\eta(x,u) + P_u J^\star\,(x)\}\,, \qquad x \in \mathsf{X} \tag{7.25}$$

$\bar\eta$ is small in the *span seminorm*:

$$\|\bar\eta\|_{\text{sp}} \overset{\text{def}}{=} \min_{\varrho} \max_{x,u} |\bar\eta(x,u) - \varrho|$$

Letting $\varrho^\circ$ denote the minimizer and $c^J(x,u) = c(x,u) - [\bar\eta(x,u) - \varrho^\circ]$, (7.25) becomes

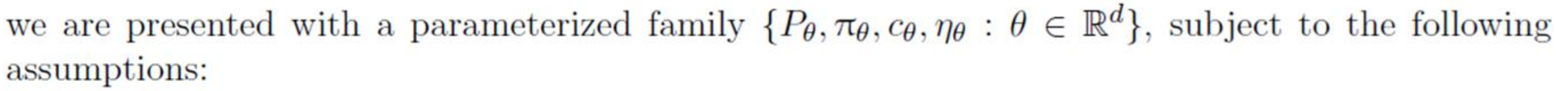
$$\varrho^\circ + J^\star(x) = \min_u \{c^J(x,u) + P_u J^\star\,(x)\} \tag{7.26}$$

Hence $J^\star$ solves the ACOE with this cost function, and average cost $\varrho^\circ$.

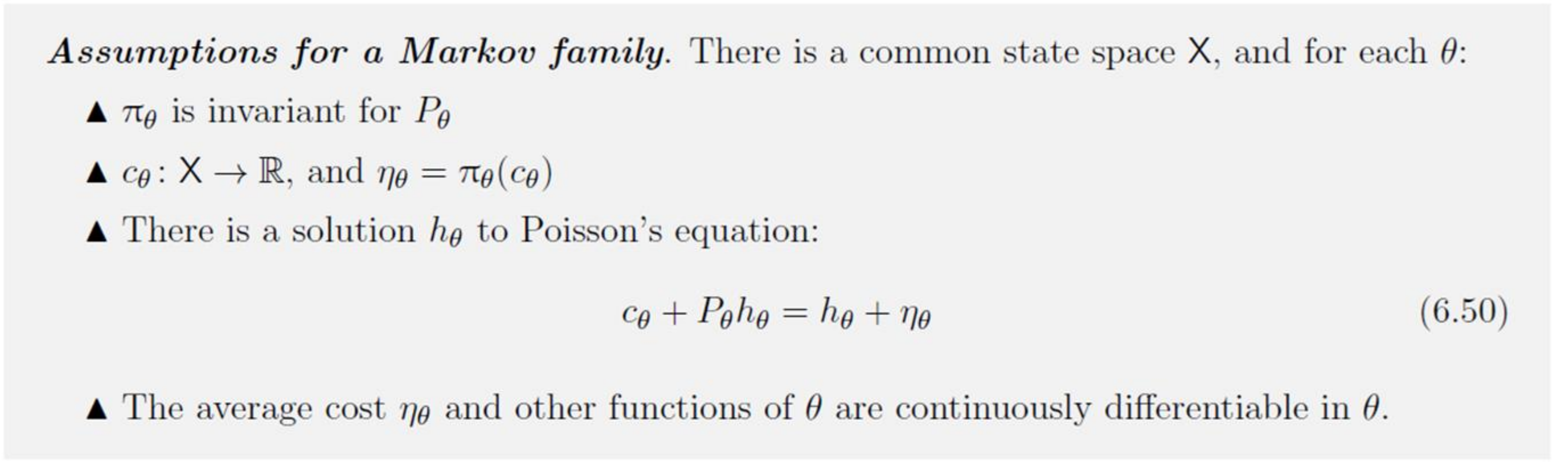
# Example: Fluid model of linear system

- Linear System: $X(k+1) = AX(k) + BU(k) + DN(k)$
- Fluid Model: $x(k+1) = Ax(k) + Bu(k)$
- Quadratic Cost: $c(x,u) = [x,u]^\top Q[x,u]$
- Optimal Cost of Fluid Model (LQR): $J^*(x) = x^\top Px$
- Discrepancy: $\overline{\eta}(x,u) \overset{\text{def}}{=} E[J^*(Ax+Bu+DN)] - J^*(Ax+Bu)$
- Observe: $\overline{\eta}(x,u) \overset{\text{def}}{=} J^*(Ax+Bu) + E[N^\top PN] - J^*(Ax+Bu)$
- Hence: $J^*(x) + E[N^\top PN] = \min_u \{c(x,u) + P_u J^*(x)\}$
- The LQR policy gives the optimal average cost: $\eta^* = E[N^\top PN]$

# Parameterized systems



we are presented with a parameterized family $\{P_\theta, \pi_\theta, c_\theta, \eta_\theta : \theta \in \mathbb{R}^d\}$, subject to the following assumptions:

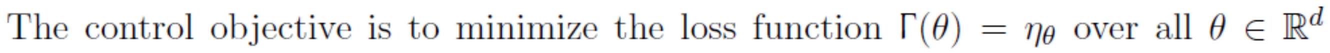


***Assumptions for a Markov family.*** There is a common state space $\mathsf{X}$, and for each $\theta$:

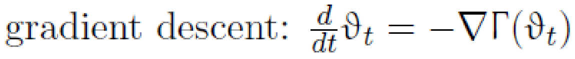
- ▲ $\pi_\theta$ is invariant for $P_\theta$
- ▲ $c_\theta\colon \mathsf{X} \to \mathbb{R}$, and $\eta_\theta = \pi_\theta(c_\theta)$
- ▲ There is a solution $h_\theta$ to Poisson's equation:

$$c_\theta + P_\theta h_\theta = h_\theta + \eta_\theta \tag{6.50}$$

- ▲ The average cost $\eta_\theta$ and other functions of $\theta$ are continuously differentiable in $\theta$.



The control objective is to minimize the loss function $\Gamma(\theta) = \eta_\theta$ over all $\theta \in \mathbb{R}^d$



gradient descent: $\frac{d}{dt}\vartheta_t = -\nabla\Gamma(\vartheta_t)$

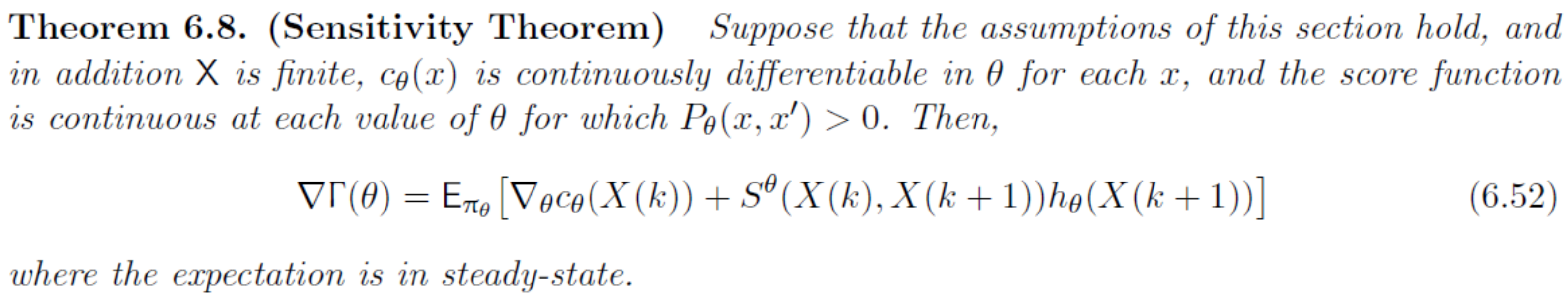
# Score function

define the *score function* : $\qquad S^{\theta}(x,x') = \nabla_{\theta} \log(P_{\theta}(x,x'))\,, \qquad x,x' \in \mathsf{X} \qquad (6.51)$

**Lemma 6.7.** *For a finite state space Markov chain, and any function* $g \colon \mathsf{X} \to \mathbb{R}$,

$$\nabla_{\theta}\Big\{\sum_{x'} P_{\theta}(x,x')g(x')\Big\} = \sum_{x'} P_{\theta}(x,x')S^{\theta}(x,x')g(x') \qquad \square$$

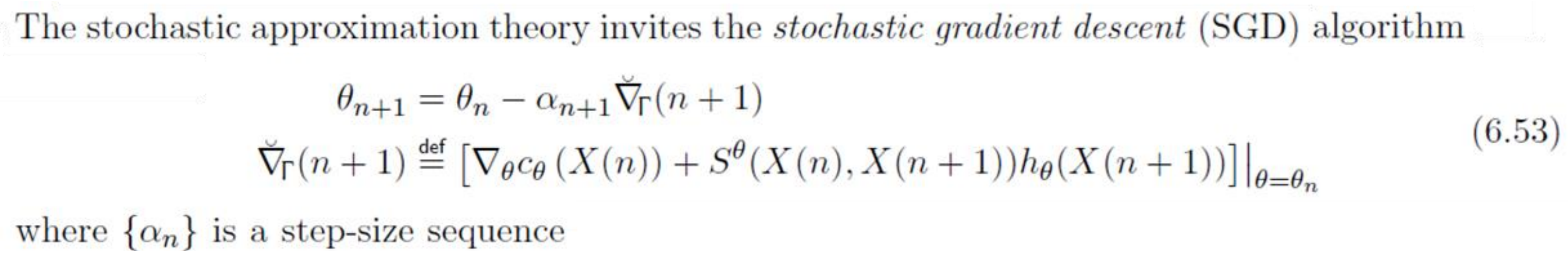
# Sensitivity theorem



**Theorem 6.8. (Sensitivity Theorem)** *Suppose that the assumptions of this section hold, and in addition* $\mathsf{X}$ *is finite,* $c_\theta(x)$ *is continuously differentiable in* $\theta$ *for each* $x$*, and the score function is continuous at each value of* $\theta$ *for which* $P_\theta(x, x') > 0$*. Then,*

$$\nabla\Gamma(\theta) = \mathsf{E}_{\pi_\theta}\big[\nabla_\theta c_\theta(X(k)) + S^\theta(X(k), X(k+1))h_\theta(X(k+1))\big] \tag{6.52}$$

*where the expectation is in steady-state.*



The stochastic approximation theory invites the *stochastic gradient descent* (SGD) algorithm

$$\begin{aligned} \theta_{n+1} &= \theta_n - \alpha_{n+1}\breve{\nabla}_{\Gamma}(n+1) \\ \breve{\nabla}_{\Gamma}(n+1) &\overset{\text{def}}{=} \big[\nabla_\theta c_\theta\,(X(n)) + S^\theta(X(n), X(n+1))h_\theta(X(n+1))\big]\big|_{\theta=\theta_n} \end{aligned} \tag{6.53}$$

where $\{\alpha_n\}$ is a step-size sequence

*Proof.*

$$h_\theta + \eta_\theta \mathbf{1} = c_\theta + P_\theta h_\theta.$$

Hence,

$$\nabla_\theta h_\theta + \nabla_\theta \eta_\theta = \nabla_\theta c_\theta + \nabla_\theta [P_\theta h_\theta].$$

It holds that

$$\begin{aligned}
\nabla_\theta \{P_\theta h_\theta\}(x) &= \nabla_\theta \left\{ \sum_{x' \in X} P_\theta(x, x') h_\theta(x') \right\} \\
&= \sum_{x' \in X} [\nabla_\theta P_\theta(x, x') h_\theta(x') + P_\theta(x, x') \nabla_\theta h_\theta(x')] \\
&= \sum_{x' \in X} [P_\theta(x, x') S_\theta(x, x') h_\theta(x') + P_\theta(x, x') \nabla_\theta h_\theta(x')].
\end{aligned}$$

Hence,

$$\begin{aligned}
\nabla_\theta \eta_\theta &= \lim_{k \to \infty} \mathsf{E} \left[ -\nabla_\theta h_\theta(x(k)) + \nabla_\theta c_\theta(x(k)) + S_\theta(x(k), x(k+1)) h_\theta(x(k)) + \nabla_\theta h_\theta(x(k+1)) \right] \\
&= \lim_{k \to \infty} \mathsf{E} \left[ \nabla_\theta c_\theta(X(k)) + S_\theta(X(k), X(k+1)) h_\theta(X(k)) \right]. \qquad \square
\end{aligned}$$

# Next course

- Learning techniques for stochastic control
  - Temporal difference:

$$0 = \mathsf{E}\big[-Q^\star(X(k),U(k)) + c(X(k),U(k)) + \gamma\underline{Q}^\star(X(k+1)) \mid \mathcal{F}_k\big] \tag{9.1}$$

  - Main challenge: estimate (9.1) from data consistently
  - Variance reduction

# References

[1] Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.