Reinforcement Learning Theory

From the Trajectory Measure to Foundation Policies

Melih Kandemir

Contents

List of Notation **viii**

0 Preliminaries **1**
0.1 Sets, Measures and Integrals 1
0.2 Kernels and the Ionescu–Tulcea Theorem 2
0.3 Conditioning and Martingales 4
0.4 Operators, Fixed Points and Neumann Series 5
0.5 Total Variation and Couplings 6
0.6 Concentration of Measure 7
Assumed without proof in this chapter 8

I Foundations **9**

1 Markov Chains I: Structure and Long-Run Behaviour **10**
1.1 Chains and Their Path Measures 10
1.2 Classification of States 11
1.3 Recurrence and Transience 12
1.4 Invariant Laws and Return Times 14
1.5 Reversibility 17
1.6 Convergence and Mixing 17
1.7 The Running Examples 20
Assumed without proof in this chapter 21

2 Markov Chains II: Potential Theory **22**
2.1 Killed Chains and Hitting Times 22
2.2 The Resolvent 25
2.3 The Abel Limit, the Deviation Matrix and the Laurent Expansion 26
2.4 The Running Examples 28
Assumed without proof in this chapter 29

3 Controlled Markov Processes and the Trajectory Measure **30**
3.1 Controlled Markov Processes and Policies 30
3.2 The Trajectory Measure 31
3.3 Returns and Values 33
3.4 Markov Policies Suffice for Marginals 34
3.5 The Running Examples 35
Assumed without proof in this chapter 35

4 Bellman Operators and Dynamic Programming **36**
4.1 Bellman Operators 36
4.2 Finite Horizon: Backward Induction 37
4.3 Discounted Infinite Horizon 38
4.4 The Performance-Difference Lemma and the Linear Program 40
4.5 The Simulation Lemma 42
4.6 The Running Examples 43
Assumed without proof in this chapter 43

5 Stochastic Shortest Paths and Average Reward **44**
5.1 Proper Policies and the Weighted Contraction 44
5.2 Average Reward and Blackwell Optimality 46
5.3 The Running Examples 47
Assumed without proof in this chapter 48

6 One Measure, Four Functionals **49**
6.1 Path Metrics and Couplings 49
6.2 The Trajectory Bellman Operator 51
6.3 Intertwining Functionals 53
6.4 The Four Functionals 54
6.5 The Running Examples 57
Assumed without proof in this chapter 58

II Learning to Control: The Classical Axis **59**

7 Stochastic Approximation **60**
7.1 Robbins–Monro Recursions 60
7.2 Asynchronous Contractions 61
7.3 The ODE Method 62
7.4 The Running Examples 63
Assumed without proof in this chapter 63

8 Temporal-Difference Learning and Q-Learning **64**
8.1 Temporal-Difference Learning 64
8.2 Q-Learning 66
8.3 Overestimation and Double Estimation 67
8.4 The Running Examples 69
Assumed without proof in this chapter 69

9 Approximate Dynamic Programming **70**
9.1 Linear Temporal-Difference Learning 70
9.2 Error Propagation in Approximate Value Iteration 72
9.3 Target Networks and Polyak Averaging 75
9.4 The Running Examples 77
Assumed without proof in this chapter 77

10 Policy-Gradient Theory **78**
10.1 The Policy-Gradient Theorem 78
10.2 Gradient Domination 80
10.3 The Natural Policy Gradient 81
10.4 The Running Examples 83
Assumed without proof in this chapter 83

III The Statistical Theory of Reinforcement Learning **84**

11 Lower Bounds and Bandits **85**
11.1 Divergences 85
11.2 Divergence of Trajectory Measures 87
11.3 Multi-Armed Bandits 88
11.4 The Running Examples 89
Assumed without proof in this chapter 90

12 Sample Complexity with a Generative Model **91**
12.1 The Plug-in Model 91
12.2 The Total Variance of the Return 93
12.3 The Sharp Bounds 94
12.4 The Running Examples 95
Assumed without proof in this chapter 95

13 Exploration by Optimism **96**
13.1 Episodic Regret . . . 96
13.2 The Sharp Results . . . 99
13.3 The Running Examples . . . 100
Assumed without proof in this chapter . . . 100

14 State Abstraction, Bisimulation and Block MDPs **101**
14.1 Abstractions and Their Hierarchy . . . 101
14.2 Bisimulation and Its Metric . . . 103
14.3 Block MDPs . . . 104
14.4 The Running Examples . . . 105
Assumed without proof in this chapter . . . 105

15 Linear and Low-Rank MDPs **106**
15.1 Linear MDPs . . . 106
15.2 Low-Rank MDPs and Bellman Rank . . . 108
15.3 The Running Examples . . . 110
Assumed without proof in this chapter . . . 110

16 Reward-Free Exploration **111**
16.1 The Protocol . . . 111
16.2 A Reward-Free Simulation Lemma . . . 111
16.3 Online Reward-Free Exploration . . . 113
16.4 The Running Examples . . . 114
Assumed without proof in this chapter . . . 114

IV Distributional Reinforcement Learning as Push-Forward **115**

17 Push-Forward Reinforcement Learning **116**
17.1 Random Rewards as Coordinates . . . 116
17.2 Three Layers of Push-Forward . . . 116
17.3 Moments and Closure . . . 118
17.4 Distributional Control . . . 119
17.5 Metrics on Path Space and Their Images . . . 120
17.6 Lipschitz Kernels . . . 121
17.7 Compounding of Model Error . . . 122
17.8 The Running Examples . . . 124
Assumed without proof in this chapter . . . 125

18 Push-Forward Policy Iteration **126**
18.1 Surrogates and Projections 126
18.2 Learning the Push-Forward from Samples 128
18.3 Risk 129
18.4 The Running Examples 130
Assumed without proof in this chapter 131

V Successor Measures and Temporal Geometry **132**

19 Successor Measures and Successor Features **133**
19.1 The Successor Measure as a Learning Target 133
19.2 Successor Features 134
19.3 The Running Examples 135
Assumed without proof in this chapter 136

20 Spectral Theory of the Successor Representation **137**
20.1 Hitting Times from the Deviation Matrix 137
20.2 Reversible Chains 138
20.3 Variational Characterisations and Representations 139
20.4 Eigenoptions 140
20.5 The Running Examples 141
Assumed without proof in this chapter 141

21 Low-Rank Structure and Entry-Wise Estimation **142**
21.1 Low-Rank Dynamics 142
21.2 Which Error Matters for Control 143
21.3 The Running Examples 144
Assumed without proof in this chapter 144

22 Forward–Backward Representations and Zero-Shot RL **145**
22.1 Exact FB Representations 145
22.2 Learning and Approximation 146
22.3 The Running Examples 147
Assumed without proof in this chapter 147

23 Hitting Times and Displacement Geometry **148**
23.1 Exact Displacement Representations 148
23.2 Approximate Closure 149
23.3 Symmetric Representations Cannot Be Directional 150

23.4 The Running Examples . . . 151
Assumed without proof in this chapter . . . 151

24 Quasimetric and Reachability Representations 152
24.1 Quasimetrics . . . 152
24.2 The Reachability Distance . . . 152
24.3 Quasimetric Learning . . . 154
24.4 The Running Examples . . . 155
Assumed without proof in this chapter . . . 155

25 Goal-Conditioned Reinforcement Learning 156
25.1 Goal-Augmented MDPs . . . 156
25.2 Shortest-Path Relaxations . . . 157
25.3 Hindsight Relabelling . . . 158
25.4 Goal Reaching as Classification . . . 158
25.5 The Running Examples . . . 160
Assumed without proof in this chapter . . . 160

26 Options and Hierarchical Reinforcement Learning 161
26.1 Options as Stopping Times . . . 161
26.2 Learning Options by Gradient . . . 163
26.3 Hierarchies . . . 165
26.4 The Running Examples . . . 166
Assumed without proof in this chapter . . . 166

VI Trajectory Measures and Foundation Policies 167

27 The Bridge: Occupancy Laws 168
27.1 The Occupancy Law . . . 168
27.2 One Law for All Rewards . . . 169
27.3 When the Occupancy Is the Trajectory . . . 170
27.4 The Running Examples . . . 171
Assumed without proof in this chapter . . . 171

28 What a Foundation Policy Must Compute 172
28.1 Four Properties . . . 172
28.2 What Each Object Computes . . . 173
28.3 Separations . . . 174
28.4 The Running Examples . . . 175

Assumed without proof in this chapter 175

29 The Trajectory-Measure Standard and Open Problems 176
29.1 The Hierarchy of Certificates 176
29.2 Open Problems 177
29.3 Closing 178
Assumed without proof in this chapter 178

List of Notation

Symbol	Meaning	First
Measures, kernels, norms		
$\mathscr{P}(E), \mathscr{M}(E), B_b(E)$	probability measures, finite signed measures, bounded measurable functions	0.1
$\langle \mu, f \rangle, \delta_x, f_\# \mu$	integral, Dirac measure, push-forward	0.1
$\mathrm{sp}(f)$	span seminorm $\sup f - \inf f$	0.1
$Kf, \mu K, KL, K^t$	kernel acting on functions, on measures, composition, powers	0.3
$\varrho(\mathsf{L})$	spectral radius	0.9
$\lVert \mu - \mu' \rVert_{\mathrm{TV}}, \Gamma(\mu, \mu')$	total variation, couplings	0.10
ε, δ	accuracy, confidence	0.12
Markov chains		
$\mathscr{X}, P, (X_t)_{t \in \mathbb{N}_0}$	state space, transition matrix, chain	1.1
$\mathbb{P}_\nu, \mathbb{P}_x, \mathbb{E}_x$	path measure from ν, from x	1.1
θ_s	shift on path space	1.1
τ_B, τ_B^+	hitting time ($t \geq 0$), return time ($t \geq 1$)	1.1
$x \to y, x \leftrightarrow y, d_x$	accessibility, communication, period	1.4
N_y, U	number of visits, potential kernel	1.6
q_B	hitting probability $\mathbb{P}_x\{\tau_B < \infty\}$	1.6
ϖ_x	excursion measure of a recurrent state	1.11
μ	invariant law	1.10
$\mathrm{dob}(P), \mathsf{d}(t), \bar{\mathsf{d}}(t), t_{\mathrm{mix}}$	Dobrushin coefficient, distances to equilibrium, mixing time	1.18
P_C, U_C	killed kernel on C, its potential	2.1
$\kappa_B, \kappa(x, y)$	mean hitting time; temporal quasimetric $\mathbb{E}_x \tau_y$	2.3, 2.5
q_B^γ	discounted reach value $\mathbb{E}_x \gamma^{\tau_B}$	2.3
$U_\gamma, \frac{1}{1-\gamma}$	resolvent $\sum_t \gamma^t P^t$, effective horizon	2.6
$P^\star, Z, D$	Abel limit, fundamental matrix $(I - P + P^\star)^{-1}$, deviation matrix	2.8, 2.9
$\bar{r}, w$	gain $P^\star r$, bias Dr	2.9
$\bar{\kappa}$	Kemeny's constant	2.11
Control		
$\mathscr{A}, \mathscr{Z} = \mathscr{X} \times \mathscr{A}, z = (x, a)$	actions, state-action pairs	3.1
$P(dx' \mid x, a), r, \gamma$	transition kernel, reward in $[0, 1]$, discount	3.1
$\mathscr{H}_t, \xi_t$	histories	3.2
$\pi; \Pi^{\mathrm{HR}}, \Pi^{\mathrm{MR}}, \Pi^{\mathrm{SR}}, \Pi^{\mathrm{SD}}$	policy; policy classes	3.2
$P_\pi, P \otimes \pi, r_\pi$	state kernel, state-action kernel, policy reward	3.2
$\Omega, Z_t = (X_t, A_t), \mathscr{F}_t$	path space $\mathscr{Z}^{\mathbb{N}_0}$, canonical process, filtration	3.3
$\mathbb{P}_\nu^\pi, \mathbb{P}_x^\pi, \mathbb{P}_z^\pi$	trajectory measure	3.3
G, G_H, R_t	return functional, H-step return, reward at t	3.5
$V^\pi, Q^\pi, V^\star, Q^\star, \pi^\star$	values, optimal values, optimal policy	3.5, 4.5

d_ν^π, $d_{\nu,\mathscr{X}}^\pi$	discounted occupancy on $\mathscr{Z}$, its state marginal	3.5
H, h	horizon, step within an episode	4.4
T^π, T^*, π_Q, $\mathbb{A}^\pi$	Bellman operators, greedy policy, advantage	4.1
K_ν	Bellman-flow polytope	4.9
The trajectory measure and its functionals		
d_γ, W, $\bar{\mathsf{W}}$	discounted path metric, transport cost, its supremum over z	6.1
W_1, $\bar{W}_1$	Wasserstein-1 distance, its supremum over z	6.1
cons_z, Φ^π	prepending map, trajectory Bellman operator	6.5
β_z, $\mathfrak{B}_F$, $\mathfrak{V}^F$	intertwining maps, Bellman operator on laws, law of F	6.7
η^π, $\mathscr{T}^\pi$	return law, distributional Bellman operator	6.9
O, M^π, $M_{\mathscr{X}}^\pi$	discounted occupancy of a path, successor measure, state successor measure	6.10
$\kappa^\pi(x,g)$, $q_g^{\gamma,\pi}$	mean hitting time and reach value under π	6.11
Learning and approximation		
ϑ, ϑ_k	a parameter (iterate, weight vector, estimand)	7.1
α_k, ζ_{k+1}	step sizes, martingale noise	7.1
δ_k, T_λ^π	temporal-difference error, λ-operator	8.1, 8.3
φ, Φ, Π_μ	features, feature matrix, projection in $\|\cdot\|_\mu$	9.1
β; $c(m)$, $C_{\nu,\beta}$	data law; concentrability coefficients	9.5
π_ϑ, Z_t	parametrised policy, NPG normaliser	10.1, 10.4
Statistical theory		
KL, kl	Kullback–Leibler divergence, Bernoulli case	11.1
R_n, $T_a(n)$, Δ_a	bandit regret, pulls, gaps	11.6
$\hat{P}$, n, N	empirical kernel, samples per pair, total samples	12.1
Σ^π, v^π	variance of the return, one-step variance	12.3
$\mathrm{Regret}(K)$, $T = KH$, b_k	episodic regret, steps, bonus	13.1, 13.2
$\mathsf{e}(\pi)$	occupancy-weighted model error	16.2
Push-forward theory		
$\mathscr{R}$, ρ, $\tilde{\mathscr{Z}}$	reward kernel, realised reward, extended pairs	17.1
F_R, ℓ_γ, L^π	reward process, discounted readout, its law	17.3
M_m, μ_j^π	moment functional, moments of the return	17.4
W_p, W_p	Wasserstein distances on (E,d) and on paths	18.1
S, P_D, ϵ_{rec}	encoder, decoder kernel, reconstruction error	17.13
Π_{C}, Π_{Q}, p_u	categorical and quantile projections, check loss	18.1, 18.5
CVaR_α	conditional value-at-risk	18.8
Representations		
m^π, ψ^π	successor density, successor features	19.1, 19.2
$\|\cdot\|_{\max}$, $\|\cdot\|_{2\to\infty}$	entry-wise and two-to-infinity norms	21.2
F, B, π_w, w_r	forward and backward maps, indexed policies, reward embedding	22.1
K, ω, $\hat{\kappa}_\omega$	latent operator, representer, displacement predictor	23.1, 23.2
d_γ, d_γ^*, s^*	reachability distances, shortest-path length	24.2–24.4
$\mathbb{O}$	space of occupancies	27.1

Added in draft 3

$J^\pi, J^\star, c, \mathsf{b}$	SSP cost-to-go, optimal cost, cost, worst expected time to the goal	5.1, 5.3
$\kappa^\star(x, g)$	optimal mean hitting time	5.5
$\bar{r}^\pi, w^\pi$	gain and bias of a policy	5.6
$\tilde{T}, \tilde{T}_2$	noisy optimality operator, double-estimation operator	8.9, 8.11
$\mathsf{C}_0, \mathsf{C}_1, \mathsf{G}, \mathsf{M}, \upsilon$	feature covariances, target regression map, its linear part, Polyak rate	9.9
$\chi, \operatorname{spec}(\mathsf{M})$	eigenvalue, spectrum	9.11
$\phi, \mathcal{X}, \mathsf{w}$	abstraction, abstract states, weighting	14.1
$d^\star, \mathsf{F}$	bisimulation metric and its operator	14.6
$\mathcal{S}, \mathsf{q}, f^\star$	latent states, emissions, decoder of a block MDP	14.7
Λ_k	Gram matrix of features	15.3
$\mathfrak{L}^F$	law of a functional F of the trajectory	6.7
f_i, χ_i	eigenvectors and eigenvalues of a reversible chain	20.2
$Q(z, g), p_+(g \mid z)$	goal-conditioned values, future-state law from time 1	25.1, 25.5
$o, \sigma_o, r^o, p^o, \mu$	option, its duration, option model, policy over options	26.1, 26.2
Ξ^π	occupancy law (distributional successor measure)	27.1

0

Preliminaries

This chapter fixes notation and collects the measure theory, operator theory and concentration inequalities that the book uses. It is the base chapter: results stated as *Facts* carry a pointer to a proof in the literature; the other results are proved. The notation fixed here, together with the List of Notation in the front matter, is binding for the whole book: a symbol introduced for a concept keeps that meaning in every later chapter, and no symbol carries two concepts.

0.1 Sets, Measures and Integrals

Definition 0.1 (Basic notation)**.**

(a) $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{N}_0 := \{0, 1, 2, \dots\}$, $\overline{\mathbb{R}} := [-\infty, \infty]$, $a \wedge b := \min(a, b)$, $a \vee b := \max(a, b)$, $a^+ := a \vee 0$. For a set B, $\mathbf{1}_B$ is its indicator and $|B|$ its cardinality; $\mathbf{1}$ is the constant function 1.

(b) A *measurable space* $(E, \mathscr{E})$ is a set with a σ-algebra. A countable set always carries its discrete σ-algebra 2^E. $(E, \mathscr{E})$ is *standard* if it is isomorphic (by a bijection measurable in both directions) to a Borel subset of $\mathbb{R}$ with its Borel σ-algebra; countable spaces, $\mathbb{R}^d$, and every complete separable metric space with its Borel σ-algebra are standard (Çınlar, 2011, §I.2).

(c) $\mathscr{P}(E)$ is the set of probability measures on $(E, \mathscr{E})$, $\mathscr{M}(E)$ the vector space of finite signed measures, $B_b(E)$ the space of bounded measurable $f : E \to \mathbb{R}$ with $\|f\|_\infty := \sup_x |f(x)|$, and $\delta_x \in \mathscr{P}(E)$ the Dirac measure at x. For $\mu \in \mathscr{M}(E)$ and $f \in B_b(E)$, $\langle \mu, f \rangle := \mu f := \int f \, d\mu$.

(d) For measurable $f : E \to F$ and a measure μ on E, the *image* or *push-forward* measure is $f_\# \mu(C) := \mu(f^{-1}(C))$, $C \in \mathscr{F}$.

(e) The *span seminorm* of $f \in B_b(E)$ is $\operatorname{sp}(f) := \sup_x f(x) - \inf_x f(x)$.

Fact 0.2 (Measure and integration)**.** *Let $(E, \mathscr{E}, \mu)$ be a measure space (Çınlar, 2011, Ch. I).*

(a) (Uniqueness.) Two finite measures that agree on a family of sets closed under finite intersections and generating $\mathscr{E}$, and on E, are equal.

(b) (Monotone convergence.) If $0 \le f_n \uparrow f$ pointwise, then $\int f_n \, d\mu \uparrow \int f \, d\mu$.

(c) (Dominated convergence.) If $f_n \to f$ pointwise and $|f_n| \le g$ with $\int g \, d\mu < \infty$, then $\int f_n \, d\mu \to \int f \, d\mu$.

(d) (Change of variables.) For measurable $f : E \to F$ and measurable $g \geq 0$ or $g \in B_b(F)$, $\int g\, d(f_\# \mu) = \int g \circ f\, d\mu$.

(e) (Tonelli and Fubini.) For σ-finite measures, iterated integrals of nonnegative (or product-integrable) functions may be exchanged.

0.2 Kernels and the Ionescu–Tulcea Theorem

Definition 0.3 (Kernels)**.** Let $(E, \mathscr{E})$, $(F, \mathscr{F})$ be measurable spaces.

(a) A *Markov kernel* from E to F is a map $K : E \times \mathscr{F} \to [0,1]$ such that $x \mapsto K(x, C)$ is measurable for every $C \in \mathscr{F}$ and $K(x, \cdot) \in \mathscr{P}(F)$ for every x. We write $K(dy|x) := K(x, dy)$. A kernel from E to E is a kernel *on* E; it is *substochastic* if $K(x, E) \leq 1$ instead of $= 1$.

(b) A kernel acts on functions from the left of the integral and on measures from the right of it:

$$Kf(x) := \int K(x, dy) f(y), \qquad \mu K(C) := \int \mu(dx) K(x, C),$$
$$KL(x, C) := \int K(x, dy) L(y, C),$$

for $f \in B_b(F)$ (or $f \geq 0$), $\mu \in \mathscr{M}(E)$, and a kernel L from F to G. On E, $K^0(x, \cdot) := \delta_x$ and $K^{t+1} := KK^t$.

(c) For $\mu \in \mathscr{P}(E)$, $\mu \otimes K \in \mathscr{P}(E \times F)$ is the measure with $(\mu \otimes K)(B \times C) = \int_B \mu(dx) K(x, C)$.

(d) If E, F are countable, a kernel is a matrix $K(x, y) := K(x, \{y\}) \geq 0$ with row sums 1 (a *stochastic matrix*); then Kf, μK, KL are the matrix–vector and matrix–matrix products with f a column and μ a row vector.

Proposition 0.4 (Calculus of kernels)**.** *Let K be a Markov kernel from E to F and L a Markov kernel from F to G.*

(a) $Kf \in B_b(E)$ with $\|Kf\|_\infty \leq \|f\|_\infty$ for $f \in B_b(F)$; $\mu K \in \mathscr{P}(F)$ for $\mu \in \mathscr{P}(E)$; KL is a Markov kernel from E to G.

(b) $(\mu K) f = \mu(Kf)$ and $(KL) f = K(Lf)$; composition is associative.

(c) K is monotone: $f \leq f'$ implies $Kf \leq Kf'$; and $K\mathbf{1} = \mathbf{1}$.

(d) For $\mu, \mu' \in \mathscr{M}(E)$, $\|\mu K - \mu' K\| \leq \|\mu - \mu'\|$ in the total-variation norm $\|\mu\| := \sup_{\|f\|_\infty \leq 1} |\mu f|$.

Proof.

(a) Measurability of Kf holds for indicators by definition, for simple f by linearity, and for bounded measurable f by monotone limits of simple

functions [0.2(b)]. $|Kf(x)| \le \int K(x,dy)|f(y)| \le \|f\|_\infty$. μK is countably additive by [0.2(b)], and $\mu K(F) = \mu(E) = 1$. $KL(\cdot, C) = K(L\mathbf{1}_C)$ is measurable by the first claim applied twice, and $KL(x,\cdot) = (\delta_x K)L$ is a probability measure.

(b) For $f = \mathbf{1}_C$ both identities are the definitions; extend by linearity and [0.2(b)]. Associativity: $((KL)M)f = K(L(Mf)) = (K(LM))f$ for all bounded f, hence as kernels.

(c) Integration against the probability $K(x,\cdot)$ is monotone and integrates $\mathbf{1}$ to 1.

(d) $|(\mu K - \mu' K)f| = |(\mu - \mu')(Kf)| \le \|\mu - \mu'\|$ for $\|f\|_\infty \le 1$, by (a) and (b).

∎

Theorem 0.5 (Ionescu-Tulcea; Ionescu Tulcea, 1949)**.** *Let $(E_t, \mathscr{E}_t)$, $t \in \mathbb{N}_0$, be measurable spaces, $\nu \in \mathscr{P}(E_0)$, and K_{t+1} a Markov kernel from $E_0 \times \cdots \times E_t$ to E_{t+1}. Let*

$$(\Omega, \mathscr{F}) := \bigotimes_{t\ge 0} (E_t, \mathscr{E}_t)$$

with coordinate maps $Y_t(\omega) := \omega_t$, and define $\pi_0 := \nu$, $\pi_t := \pi_{t-1} \otimes K_t$ on $E_0 \times \cdots \times E_t$. There is a unique $\mathbb{P} \in \mathscr{P}(\Omega)$ such that

$$\mathbb{P}\{(Y_0, \dots, Y_t) \in B\} = \pi_t(B) \qquad \text{for all } t \in \mathbb{N}_0,\ B \in \mathscr{E}_0 \otimes \cdots \otimes \mathscr{E}_t.$$

No topological assumption on the spaces is needed.

Proof. The proof follows Çınlar (2011, Thm. IV.4.7).

(1) *Set-up.* Write $y_{0:t} := (y_0, \dots, y_t)$. The cylinders $C = \{Y_{0:t} \in B\}$ form an algebra $\mathscr{F}_0$ that generates $\mathscr{F}$. $\mathbb{P}(C) := \pi_t(B)$ is well defined, because $\pi_{t+1}(B \times E_{t+1}) = \pi_t(B)$, and it is finitely additive on $\mathscr{F}_0$. By Carathéodory's extension theorem and [0.2(a)] (Çınlar, 2011, Thm. I.3.19) it suffices to prove continuity at $\emptyset$: $C_k \in \mathscr{F}_0$, $C_k \downarrow \emptyset$ imply $\mathbb{P}(C_k) \to 0$.

(2) *Conditional masses.* For C with base B at level t put $Q_m(y_{0:m}; C) := \mathbf{1}_B(y_{0:t})$ for $m \ge t$, and backward

$$Q_m(y_{0:m}; C) := \int K_{m+1}(y_{0:m}; dy)\, Q_{m+1}(y_{0:m}, y; C)$$

for $m < t$. These are measurable [0.4(a)] with values in $[0,1]$, do not depend on the level chosen for C, satisfy the recursion for all m, and $\mathbb{P}(C) = \int \nu(dy_0) Q_0(y_0; C)$.

(3) *Contradiction.* Suppose $\mathbb{P}(C_k) \downarrow c > 0$. Each $Q_m(\cdot; C_k)$ decreases in k. By dominated convergence [0.2(c)] there is y_0^* with $\lim_k Q_0(y_0^*; C_k) > 0$; inductively, by the recursion and [0.2(c)], there is y_{m+1}^* with $\lim_k Q_{m+1}(y_{0:m+1}^*; C_k) > 0$. If C_k has its base at level t_k, then $\mathbf{1}_{C_k}(y_0^*, y_1^*, \dots) = Q_{t_k}(y_{0:t_k}^*; C_k) > 0$, so $(y_0^*, y_1^*, \dots) \in \bigcap_k C_k = \emptyset$, a contradiction.

Uniqueness follows from [0.2(a)] since $\mathscr{F}_0$ is closed under intersections. ∎

Intuition. The kernels K_{t+1} say how to draw the next coordinate given the past. Ionescu-Tulcea glues these one-step laws into one measure on infinite sequences. Every object of this book (value, successor measure, return law, hitting time) is a functional of a measure built this way.

0.3 Conditioning and Martingales

Fact 0.6 (Conditional expectation and martingales)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $\mathscr{G} \subseteq \mathscr{F}$ a sub-σ-algebra, and Y integrable (Çınlar, 2011, Ch. IV–V).*

(a) There is an a.s. unique $\mathscr{G}$-measurable integrable $\mathbb{E}[Y \mid \mathscr{G}]$ with $\mathbb{E}[Y\mathbf{1}_B] = \mathbb{E}[\mathbb{E}[Y \mid \mathscr{G}]\mathbf{1}_B]$ for all $B \in \mathscr{G}$. It is linear and monotone, $\mathbb{E}[\mathbb{E}[Y \mid \mathscr{G}]] = \mathbb{E}Y$, $\mathbb{E}[ZY \mid \mathscr{G}] = Z\mathbb{E}[Y \mid \mathscr{G}]$ for bounded $\mathscr{G}$-measurable Z, and

$$\mathbb{E}[\mathbb{E}[Y \mid \mathscr{G}] \mid \mathscr{H}] = \mathbb{E}[Y \mid \mathscr{H}]$$

for $\mathscr{H} \subseteq \mathscr{G}$ (tower property).

(b) A filtration *is an increasing sequence $(\mathscr{F}_t)_{t \in \mathbb{N}_0}$ of sub-σ-algebras. A random time $\tau : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is a* stopping time *if $\{\tau = t\} \in \mathscr{F}_t$ for all t;*

$$\mathscr{F}_\tau := \{B : B \cap \{\tau = t\} \in \mathscr{F}_t \ \forall t\}.$$

(c) An adapted integrable process (M_t) is a martingale *if $\mathbb{E}[M_{t+1} \mid \mathscr{F}_t] = M_t$; then $D_t := M_t - M_{t-1}$ are* martingale differences*. (Optional stopping.) For bounded stopping times $\sigma \le \tau$, $\mathbb{E}[M_\tau] = \mathbb{E}[M_\sigma]$.*

Fact 0.7 (Strong law of large numbers)**.** *If $Y_1, Y_2, \dots$ are i.i.d. real random variables with $\mathbb{E}|Y_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^n Y_i \to \mathbb{E}Y_1$ almost surely. If $Y_i \ge 0$ and $\mathbb{E}Y_1 = \infty$, then $\frac{1}{n}\sum_{i=1}^n Y_i \to \infty$ almost surely (Çınlar, 2011, Thm. III.6.4, Prop. III.6.3).*

0.4 Operators, Fixed Points and Neumann Series

Theorem 0.8 (Banach fixed point; Banach, 1922)**.** *Let* $(\mathbb{V}, \|\cdot\|)$ *be a Banach space,* $C \subseteq \mathbb{V}$ *closed and nonempty, and* $\mathsf{T} : C \to C$ *a* contraction *with modulus* $c \in [0,1)$*:* $\|\mathsf{T}v - \mathsf{T}v'\| \le c\|v - v'\|$.

(a) T *has a unique fixed point* $v^* \in C$.

(b) For every $v_0 \in C$ *the iterates* $v_{k+1} := \mathsf{T}v_k$ *satisfy* $\|v_k - v^*\| \le c^k\|v_0 - v^*\|$ *and*

$$\|v_k - v^*\| \le \frac{c^k}{1-c}\|v_1 - v_0\|.$$

(c) For every $v \in C$, $\|v - v^*\| \le \frac{1}{1-c}\|v - \mathsf{T}v\|$.

Uniqueness and the first bound in (b) hold on any metric space; completeness is used only for existence.

Proof.

(a) By induction $\|v_{k+1} - v_k\| \le c^k\|v_1 - v_0\|$, so for $m > k$,

$$\|v_m - v_k\| \le \sum_{j \ge k} c^j\|v_1 - v_0\| = \frac{c^k}{1-c}\|v_1 - v_0\|.$$

The sequence is Cauchy, converges to some $v^* \in C$ (C closed), and $\mathsf{T}v^* = \lim \mathsf{T}v_k = v^*$ by continuity of T. If $\mathsf{T}v' = v'$, then $\|v^* - v'\| \le c\|v^* - v'\|$ forces $v' = v^*$.

(b)

$$\|v_k - v^*\| = \|\mathsf{T}v_{k-1} - \mathsf{T}v^*\| \le c\|v_{k-1} - v^*\|;$$

iterate. The second bound is the Cauchy estimate of (a) with $m \to \infty$.

(c)

$$\|v - v^*\| \le \|v - \mathsf{T}v\| + \|\mathsf{T}v - \mathsf{T}v^*\| \le \|v - \mathsf{T}v\| + c\|v - v^*\|.$$

∎

Proposition 0.9 (Neumann series)**.** *Let* L *be a bounded linear operator on a Banach space with operator norm* $\|\mathsf{L}\|$.

(a) If $\mathsf{L}^t \to 0$ *in operator norm, in particular if* $\|\mathsf{L}^m\| < 1$ *for some* $m \ge 1$*, then* $I - \mathsf{L}$ *is invertible and* $(I - \mathsf{L})^{-1} = \sum_{t \ge 0} \mathsf{L}^t$ *(norm-convergent).*

(b) If $\|\mathsf{L}\| \le c < 1$*, then* $\|(I - \mathsf{L})^{-1}\| \le \frac{1}{1-c}$.

(c) (Fact.) The spectral radius $\varrho(\mathsf{L}) := \lim_t \|\mathsf{L}^t\|^{1/t}$ *exists, is independent of the norm on a finite-dimensional space, and for a matrix equals the largest modulus of an eigenvalue;* $\mathsf{L}^t \to 0$ *iff* $\varrho(\mathsf{L}) < 1$ *(Puterman and Chan, 2025, App. B.5).*

Proof.

(a) If $\|\mathsf{L}^m\| =: c < 1$, then

$$\|\mathsf{L}^{qm+j}\| \le c^q \max_{j<m} \|\mathsf{L}^j\|,$$

so $\sum_t \|\mathsf{L}^t\| < \infty$ and $S := \sum_t \mathsf{L}^t$ converges. Telescoping,

$$(I - \mathsf{L}) \sum_{t<n} \mathsf{L}^t = \sum_{t<n} \mathsf{L}^t (I - \mathsf{L}) = I - \mathsf{L}^n \to I,$$

so S is a two-sided inverse. If only $\mathsf{L}^t \to 0$, then $\|\mathsf{L}^m\| < 1$ for some m.

(b) $\|\sum_t \mathsf{L}^t\| \le \sum_t c^t$.

(c) Standard.

■

0.5 Total Variation and Couplings

Definition 0.10 (Total variation and couplings)**.** Let E be countable and $\mu, \mu' \in \mathscr{P}(E)$.

(a)

$$\|\mu - \mu'\|_{\mathrm{TV}} := \sup_{B \subseteq E} |\mu(B) - \mu'(B)|.$$

(b) A *coupling* of μ, μ' is a $\varsigma \in \mathscr{P}(E \times E)$ with marginals μ and μ'; $\Gamma(\mu, \mu')$ denotes the set of couplings.

Proposition 0.11 (Total variation)**.** *Let E be countable and $\mu, \mu' \in \mathscr{P}(E)$.*

(a)

$$\begin{aligned} \|\mu - \mu'\|_{\mathrm{TV}} &= \frac{1}{2} \sum_x |\mu(x) - \mu'(x)| \\ &= \frac{1}{2} \sup_{\|f\|_\infty \le 1} |\mu f - \mu' f| \\ &= 1 - \sum_x \mu(x) \wedge \mu'(x). \end{aligned}$$

(b)

$$\|\mu - \mu'\|_{\mathrm{TV}} = \min_{\varsigma \in \Gamma(\mu, \mu')} \varsigma\{(x, y) : x \neq y\},$$

and the minimum is attained.

(c) For $f \in B_b(E)$, $|\mu f - \mu' f| \le \mathrm{sp}(f)\, \|\mu - \mu'\|_{\mathrm{TV}}$.

Proof.

(a) Let $B^* := \{\mu > \mu'\}$. For any B,

$$\mu(B) - \mu'(B) \le \mu(B \cap B^*) - \mu'(B \cap B^*) \le \mu(B^*) - \mu'(B^*),$$

and symmetrically for $\mu'(B) - \mu(B)$ with $\{\mu' > \mu\}$. Both maxima equal $\frac{1}{2}\sum_x |\mu(x) - \mu'(x)|$, since $\sum_x (\mu - \mu')(x) = 0$. For $\|f\|_\infty \le 1$, $|\sum_x f(x)(\mu - \mu')(x)| \le \sum_x |\mu - \mu'|(x)$, with equality at $f = \mathbf{1}_{B^*} - \mathbf{1}_{(B^*)^c}$. Finally

$$\sum_x |\mu - \mu'| = \sum_x (\mu + \mu' - 2\mu \wedge \mu') = 2 - 2\sum_x \mu \wedge \mu'.$$

(b) For a coupling and any B,

$$\mu(B) - \mu'(B) = \varsigma(B \times E) - \varsigma(E \times B) \le \varsigma\{x \in B, y \notin B\} \le \varsigma\{x \ne y\}.$$

Conversely let $p := \sum_x \mu \wedge \mu'(x)$ and $\epsilon := 1 - p = \|\mu - \mu'\|_{\mathrm{TV}}$. If $\epsilon = 0$ put $\varsigma(x,x) := \mu(x)$. Otherwise put $\varsigma(x,x) := \mu(x) \wedge \mu'(x)$ and $\varsigma(x,y) := (\mu - \mu')^+(x)(\mu' - \mu)^+(y)/\epsilon$ for $x \ne y$; the marginals are μ, μ' because $\sum_y (\mu' - \mu)^+(y) = \epsilon$, and the diagonal carries only $\mu \wedge \mu'$ (the supports of $(\mu - \mu')^+$ and $(\mu' - \mu)^+$ are disjoint). So $\varsigma\{x \ne y\} = \epsilon$.

(c) With $c := (\sup f + \inf f)/2$, $\|f - c\|_\infty = \mathrm{sp}(f)/2$ and $\mu f - \mu' f = \mu(f - c) - \mu'(f - c)$; apply (a).

■

0.6 Concentration of Measure

Fact 0.12 (Concentration inequalities)**.** *Let $\delta \in (0,1)$.*

(a) (Hoeffding (1963).) If $Y_1, \dots, Y_n$ are independent with $Y_i \in [a_i, b_i]$ a.s., then for $s > 0$,

$$\mathbb{P}\Big\{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i) \ge s\Big\} \le \exp\Big(-\frac{2s^2}{\sum_i (b_i - a_i)^2}\Big)$$

(Boucheron et al., 2013, Thm. 2.8). In particular, for $Y_i \in [0,b]$ i.i.d., with probability at least $1 - \delta$, $|\frac{1}{n}\sum_i Y_i - \mathbb{E}Y_1| \le b\sqrt{\ln(2/\delta)/(2n)}$.

(b) (Bernstein.) If $Y_1, \dots, Y_n$ are independent with $Y_i - \mathbb{E}Y_i \le c$ a.s. and $\sum_i \mathrm{Var}(Y_i) \le v$, then with probability at least $1 - \delta$,

$$\sum_i (Y_i - \mathbb{E}Y_i) \le \sqrt{2v\ln(1/\delta)} + \frac{c}{3}\ln(1/\delta)$$

(Boucheron et al., 2013, Thm. 2.10 and the remark after Cor. 2.11).

(c) *(Azuma–Hoeffding (1967).) If $(D_i)_{i\le n}$ are martingale differences for $(\mathcal{F}_i)$ and $D_i \in [\ell_i, \ell_i + M_i]$ with ℓ_i $\mathcal{F}_{i-1}$-measurable and constants M_i, then with probability at least $1-\delta$, $\sum_i D_i < \sqrt{\frac{1}{2}\sum_i M_i^2 \ln(1/\delta)}$ (Zhang, 2023, Thm. 13.4).*

(d) *(Freedman (1975).) If (D_i) are martingale differences with $D_i \le c$ a.s. and $\mathsf{V}_n := \sum_{i\le n} \mathbb{E}[D_i^2 \mid \mathcal{F}_{i-1}]$, then for every $\sigma > 0$, with probability at least $1-\delta$: either*

$$\sum_{i\le n} D_i < \sigma\sqrt{2\ln(1/\delta)} + \frac{c}{3}\ln(1/\delta),$$

or $\mathsf{V}_n > \sigma^2$ (Zhang, 2023, Thm. 13.6).

(e) *(Union bound.) If events $B_1, \dots, B_m$ each have probability at least $1-\delta/m$, then $\mathbb{P}(\bigcap_j B_j) \ge 1-\delta$.*

Intuition. Hoeffding pays for the range of the summands; Bernstein and Freedman pay for their variance and only a lower-order term for the range. The second kind of bound is what removes a factor $1/(1-\gamma)$ from the sample complexity of reinforcement learning in Part II.

Assumed without proof in this chapter

[0.2] measure and integration; [0.6] conditional expectation, stopping times, optional stopping; [0.7] the strong law of large numbers; [0.9(c)] the spectral-radius formula; [0.12] the concentration inequalities of Hoeffding, Bernstein, Azuma–Hoeffding and Freedman. Each item carries its source.

Part I

Foundations

1

Markov Chains I: Structure and Long-Run Behaviour

A stationary policy turns a controlled Markov process into a Markov chain (Chapter 3), so every reinforcement-learning quantity of this book is a functional of a chain. This chapter develops the properties of chains that reinforcement learning uses: classification of states, recurrence, invariant laws, convergence and mixing. The order follows Mannor et al. (2025, Ch. 4); the proofs follow Levin and Peres (2017, Ch. 1, 4) and close three gaps of that presentation: invariant laws exist without aperiodicity, convergence holds without diagonalisability, and the mixing time is defined uniformly in the initial state. Throughout, $\mathcal{X}$ is countable and P is a stochastic matrix on $\mathcal{X}$ [0.3(d)].

1.1 Chains and Their Path Measures

Definition 1.1 (Markov chain)**.** Let $\Omega := \mathcal{X}^{\mathbb{N}_0}$ with the product σ-algebra $\mathcal{F}$, coordinates $X_t(\omega) := \omega_t$ and natural filtration $\mathcal{F}_t := \sigma(X_0, \dots, X_t)$.

(a) For $\nu \in \mathcal{P}(\mathcal{X})$, $\mathbb{P}_\nu$ is the measure of Ionescu-Tulcea [0.5] with initial law ν and kernels $K_{t+1}(x_{0:t}; \cdot) := P(x_t, \cdot)$; $\mathbb{P}_x := \mathbb{P}_{\delta_x}$, and $\mathbb{E}_\nu$, $\mathbb{E}_x$ are the expectations. Under $\mathbb{P}_\nu$, (X_t) is the *Markov chain with transition matrix P and initial law ν*.

(b) The *shift* $\theta_s : \Omega \to \Omega$ is $(\theta_s\omega)_t := \omega_{s+t}$.

(c) For $B \subseteq \mathcal{X}$, the *hitting time* is $\tau_B := \inf\{t \geq 0 : X_t \in B\}$ and the *return time* $\tau_B^+ := \inf\{t \geq 1 : X_t \in B\}$ ($\inf \emptyset := \infty$); $\tau_x := \tau_{\{x\}}$, $\tau_x^+ := \tau_{\{x\}}^+$. Both are stopping times [0.6(b)].

Proposition 1.2 (Markov and strong Markov properties)**.**

(a)
$$\mathbb{P}_\nu\{X_0 = x_0, \dots, X_t = x_t\} = \nu(x_0)P(x_0, x_1)\cdots P(x_{t-1}, x_t);$$
hence $\mathbb{P}_\nu\{X_t = y\} = (\nu P^t)(y)$ *and* $\mathbb{E}_x f(X_t) = P^t f(x)$.

(b) (Markov property.) For every bounded measurable $F : \Omega \to \mathbb{R}$ and $s \in \mathbb{N}_0$, $\mathbb{E}_\nu[F \circ \theta_s \mid \mathcal{F}_s] = \mathbb{E}_{X_s} F$ a.s., where $\mathbb{E}_{X_s} F$ is $y \mapsto \mathbb{E}_y F$ evaluated at X_s.

(c) (Strong Markov property.) For a stopping time τ,
$$\mathbb{E}_\nu[F \circ \theta_\tau \mathbf{1}\{\tau < \infty\} \mid \mathcal{F}_\tau] = \mathbf{1}\{\tau < \infty\}\mathbb{E}_{X_\tau} F$$
a.s.

Proof.

(a) This is the defining property of [0.5] with $\pi_t = \nu \otimes P \otimes \cdots \otimes P$; sum over $x_0, \dots, x_{t-1}$.

(b) It suffices to check $\mathbb{E}_\nu[F \circ \theta_s \mathbf{1}_B] = \mathbb{E}_\nu[\mathbb{E}_{X_s} F \, \mathbf{1}_B]$ for $B = \{X_0 = x_0, \dots, X_s = x_s\}$, since these sets and $\emptyset$ form an intersection-closed generator of $\mathcal{F}_s$ and every $B \in \mathcal{F}_s$ is a countable disjoint union of them. For $F = \mathbf{1}\{X_0 = y_0, \dots, X_m = y_m\}$, both sides equal

$$\nu(x_0) \prod_{i<s} P(x_i, x_{i+1}) \, \mathbf{1}\{y_0 = x_s\} \prod_{j<m} P(y_j, y_{j+1})$$

by (a). The class of bounded F for which the identity holds is a vector space closed under bounded monotone limits [0.2(b)] containing the indicators of cylinders, so it contains all bounded measurable F (Çınlar, 2011, Thm. I.2.19).

(c) For $B \in \mathcal{F}_\tau$, $B \cap \{\tau = s\} \in \mathcal{F}_s$, and on $\{\tau = s\}$, $F \circ \theta_\tau = F \circ \theta_s$. By (b),

$$\mathbb{E}_\nu[F \circ \theta_\tau \mathbf{1}_{B \cap \{\tau = s\}}] = \mathbb{E}_\nu[\mathbb{E}_{X_s} F \, \mathbf{1}_{B \cap \{\tau = s\}}].$$

Sum over $s \in \mathbb{N}_0$ [0.2(c)]. ∎

Example 1.3 (A two-state chain; Mannor et al., 2025, Ex. 4.1)**.** On $\mathcal{X} = \{1, 2\}$ let

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}$$

and $\nu = (0.5, 0.5)$. Then $\nu P = (0.3, 0.7)$, $\nu P^2 = (0.26, 0.74)$, and $\mu := (0.25, 0.75)$ satisfies $\mu P = \mu$: $0.25 \cdot 0.4 + 0.75 \cdot 0.2 = 0.25$. The two rows of P differ by 0.2 in total variation; this number reappears as the convergence rate of νP^t in the convergence theorem of Section 1.6.

1.2 Classification of States

Definition 1.4 (Accessibility, classes, period)**.**

(a) y is *accessible* from x, $x \to y$, if $P^t(x, y) > 0$ for some $t \in \mathbb{N}_0$; x and y *communicate*, $x \leftrightarrow y$, if $x \to y$ and $y \to x$.

(b) A *class* is an equivalence class of $\leftrightarrow$. $C \subseteq \mathcal{X}$ is *closed* if $P(x, C) = 1$ for all $x \in C$; a closed one-point set $\{x\}$ is an *absorbing* state. P is *irreducible* if $\mathcal{X}$ is a single class.

(c) The *period* of x is $d_x := \gcd \mathbb{T}(x)$, $\mathbb{T}(x) := \{t \geq 1 : P^t(x, x) > 0\}$ ($d_x := \infty$ if $\mathbb{T}(x) = \emptyset$); x is *aperiodic* if $d_x = 1$, and P is aperiodic if all states are.

Lemma 1.5 (Structure of accessibility)**.**

(a) $P^{s+t}(x,z) \geq P^s(x,y)P^t(y,z)$*; hence* $\to$ *is transitive and* $\leftrightarrow$ *is an equivalence relation.*

(b) (Period is a class property; Mannor et al., 2025, Claim 4.1.) If $x \leftrightarrow y$ *and* $x \neq y$*, then* $d_x = d_y < \infty$*.*

(c) If $d_x = 1$*, there is* t_x *with* $P^t(x,x) > 0$ *for all* $t \geq t_x$*.*

(d) If $\mathcal{X}$ *is finite and* P *is irreducible and aperiodic, there is* m *with* $P^m(x,y) > 0$ *for all* x, y*.*

Proof.

(a) $P^{s+t}(x,z) = \sum_w P^s(x,w)P^t(w,z) \geq P^s(x,y)P^t(y,z)$.

(b) Choose $r, s \geq 1$ with $P^r(x,y) > 0$ and $P^s(y,x) > 0$ (possible as $x \neq y$). By (a), $r + s \in \mathbb{T}(x)$, and for $t \in \mathbb{T}(y)$ also $r + t + s \in \mathbb{T}(x)$. So $d_x < \infty$ divides $r + s$ and $r + s + t$, hence every $t \in \mathbb{T}(y)$, hence d_y. By symmetry $d_y \mid d_x$.

(c) $\mathbb{T}(x)$ is closed under addition by (a). Its gcd 1 is already the gcd of finitely many $a_1, \dots, a_m \in \mathbb{T}(x)$, so Bézout gives $1 = p - q$ with p, q nonnegative integer combinations of the a_i, hence in $\mathbb{T}(x) \cup \{0\}$. If $q = 0$, $1 \in \mathbb{T}(x)$ and $t_x = 1$. Otherwise every $t \geq q^2$ is $t = aq + b$ with $0 \leq b < q \leq a$, and $t = (a - b)q + bp \in \mathbb{T}(x)$.

(d) For x, y choose m_{xy} with $P^{m_{xy}}(x,y) > 0$; by (a) and (c), $P^t(x,y) \geq P^{t-m_{xy}}(x,x)P^{m_{xy}}(x,y) > 0$ for $t \geq t_x + m_{xy}$. Take $m := \max_{x,y}(t_x + m_{xy})$. ∎

1.3 Recurrence and Transience

Definition 1.6 (Recurrence and the potential kernel)**.**

(a) x is *recurrent* if $\mathbb{P}_x\{\tau_x^+ < \infty\} = 1$ and *transient* otherwise. A recurrent x is *positive recurrent* if $\mathbb{E}_x\tau_x^+ < \infty$ and *null recurrent* otherwise.

(b) $N_y := \sum_{t \geq 0} \mathbf{1}\{X_t = y\}$ is the *number of visits* to y, and $U(x,y) := \mathbb{E}_x N_y = \sum_{t \geq 0} P^t(x,y) \in [0, \infty]$ is the *potential kernel* (Green's function) of P.

(c) $q_B(x) := \mathbb{P}_x\{\tau_B < \infty\}$ is the *hitting probability* of B.

Theorem 1.7 (Recurrence through visits)**.** *Let* $f_x := \mathbb{P}_x\{\tau_x^+ < \infty\}$*.*

(a) $\mathbb{P}_x\{N_x \geq m + 1\} = f_x^m$ *for* $m \in \mathbb{N}_0$*.*

(b) x *is recurrent iff* $\mathbb{P}_x\{N_x = \infty\} = 1$ *iff* $U(x,x) = \infty$*. If* x *is transient,* N_x *is geometric under* $\mathbb{P}_x$ *and* $U(x,x) = 1/(1 - f_x) < \infty$*.*

(c) $U(x,y) = q_y(x)\,U(y,y)$ *for all* x, y. *In particular* $U(x,y) \le U(y,y)$, *and if* y *is transient, then* $U(x,y) < \infty$ *and* $P^t(x,y) \to 0$ *for every* x.

Proof.

(a) Let $\tau^{(0)} := 0$ and $\tau^{(m+1)} := \inf\{t > \tau^{(m)} : X_t = x\}$; these are stopping times and $\{N_x \ge m+2\} = \{\tau^{(m+1)} < \infty\}$ under $\mathbb{P}_x$. On $\{\tau^{(m)} < \infty\}$,

$$\tau^{(m+1)} = \tau^{(m)} + \tau_x^+ \circ \theta_{\tau^{(m)}}$$

and $X_{\tau^{(m)}} = x$, so by [1.2(c)] with $F = \mathbf{1}\{\tau_x^+ < \infty\}$,

$$\mathbb{P}_x\{\tau^{(m+1)} < \infty\} = \mathbb{P}_x\{\tau^{(m)} < \infty\} f_x.$$

Induction from $\mathbb{P}_x\{\tau^{(0)} < \infty\} = 1$.

(b) If $f_x = 1$, $\mathbb{P}_x\{N_x = \infty\} = \lim_m f_x^m = 1$ and $U(x,x) = \infty$. If $f_x < 1$,

$$\mathbb{E}_x N_x = \sum_{m \ge 0} \mathbb{P}_x\{N_x \ge m+1\} = \sum_m f_x^m = 1/(1-f_x).$$

(c) Under $\mathbb{P}_x$, $N_y = N_y \circ \theta_{\tau_y}$ on $\{\tau_y < \infty\}$ and $N_y = 0$ otherwise. By [1.2(c)] with $F = N_y \wedge k$ and monotone convergence in k [0.2(b)], $U(x,y) = \mathbb{P}_x\{\tau_y < \infty\}\mathbb{E}_y N_y$. If y is transient, the series $\sum_t P^t(x,y) = U(x,y) \le U(y,y) < \infty$ converges, so its terms tend to 0.

■

Proposition 1.8 (Recurrence is a class property)**.**

(a) If x *is recurrent and* $x \to y$, *then* $q_x(y) = 1$; *in particular* $y \to x$.

(b) If x *is recurrent and* $x \leftrightarrow y$, *then* y *is recurrent and* $q_y(x) = q_x(y) = 1$.

(c) A class containing a recurrent state is closed.

Proof.

(a) Let $y \ne x$ (otherwise trivial) and choose a path $x = x_0, x_1, \dots, x_s = y$ with $P(x_{i-1}, x_i) > 0$ and s minimal; then $x_i \ne x$ for $i \ge 1$, so $\epsilon := \mathbb{P}_x\{X_1 = x_1, \dots, X_s = y\} > 0$ is the probability of a path that reaches y before returning to x. On this event, $\tau_x^+ = s + \tau_x \circ \theta_s$. By [1.2(b)],

$$1 - f_x = \mathbb{P}_x\{\tau_x^+ = \infty\} \ge \epsilon\, \mathbb{P}_y\{\tau_x = \infty\}.$$

Since $f_x = 1$, $\mathbb{P}_y\{\tau_x = \infty\} = 0$.

(b) Choose r, s with $P^r(y,x) > 0$, $P^s(x,y) > 0$. By [1.5(a)],

$$U(y,y) \ge \sum_t P^{r+t+s}(y,y) \ge P^r(y,x)\,U(x,x)\,P^s(x,y) = \infty,$$

so y is recurrent by [1.7(b)]. Apply (a) twice.

(c) If x is recurrent, x' is in its class and $P(x', y) > 0$, then x' is recurrent by (b), $x' \to y$, and $y \to x'$ by (a); so y is in the class.

■

Theorem 1.9 (Decomposition of a finite chain)**.** *Let $\mathcal{X}$ be finite.*

(a) Every nonempty closed set contains a recurrent state.

(b) $\mathcal{X} = \mathsf{T} \cup C_1 \cup \cdots \cup C_M$ with $M \geq 1$, where $C_1, \dots, C_M$ are the closed classes, each consisting of recurrent states, and T is the set of transient states.

(c) For every ν, $\mathbb{P}_\nu$-a.s. the chain visits T finitely often, so it enters some C_j and stays there.

Proof.

(a) Let C be closed and $x \in C$. Then $\sum_{y \in C} U(x, y) = \sum_t P^t(x, C) = \infty$, and C is finite, so $U(x, y) = \infty$ for some $y \in C$. By [1.7(c)], $U(y, y) \geq U(x, y) = \infty$: y is recurrent.

(b) Classes of recurrent states are closed [1.8(c)]. Conversely a closed class contains a recurrent state by (a), hence consists of recurrent states [1.8(b)]. $\mathcal{X}$ is closed, so $M \geq 1$. Remaining states are transient by definition.

(c) For $y \in \mathsf{T}$, $\mathbb{E}_\nu N_y = \sum_x \nu(x) U(x, y) \leq U(y, y) < \infty$ by [1.7(c)], so $N_y < \infty$ a.s.; T is finite. After its last visit to T the chain is in $\bigcup_j C_j$, and the C_j are closed.

■

Intuition. In a finite chain, transient states are visited finitely often and the chain ends up in one closed class. For a goal-reaching problem with an absorbing goal g this is the statement that the goal is eventually reached, provided $\{g\}$ is the only closed class.

1.4 Invariant Laws and Return Times

Definition 1.10 (Invariant measures)**.** A measure $\varpi : \mathcal{X} \to [0, \infty)$ is *invariant* for P if $\varpi P = \varpi$; an invariant $\mu \in \mathcal{P}(\mathcal{X})$ is an *invariant law* (stationary distribution). If $X_0 \sim \mu$, then $X_t \sim \mu$ for all t [1.2(a)].

Theorem 1.11 (The excursion measure)**.** *Let x be recurrent and*

$$\varpi_x(y) := \mathbb{E}_x \sum_{t=0}^{\tau_x^+ - 1} \mathbf{1}\{X_t = y\} = \sum_{t \geq 0} \mathbb{P}_x\{X_t = y,\ \tau_x^+ > t\}.$$

Then $\varpi_x P = \varpi_x$, $\varpi_x(x) = 1$, and $\sum_y \varpi_x(y) = \mathbb{E}_x \tau_x^+$. Moreover $\varpi_x(y) < \infty$ for every y with $y \to x$.

Proof. Write $\tau := \tau_x^+$. Since $\{\tau > t\} \in \mathcal{F}_t$, the Markov property [1.2(b)] gives

$$\mathbb{P}_x\{X_t = y, \tau > t, X_{t+1} = z\} = \mathbb{P}_x\{X_t = y, \tau > t\} P(y,z).$$

Summing over y and t (nonnegative terms, [0.2(b)]),

$$\begin{aligned}(\varpi_x P)(z) &= \sum_{t \ge 0} \mathbb{P}_x\{X_{t+1} = z,\ \tau \ge t+1\} \\ &= \sum_{s \ge 1} \mathbb{P}_x\{X_s = z, \tau > s\} + \sum_{s \ge 1} \mathbb{P}_x\{X_s = z, \tau = s\}.\end{aligned}$$

The second sum is $\mathbb{P}_x\{X_\tau = z, \tau < \infty\} = \mathbf{1}\{z = x\}$, because $\tau < \infty$ a.s. and $X_\tau = x$. The first sum is $\varpi_x(z) - \mathbb{P}_x\{X_0 = z\} = \varpi_x(z) - \mathbf{1}\{z = x\}$, where the term $t = 0$ of $\varpi_x(z)$ is finite. Hence $\varpi_x P = \varpi_x$. Before τ the chain is at x only at $t = 0$, so $\varpi_x(x) = 1$, and $\sum_y \varpi_x(y) = \sum_t \mathbb{P}_x\{\tau > t\} = \mathbb{E}_x \tau$. Finally, if $P^t(y,x) > 0$, then $1 = \varpi_x(x) = (\varpi_x P^t)(x) \ge \varpi_x(y) P^t(y,x)$. ∎

Theorem 1.12 (Uniqueness of invariant measures)**.** *Let P be irreducible and recurrent, and $\varpi \neq 0$ an invariant measure. Then $\varpi(y) > 0$ for all y, and $\varpi = \varpi(x)\, \varpi_x$ for every x.*

Proof. If $\varpi(x) = 0$ for some x, then $0 = \varpi(x) \ge \varpi(y) P^t(y,x)$ for all y, t, and irreducibility gives $\varpi = 0$. So $\varpi > 0$; normalise $\varpi(x) = 1$. *Claim:* $\varpi \ge \varpi_x$. For $z \neq x$, expand

$$\varpi(z) = \varpi(x) P(x,z) + \sum_{y_0 \neq x} \varpi(y_0) P(y_0, z)$$

and substitute the same expansion for $\varpi(y_0)$, n times; dropping the last remainder (nonnegative),

$$\varpi(z) \ge \sum_{s=1}^{n} \sum_{y_1, \dots, y_{s-1} \neq x} P(x,y_1) \cdots P(y_{s-1}, z) = \sum_{s=1}^{n} \mathbb{P}_x\{X_s = z,\ \tau_x^+ \ge s\}.$$

As $z \neq x$, $\{X_s = z, \tau_x^+ \ge s\} = \{X_s = z, \tau_x^+ > s\}$; letting $n \to \infty$, $\varpi(z) \ge \varpi_x(z)$, and $\varpi(x) = 1 = \varpi_x(x)$. *Conclusion:* ϖ_x is finite [1.11], so $\varpi' := \varpi - \varpi_x \ge 0$ is an invariant measure [1.11] with $\varpi'(x) = 0$; by the first step $\varpi' = 0$. ∎

Theorem 1.13 (Invariant laws and Kac's formula)**.** *Let P be irreducible.*

(a) If some state is positive recurrent, then all states are, P has a unique invariant law μ, and $\mu(y) = 1/\mathbb{E}_y\tau_y^+ > 0$ for every y.
(b) If P is recurrent and some state is null recurrent, all states are null recurrent and there is no invariant law.
(c) If P is transient, there is no invariant law.

Proof.

(a) Let x be positive recurrent. Then P is recurrent [1.8(b)], and $\mu := \varpi_x/\mathbb{E}_x\tau_x^+$ is an invariant law [1.11]. Any invariant law is a multiple of ϖ_x [1.12], so μ is unique. For any y, $\mu = \mu(y)\varpi_y$ [1.12] with $\mu(y) > 0$; summing, $1 = \mu(y)\mathbb{E}_y\tau_y^+$.
(b) If an invariant law existed, then for the null-recurrent x, $\mu = \mu(x)\varpi_x$ with $\mu(x) > 0$ would have mass $\mu(x)\mathbb{E}_x\tau_x^+ = \infty$. The claim on all states follows from (a).
(c) If μ were an invariant law, then $\mu(y) = \sum_x \mu(x)P^t(x,y) \to 0$ by dominated convergence [0.2(c)] and [1.7(c)], for every y.

■

Corollary 1.14 (Finite irreducible chains)**.** *If $\mathcal{X}$ is finite and P is irreducible, then every state is positive recurrent and P has a unique invariant law μ, with $\mu(y) = 1/\mathbb{E}_y\tau_y^+ > 0$. No aperiodicity is needed.*

Proof. $\mathcal{X}$ is a closed class, so it consists of recurrent states [1.9(b)]. For any x, $\varpi_x(y) < \infty$ for every y [1.11], and $\mathcal{X}$ is finite, so $\mathbb{E}_x\tau_x^+ = \sum_y \varpi_x(y) < \infty$. Apply [1.13(a)]. ■

Example 1.15 (Simple random walk; Mannor et al., 2025, Ex. 4.2)**.** On $\mathcal{X} = \mathbb{Z}$ let $P(x, x \pm 1) = \frac{1}{2}$. P is irreducible with period 2. It is recurrent: $P^{2t}(0,0) = a_t := \binom{2t}{t}4^{-t}$, and $a_1 = \frac{1}{2}$, $a_{t+1} = a_t\frac{2t+1}{2t+2}$; by induction $a_t \geq \frac{1}{2\sqrt{t}}$, because $(2t+1)^2(t+1) \geq (2t+2)^2 t$. So $U(0,0) \geq \sum_t \frac{1}{2\sqrt{t}} = \infty$ [1.7(b)]. Counting measure $\varpi \equiv 1$ is invariant, since every column of P sums to 1. By [1.12] every invariant measure is a multiple of ϖ, so there is no invariant law, and the walk is null recurrent [1.13(b)].

Example 1.16 (Random walk with resets; Mannor et al., 2025, Ex. 4.3)**.** On $\mathcal{X} = \mathbb{N}_0$ let $P(i, i+1) = P(i, 0) = \frac{1}{2}$. P is irreducible and aperiodic ($P(0,0) > 0$). $\mu(i) := 2^{-(i+1)}$ is an invariant law: $(\mu P)(0) = \frac{1}{2}\sum_i \mu(i) = \frac{1}{2}$ and $(\mu P)(i+1) = \frac{1}{2}\mu(i)$. By [1.13(a)] all states are positive recurrent and, exactly, $\mathbb{E}_i\tau_i^+ = 2^{i+1}$. This sharpens the bound $\mathbb{E}_i\tau_i^+ \leq 2 + 2\cdot 2^i$ of Mannor et al. (2025).

1.5 Reversibility

Proposition 1.17 (Detailed balance)**.**

(a) If $\mu \in \mathcal{P}(\mathcal{X})$ satisfies $\mu(x)P(x,y) = \mu(y)P(y,x)$ for all x, y, then μ is invariant; P is then called reversible *with respect to μ.*

(b) (Discrete-time queue; Mannor et al., 2025, Ex. 4.4.) On $\mathbb{N}_0$ let $P(i,i+1) = \varpi_\uparrow$, $P(i+1,i) = \varpi_\downarrow$ and $P(i,i) = 1 - \varpi_\uparrow - \varpi_\downarrow \mathbf{1}\{i > 0\}$, with $\varpi_\uparrow, \varpi_\downarrow > 0$ and $\varpi_\uparrow + \varpi_\downarrow \le 1$. If $\varsigma := \varpi_\uparrow / \varpi_\downarrow < 1$, then $\mu(i) = (1-\varsigma)\varsigma^i$ is the unique invariant law, and P is reversible with respect to it.

Proof.

(a) $(\mu P)(y) = \sum_x \mu(y)P(y,x) = \mu(y)$.

(b) Detailed balance holds trivially for $|i-j| \neq 1$, and for $j = i+1$ it reads $\mu(i)\varpi_\uparrow = \mu(i+1)\varpi_\downarrow$, true for $\mu(i) \propto \varsigma^i$. Apply (a); P is irreducible, so μ is unique [1.13(a)].

■

1.6 Convergence and Mixing

Definition 1.18 (Dobrushin coefficient, mixing time)**.**

(a) The *Dobrushin coefficient* of a stochastic matrix is

$$\operatorname{dob}(P) := \sup_{x,y} \| P(x,\cdot) - P(y,\cdot) \|_{\mathrm{TV}} \in [0,1].$$

(b) If P has an invariant law μ: $\mathsf{d}(t) := \sup_x \| P^t(x,\cdot) - \mu \|_{\mathrm{TV}}$, $\bar{\mathsf{d}}(t) := \operatorname{dob}(P^t)$, and the *mixing time* is

$$t_{\mathrm{mix}}(\varepsilon) := \min\{ t : \mathsf{d}(t) \le \varepsilon \},$$

$t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4)$.

Proposition 1.19 (Contraction in total variation)**.** *For stochastic matrices P, P' on countable $\mathcal{X}$ and $\nu, \nu' \in \mathcal{P}(\mathcal{X})$:*

(a)

$$\| \nu P - \nu' P \|_{\mathrm{TV}} \le \operatorname{dob}(P)\, \| \nu - \nu' \|_{\mathrm{TV}};$$

(b)

$$\operatorname{dob}(PP') \le \operatorname{dob}(P)\, \operatorname{dob}(P');$$

(c) if μ is invariant for P, then $\mathsf{d}(t) \le \bar{\mathsf{d}}(t) \le 2\mathsf{d}(t)$, and $\bar{\mathsf{d}}$ is submultiplicative:

$$\bar{\mathsf{d}}(s+t) \le \bar{\mathsf{d}}(s)\bar{\mathsf{d}}(t).$$

Proof.

(a) Let $\epsilon := \|\nu - \nu'\|_{\mathrm{TV}} > 0$ (else trivial) and $\alpha := (\nu - \nu')^+/\epsilon$, $\beta := (\nu' - \nu)^+/\epsilon$, probability vectors by [0.11(a)], with $\nu - \nu' = \epsilon(\alpha - \beta)$. Then

$$\alpha P - \beta P = \sum_{x,y} \alpha(x)\beta(y)\big(P(x,\cdot) - P(y,\cdot)\big),$$

and the triangle inequality for $\frac{1}{2}\|\cdot\|_{\ell^1}$ [0.11(a)] gives

$$\|\nu P - \nu' P\|_{\mathrm{TV}} \le \epsilon \sum_{x,y} \alpha(x)\beta(y)\,\mathrm{dob}(P).$$

(b) Apply (a) to $\nu = P(x,\cdot)$, $\nu' = P(y,\cdot)$ and the matrix P'.

(c)

$$P^t(x,\cdot) - \mu = \sum_y \mu(y)\big(P^t(x,\cdot) - P^t(y,\cdot)\big)$$

gives the first inequality, the triangle inequality the second, and (b) the submultiplicativity.

■

Theorem 1.20 (Convergence theorem; Doeblin, 1938)**.** *Let P have an invariant law μ.*

(a) If $\mathrm{dob}(P^m) < 1$ for some $m \ge 1$, then for all t and ν,

$$\|\nu P^t - \mu\|_{\mathrm{TV}} \le \mathsf{d}(t) \le \mathrm{dob}(P^m)^{\lfloor t/m \rfloor},$$

and μ is the only invariant law.

(b) If $\mathcal{X}$ is finite and P is irreducible and aperiodic, then $\mathrm{dob}(P^m) < 1$ for some m. Hence $P^t(x,\cdot) \to \mu$ geometrically fast, uniformly in x.

Proof.

(a) By [1.19(a)] and $\mu = \mu P^t$,

$$\|\delta_x P^t - \mu P^t\|_{\mathrm{TV}} \le \mathrm{dob}(P^t) \le \mathrm{dob}(P^m)^{\lfloor t/m \rfloor},$$

using [1.19(b)] and $\mathrm{dob} \le 1$. The left bound follows by mixing over $x \sim \nu$ (convexity of $\|\cdot\|_{\mathrm{TV}}$). If μ' is invariant, $\|\mu' - \mu\|_{\mathrm{TV}} = \|\mu' P^t - \mu P^t\|_{\mathrm{TV}} \to 0$.

(b) By [1.5(d)], P^m has all entries positive. For two rows with full support,

$$\|P^m(x,\cdot) - P^m(y,\cdot)\|_{\mathrm{TV}} = 1 - \sum_z P^m(x,z) \wedge P^m(y,z) < 1$$

[0.11(a)]; finitely many pairs give $\mathrm{dob}(P^m) < 1$. An invariant law exists by [1.14]; apply (a).

∎

Corollary 1.21 (Mixing-time amplification)**.** *If P has an invariant law and $t_{\mathrm{mix}} < \infty$, then $\mathsf{d}(kt_{\mathrm{mix}}) \le 2^{-k}$ for $k \in \mathbb{N}$, and*

$$t_{\mathrm{mix}}(\varepsilon) \le \lceil \log_2(1/\varepsilon) \rceil \, t_{\mathrm{mix}}$$

for $\varepsilon \in (0,1)$.

Proof. With $t := t_{\mathrm{mix}}$, $\bar{\mathsf{d}}(t) \le 2\mathsf{d}(t) \le \frac{1}{2}$ and

$$\mathsf{d}(kt) \le \bar{\mathsf{d}}(kt) \le \bar{\mathsf{d}}(t)^k \le 2^{-k}$$

by [1.19(c)]. ∎

Remark 1.22 (On the mixing time)**.** Mannor et al. (2025, Sec. 4.3.2) define the mixing time τ for one fixed initial law s_0 by $\|s_0 P^{\tau} - \mu\|_{\mathrm{TV}} \le \frac{1}{4}\|s_0 - \mu\|_{\mathrm{TV}}$, and then apply the same τ to the initial law $s_0 P^{\tau}$ to conclude $\|s_0 P^{k\tau} - \mu\|_{\mathrm{TV}} \le 4^{-k}\|s_0 - \mu\|_{\mathrm{TV}}$. That step is not justified, because τ depends on the initial law. Definition [1.18(b)] takes the worst case over initial states, and submultiplicativity [1.19(c)] then gives the amplification [1.21], with rate 2^{-k} instead of 4^{-k}.

Theorem 1.23 (Ergodic theorem)**.** *Let P be irreducible and positive recurrent with invariant law μ, and $f \in B_b(\mathcal{X})$. For every initial law ν,*

$$\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) \longrightarrow \langle \mu, f \rangle \qquad \mathbb{P}_\nu\text{-almost surely.}$$

In particular the fraction of time spent in y tends to $\mu(y) = 1/\mathbb{E}_y \tau_y^+$, with no aperiodicity assumption. Consequently:

(a) (Cesàro convergence.) $\frac{1}{t}\sum_{s<t} P^s(x,y) \to \mu(y)$ for all x, y;

(b) (Periodic chains do not converge.) If the period is $d \ge 2$, then $P^t(x,x)$ does not converge as $t \to \infty$.

Proof. Write $f = f^+ - f^-$; it suffices to treat $f \ge 0$. Fix x. Let $\tau^{(0)} := \tau_x$, finite $\mathbb{P}_\nu$-a.s. by [1.8(a)], and

$$\tau^{(k+1)} := \tau^{(k)} + \tau_x^+ \circ \theta_{\tau^{(k)}}.$$

For $k \ge 1$ let $L_k := \tau^{(k)} - \tau^{(k-1)}$ and $Y_k := \sum_{s=\tau^{(k-1)}}^{\tau^{(k)}-1} f(X_s)$. *Regeneration:* by the strong Markov property [1.2(c)] at $\tau^{(k-1)}$, where $X_{\tau^{(k-1)}} = x$, the pair (L_k, Y_k) is independent of $\mathcal{F}_{\tau^{(k-1)}}$ and has the law of $(\tau_x^+, \sum_{s<\tau_x^+} f(X_s))$ under $\mathbb{P}_x$; since

$(L_j, Y_j)_{j<k}$ is $\mathcal{F}_{\tau^{(k-1)}}$-measurable, the pairs are i.i.d. Their means are $\mathbb{E}_x\tau_x^+ < \infty$ and $\langle \varpi_x, f\rangle \le \|f\|_\infty \mathbb{E}_x\tau_x^+$ [1.11]. By [0.7], $\frac{1}{n}\sum_{k\le n} L_k \to \mathbb{E}_x\tau_x^+$ and $\frac{1}{n}\sum_{k\le n} Y_k \to \langle \varpi_x, f\rangle$ a.s. Let $S_t := \sum_{s<t} f(X_s)$ and $n(t) := \max\{k : \tau^{(k)} \le t\}$ for $t \ge \tau^{(0)}$, so $n(t) \to \infty$ and $\tau^{(n(t))} \le t < \tau^{(n(t)+1)}$. As $f \ge 0$,

$$\frac{S_{\tau^{(0)}} + \sum_{k\le n(t)} Y_k}{\tau^{(0)} + \sum_{k\le n(t)+1} L_k} \le \frac{S_t}{t} \le \frac{S_{\tau^{(0)}} + \sum_{k\le n(t)+1} Y_k}{\tau^{(0)} + \sum_{k\le n(t)} L_k}.$$

Both bounds tend to $\langle \varpi_x, f\rangle / \mathbb{E}_x\tau_x^+ = \langle \mu, f\rangle$ [1.13(a)]. For (a) take $f = \mathbf{1}_{\{y\}}$, $\nu = \delta_x$, and integrate, by dominated convergence [0.2(c)], since $\frac{1}{t}\sum_{s<t} \mathbf{1}\{X_s = y\} \in [0,1]$. For (b), $P^t(x,x) = 0$ for $t \notin d\mathbb{N}$, so a limit would be 0, and then the Cesàro averages would tend to $0 \ne \mu(x)$ [1.13(a)], contradicting (a). ∎

1.7 The Running Examples

Example 1.24 (The three-state goal chain)**.** On $\mathcal{X} = \{1, 2, g\}$ let

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{(rows and columns ordered } 1, 2, g\text{)}.$$

The classes are $\mathsf{T} = \{1, 2\}$ and $C_1 = \{g\}$; g is absorbing, 1 and 2 are transient, and $d_1 = d_2 = 1$ since $P(1,1) > 0$. By [1.9(c)] the chain reaches g a.s. from every state. δ_g is the unique invariant law: if μ is invariant and y transient, $\mu(y) = \sum_x \mu(x)P^t(x,y) \to 0$ by [1.7(c)] and dominated convergence [0.2(c)]. Since g is absorbing, $P^t(x,g) = \mathbb{P}_x\{\tau_g \le t\}$, so $\|P^t(x,\cdot) - \delta_g\|_{\mathrm{TV}} = \mathbb{P}_x\{\tau_g > t\}$; Chapter 2 computes its tail and mean exactly. This chain, enriched with a second action, is the running example of the book.

Example 1.25 (The 5×5 gridworld under the uniform policy)**.** Let $\mathcal{X} = \{1, \dots, 5\}^2$. From x, each of the four compass moves is chosen with probability $\frac{1}{4}$; a move that would leave the grid leaves the state unchanged. Then $P(x,y) = \frac{1}{4}$ for neighbours y of x, $P(x,x) = \frac{1}{4} \cdot \#\{\text{blocked moves at } x\}$, and P is symmetric, so P is doubly stochastic and reversible with respect to the uniform law $\mu \equiv \frac{1}{25}$ [1.17(a)]. P is irreducible and aperiodic ($P(x,x) > 0$ at the boundary and [1.5(b)]). Hence $\mathbb{E}_x\tau_x^+ = 25$ for all x [1.14], and $P^t(x,\cdot) \to \mu$ geometrically [1.20(b)]. This also corrects the impression from Mannor's Theorem 4.7 that aperiodicity is needed for the invariant law: it is needed only for convergence [1.23(b)], not for existence, uniqueness or the ergodic averages [1.14], [1.23].

Assumed without proof in this chapter

Nothing beyond Chapter 0; the monotone-class theorem for functions is quoted from Çınlar (2011, Thm. I.2.19) in [1.2(b)].

2

Markov Chains II: Potential Theory

Reinforcement learning computes expectations of sums along a chain: discounted sums (values), sums up to a stopping time (goal-reaching times), and long-run averages. The operators that produce these sums are the *resolvent* $U_\gamma = \sum_t \gamma^t P^t$, the *killed potential* $U_C = \sum_t P_C^t$ and the *Abel limit* P^* with its deviation matrix D. This chapter develops them, following Puterman and Chan (2025, App. B), Kemeny and Snell (1960) and Levin and Peres (2017, Ch. 10–10). Two results go beyond these sources: mean hitting times form a quasimetric, the structure that Part V exploits, and the Laurent expansion of the resolvent holds without aperiodicity. $\mathcal{X}$ is countable, P a stochastic matrix on it.

2.1 Killed Chains and Hitting Times

Definition 2.1 (Killed kernels)**.** For $C \subseteq \mathcal{X}$ with complement $B := \mathcal{X} \setminus C$, the *killed kernel* P_C is the substochastic matrix $P_C(x,y) := P(x,y)$, $x, y \in C$, and $U_C := \sum_{t \geq 0} P_C^t$ is its potential kernel ($[0,\infty]$-valued).

Proposition 2.2 (Killed chains count visits before the target)**.** *Let $B = \mathcal{X} \setminus C$ and $x, y \in C$.*

(a) $P_C^t(x,y) = \mathbb{P}_x\{X_t = y,\ \tau_B > t\}$.

(b) $U_C(x,y) = \mathbb{E}_x \sum_{t<\tau_B} \mathbf{1}\{X_t = y\}$, *and* $\sum_{y\in C} U_C(x,y) = \mathbb{E}_x \tau_B$.

(c)

$$\mathbb{P}_x\{X_{\tau_B} = b,\ \tau_B < \infty\} = \sum_{y \in C} U_C(x,y) P(y,b)$$

for $b \in B$.

Proof.

(a) Induction on t; $t = 0$ is trivial.

$$\{X_{t+1} = y, \tau_B > t+1\} = \{X_t \in C, \tau_B > t, X_{t+1} = y\}$$

for $y \in C$, and by the Markov property [1.2(b)] its probability is

$$\sum_{w \in C} \mathbb{P}_x\{X_t = w, \tau_B > t\} P(w,y) = P_C^{t+1}(x,y).$$

(b) Sum (a) over t [0.2(b)]; then sum over $y \in C$: $\sum_t \mathbb{P}_x\{\tau_B > t\} = \mathbb{E}_x\tau_B$.

(c) $\{X_{\tau_B} = b, \tau_B = t+1\} = \{\tau_B > t, X_{t+1} = b\}$; by [1.2(b)] and (a) its probability is $\sum_{y\in C} P_C^t(x,y)P(y,b)$. Sum over t.

■

Theorem 2.3 (First-step analysis and minimality)**.** *Let $B \subseteq \mathcal{X}$, $C := \mathcal{X} \setminus B$, and $\gamma \in [0,1)$.*

(a) The hitting probability q_B is the minimal $[0,1]$-valued solution of $v = 1$ on B, $v = Pv$ on C.

(b) The mean hitting time $\kappa_B(x) := \mathbb{E}_x\tau_B$ is the minimal $[0,\infty]$-valued solution of the Poisson equation for hitting times*: $v = 0$ on B, $v = \mathbf{1} + Pv$ on C.*

(c) The discounted reach value *$q_B^\gamma(x) := \mathbb{E}_x\gamma^{\tau_B}$ (with $\gamma^\infty := 0$) is the unique bounded solution of $v = 1$ on B, $v = \gamma Pv$ on C; on C, $q_B^\gamma = (I - \gamma P_C)^{-1}\gamma P(\cdot, B)$.*

Proof. On $\{X_0 \in C\}$, $\tau_B = 1 + \tau_B \circ \theta_1$. Hence by the Markov property [1.2(b)], applied to the bounded functionals $\mathbf{1}\{\tau_B < \infty\}$, $\tau_B \wedge k$ (then $k \to \infty$ by [0.2(b)]) and γ^{τ_B}, the functions q_B, κ_B, q_B^γ solve the stated equations.

(a) Let $v \geq 0$ solve the system. *Claim:* $v(x) \geq \mathbb{P}_x\{\tau_B \leq t\}$ for all t. For $t = 0$ this is $v \geq \mathbf{1}_B$. If it holds for t, then for $x \in C$,

$$v(x) = \sum_y P(x,y)v(y) \geq \sum_y P(x,y)\mathbb{P}_y\{\tau_B \leq t\} = \mathbb{P}_x\{\tau_B \leq t+1\}$$

by the Markov property; for $x \in B$ it is trivial. Let $t \to \infty$.

(b) Let $v \geq 0$ solve the system. *Claim:* $v(x) \geq \mathbb{E}_x[\tau_B \wedge t]$. For $t = 0$ both sides are $\geq 0 = 0$. Inductively, for $x \in C$, $\tau_B \wedge (t+1) = 1 + (\tau_B \circ \theta_1) \wedge t$, so

$$\mathbb{E}_x[\tau_B \wedge (t+1)] = 1 + \sum_y P(x,y)\mathbb{E}_y[\tau_B \wedge t] \leq 1 + Pv(x) = v(x).$$

Let $t \to \infty$ [0.2(b)].

(c) On $B_b(C)$ with the sup norm, $\|\gamma P_C\| \leq \gamma < 1$, so $v_C = \gamma P_C v_C + \gamma P(\cdot, B)$ has the unique bounded solution $(I - \gamma P_C)^{-1}\gamma P(\cdot, B)$ [0.9(a),(b)].

■

Intuition. Hitting probabilities and hitting times satisfy the same one-step equations as values; what distinguishes them is the boundary condition on the target and, for $\gamma = 1$, the need to select the *minimal* solution. When the target is reachable from everywhere in a finite chain, the solution is unique and minimality is automatic, as the next theorem shows.

Theorem 2.4 (The fundamental matrix of an absorbing chain; Kemeny and Snell, 1960)**.** *Let $C \subseteq \mathcal{X}$ be finite with $|C| = m$, $B := \mathcal{X} \setminus C$, and suppose B is accessible from every $x \in C$. Let $\epsilon := \min_{x\in C} \mathbb{P}_x\{\tau_B \le m\}$.*

(a) $\epsilon > 0$ *and*

$$\sup_{x\in C} \mathbb{P}_x\{\tau_B > t\} \le (1-\epsilon)^{\lfloor t/m \rfloor}.$$

(b) $\varrho(P_C) < 1$, $I - P_C$ *is invertible, and* $U_C = (I-P_C)^{-1}$.

(c) $\kappa_B = U_C\mathbf{1} < \infty$ *on* C*, and it is the unique solution of* $(I-P_C)v = \mathbf{1}$ *on* C.

(d) The absorption probabilities are $\mathbb{P}_x\{X_{\tau_B} = b\} = (U_C P_{CB})(x,b)$*, where* P_{CB} *is the* $C \times B$ *block of* P*; they sum to* 1 *over* $b \in B$.

Proof.

(a) From $x \in C$ a shortest positive-probability path to B visits distinct states of C, so it has length at most m; hence $\mathbb{P}_x\{\tau_B \le m\} > 0$, and $\epsilon > 0$ as a minimum over the finite set C. By the Markov property at time km [1.2(b)],

$$\mathbb{P}_x\{\tau_B > (k+1)m\} = \mathbb{E}_x[\mathbf{1}\{\tau_B > km\}\mathbb{P}_{X_{km}}\{\tau_B > m\}] \le (1-\epsilon)\mathbb{P}_x\{\tau_B > km\}.$$

(b) The operator norm of P_C^t on $B_b(C)$ is $\max_x P_C^t(x,C) = \max_x \mathbb{P}_x\{\tau_B > t\}$ [2.2(a)], which is $\le 1-\epsilon < 1$ for $t = m$. Apply [0.9(a),(c)].

(c) By [2.2(b)] and (b), $\kappa_B = U_C\mathbf{1} = (I-P_C)^{-1}\mathbf{1}$, which is the unique solution. 2.2(c) with (b). The sum is $\mathbb{P}_x\{\tau_B < \infty\} = 1$ by (a).

∎

Theorem 2.5 (Mean hitting times form a quasimetric)**.** *Let $\kappa(x,y) := \mathbb{E}_x\tau_y \in [0,\infty]$.*

(a) $\kappa(x,x) = 0$*, and* $\kappa(x,y) \ge 1$ *for* $x \ne y$.

(b) (Triangle inequality.) $\kappa(x,z) \le \kappa(x,y) + \kappa(y,z)$ *for all* x,y,z.

(c) κ *is not symmetric in general, but the* commute time $\kappa(x,y)+\kappa(y,x)$ *is a (possibly infinite) metric.*

So κ is a quasimetric on $\mathcal{X}$ with values in $[0,\infty]$: the temporal distance of the chain.

Proof.

(a) $\tau_x = 0$ under $\mathbb{P}_x$; for $x \ne y$, $\tau_y \ge 1$ under $\mathbb{P}_x$.

(b) If $\kappa(x,y) = \infty$ there is nothing to show. Otherwise $\tau_y < \infty$ $\mathbb{P}_x$-a.s., and $X_{\tau_y} = y$. After time τ_y the chain hits z no later than $\tau_y + \tau_z \circ \theta_{\tau_y}$, so $\tau_z \le \tau_y + \tau_z \circ \theta_{\tau_y}$. By the strong Markov property [1.2(c)] (applied to $\tau_z \wedge k$, then $k \to \infty$), $\mathbb{E}_x\tau_z \le \mathbb{E}_x\tau_y + \mathbb{E}_y\tau_z$.

(c) Symmetric by construction; the triangle inequality follows from (b) applied in both directions; positivity off the diagonal from (a). The examples at the end of this chapter show that κ itself is asymmetric, even for reversible chains.

■

2.2 The Resolvent

Definition 2.6 (Resolvent)**.** For $\gamma \in [0,1)$, the *resolvent* of P is the kernel $U_\gamma := \sum_{t\ge 0} \gamma^t P^t$, i.e. $U_\gamma(x,y) = \mathbb{E}_x \sum_{t\ge 0} \gamma^t \mathbf{1}\{X_t = y\}$. The quantity $1/(1-\gamma)$ is the *effective horizon*.

Proposition 2.7 (Calculus of the resolvent)**.** *Let $\gamma, \beta \in [0,1)$ and let P' be a second stochastic matrix on $\mathcal{X}$, with resolvent U'_γ.*

(a) On $B_b(\mathcal{X})$, $U_\gamma = (I-\gamma P)^{-1}$ and $\|U_\gamma\| = \frac{1}{1-\gamma}$. $(1-\gamma)U_\gamma$ is a stochastic matrix: $(1-\gamma)U_\gamma(x,\cdot)$ is the law of X_σ under $\mathbb{P}_x$, where σ is independent of the chain with $\mathbb{P}\{\sigma = t\} = (1-\gamma)\gamma^t$.

(b) For $f \in B_b(\mathcal{X})$, $U_\gamma f(x) = \mathbb{E}_x \sum_t \gamma^t f(X_t)$ and $\operatorname{sp}(U_\gamma f) \le \frac{\operatorname{sp}(f)}{1-\gamma}$.

(c) (Resolvent identity.) $U_\gamma - U_\beta = (\gamma - \beta)\, U_\gamma P U_\beta$.

(d) (Perturbation identity.) $U'_\gamma - U_\gamma = \gamma\, U'_\gamma (P' - P) U_\gamma$.

(e) (Perturbation bound.) With $\epsilon_P := \sup_x \|P'(x,\cdot) - P(x,\cdot)\|_{\mathrm{TV}}$ and $f \in B_b(\mathcal{X})$,

$$\|U'_\gamma f - U_\gamma f\|_\infty \le \frac{\gamma}{1-\gamma}\, \|(P'-P)U_\gamma f\|_\infty \le \frac{\gamma\, \epsilon_P\, \operatorname{sp}(f)}{(1-\gamma)^2}.$$

Proof.

(a) $\|\gamma P\| = \gamma$ on $B_b(\mathcal{X})$ [0.4(a)]; apply [0.9(a),(b)]. $U_\gamma \mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$ gives the norm. For the sampling interpretation, $\mathbb{P}_x\{X_\sigma = y\} = \sum_t (1-\gamma)\gamma^t P^t(x,y)$ by independence.

(b) Sum [1.2(a)] against γ^t and exchange sum and expectation [0.2(c)]. $\operatorname{sp}(P^t f) \le \operatorname{sp}(f)$ because $P^t f(x)$ is an average of values of f; hence $\operatorname{sp}(U_\gamma f) \le \sum_t \gamma^t \operatorname{sp}(f)$.

(c) $U_\gamma(I-\beta P)U_\beta = U_\gamma$ and $U_\gamma(I-\gamma P)U_\beta = U_\beta$; subtracting, $(\gamma-\beta)U_\gamma P U_\beta = U_\gamma - U_\beta$.

(d) Similarly $U'_\gamma(I-\gamma P)U_\gamma = U'_\gamma$ and $U'_\gamma(I-\gamma P')U_\gamma = U_\gamma$; subtract.

(e) By (d) and $\|U'_\gamma\| = \frac{1}{1-\gamma}$; then [0.11(c)] row by row, and (b).

■

2.3 The Abel Limit, the Deviation Matrix and the Laurent Expansion

Theorem 2.8 (The Abel limit)**.** *Let $\mathcal{X}$ be finite. Then $P^* := \lim_{\gamma\uparrow 1}(1-\gamma)U_\gamma$ exists, and:*

(a) P^* is a stochastic matrix with $PP^* = P^*P = P^*P^* = P^*$;

(b) $Z := (I - P + P^*)^{-1}$ exists and $P^*Z = P^*$;

(c) (Cesàro limit.) $\frac{1}{T}\sum_{t<T} P^t = P^* + \frac{1}{T}(I - P^T)Z \to P^*$;

(d) every row of P^* is an invariant law; if P is irreducible, every row equals the unique invariant law μ; if moreover P is aperiodic, $P^t \to P^*$.

Proof. Existence. By Cramer's rule each entry of $(I-\gamma P)^{-1}$ is a ratio of two polynomials in γ, with denominator $\det(I-\gamma P) \neq 0$ on $[0,1)$ [2.7(a)]. So $R(\gamma) := (1-\gamma)U_\gamma(x,y)$ is a rational function with values in $[0,1]$ on $[0,1)$. Write $R(\gamma) = (1-\gamma)^j c(\gamma)$ with $j \in \mathbb{Z}$ and c rational, continuous and nonzero at $\gamma = 1$ (or $R \equiv 0$). Boundedness forces $j \geq 0$, so $\lim_{\gamma\uparrow 1} R(\gamma)$ exists.

(a) Limits of stochastic matrices are stochastic. From $(I-\gamma P)U_\gamma = I = U_\gamma(I-\gamma P)$, multiplying by $1-\gamma$ and letting $\gamma \uparrow 1$ gives $P^* - PP^* = 0 = P^* - P^*P$. Hence $P^tP^* = P^*$, so $U_\gamma P^* = P^*/(1-\gamma)$ and $(1-\gamma)U_\gamma P^* = P^*$; let $\gamma \uparrow 1$: $P^*P^* = P^*$.

(b) If $(I - P + P^*)v = 0$, apply P^* and use (a): $P^*v = 0$, hence $v = Pv$, hence $v = (1-\gamma)U_\gamma v \to P^*v = 0$. $\mathcal{X}$ is finite, so $I - P + P^*$ is invertible. From (a), $P^*(I - P + P^*) = P^*$; multiply by Z.

(c) $\sum_{t<T} P^t(I-P) = I - P^T$ and $\sum_{t<T} P^tP^* = TP^*$; so $\sum_{t<T} P^t(I-P+P^*) = I - P^T + TP^*$. Multiply by Z and use (b). P^T is bounded.

(d) By (a), $P^*(x,\cdot)P = P^*(x,\cdot)$. Uniqueness [1.14] and the convergence theorem [1.20(b)].

■

Definition 2.9 (Deviation matrix, gain and bias)**.** For finite $\mathcal{X}$, the *deviation matrix* is $D := Z - P^* = (I - P + P^*)^{-1} - P^*$. For $r \in B_b(\mathcal{X})$ the *gain* is $\bar{r} := P^*r$ and the *bias* is $w := Dr$.

Theorem 2.10 (Laurent expansion of the resolvent; Blackwell, 1962)**.** *Let $\mathcal{X}$ be finite. Then for $\gamma \in [0,1)$,*

$$U_\gamma = \frac{P^*}{1-\gamma} + D + E(\gamma), \qquad \|E(\gamma)\| \leq c_P\,(1-\gamma),$$

with a constant c_P depending only on P. No aperiodicity is required. Moreover:

(a) $(I-P)D = D(I-P) = I - P^*$ *and* $P^*D = DP^* = 0$*;*

(b) (Poisson equation.) For $r \in B_b(\mathcal{X})$*, the bias* $w = Dr$ *solves* $w + \bar{r} = r + Pw$ *with* $P^*w = 0$*, and it is the only solution with* $P^*w = 0$*;*

(c) for every r*,* $U_\gamma r = \frac{\bar{r}}{1-\gamma} + w + O(1-\gamma)$*, so* $(1-\gamma)U_\gamma r \to \bar{r}$ *and* $U_\gamma r - \frac{\bar{r}}{1-\gamma} \to w$*.*

Proof. Let $Q := P - P^*$. By [2.8(a)], $(P^t - P^*)Q = P^{t+1} - P^*$, so $Q^t = P^t - P^*$ for $t \geq 1$ (induction). Hence, for $\gamma < 1$,

$$U_\gamma = \sum_t \gamma^t P^* + \sum_t \gamma^t (P^t - P^*) = \frac{P^*}{1-\gamma} - P^* + \sum_{t \geq 0} \gamma^t Q^t.$$

The last series converges, because Q^t is bounded, and $(I-\gamma Q)\sum_{t<n} \gamma^t Q^t = I - \gamma^n Q^n \to I$; so it equals $Z_\gamma := (I-\gamma Q)^{-1}$. At $\gamma = 1$, $I - Q = I - P + P^*$ is invertible [2.8(b)]. So $I - \gamma Q$ is invertible for every $\gamma \in [0,1]$, $\gamma \mapsto Z_\gamma$ is continuous on the compact $[0,1]$ (inversion is continuous), and $c := \sup_{\gamma \in [0,1]} \|Z_\gamma\| < \infty$. From

$$Z_\gamma - Z_1 = Z_\gamma\big[(I-Q) - (I-\gamma Q)\big]Z_1 = -(1-\gamma)Z_\gamma Q Z_1$$

we get $E(\gamma) := Z_\gamma - Z_1$ with $\|E(\gamma)\| \leq (1-\gamma)c^2\|Q\|$, and $U_\gamma = \frac{P^*}{1-\gamma} + (Z_1 - P^*) + E(\gamma)$ with $Z_1 = Z$.

(a) $(I-P)Z = (I - P + P^*)Z - P^*Z = I - P^*$ [2.8(b)], and $(I-P)P^* = 0$. Symmetrically $Z(I-P+P^*) = I$ gives $Z(I-P) = I - ZP^*$, and $ZP^* = P^*$ because $(I - P + P^*)P^* = P^*$. Finally $P^*D = P^*Z - P^* = 0$ and $DP^* = ZP^* - P^* = 0$.

(b) Apply (a) to r: $(I-P)w = r - \bar{r}$ and $P^*w = 0$. If w' is another solution with $P^*w' = 0$, then $v := w - w'$ satisfies $v = Pv$, hence $v = (1-\gamma)U_\gamma v \to P^*v = 0$.

(c) Apply the expansion to r.

■

Intuition. For γ close to 1, the discounted value $U_\gamma r$ consists of a large part $\bar{r}/(1-\gamma)$, the long-run reward rate times the effective horizon, and a finite correction w that records how much better or worse each start state is in the transient phase. The bias solves the same kind of Poisson equation as the mean hitting time [2.3(b)]; for an absorbing goal and unit cost the two coincide (the first running example below).

Theorem 2.11 (Random-target lemma; Levin and Peres, 2017, Lemma 10.1)**.** *Let* $\mathcal{X}$ *be finite and* P *irreducible with invariant law* μ*, and* $\kappa(x,y) = \mathbb{E}_x \tau_y$*. Then* $\bar{\kappa} := \sum_y \mu(y)\kappa(x,y)$ *does not depend on* x *(Kemeny's constant).*

Proof. Let $h(x) := \sum_y \mu(y)\kappa(x,y)$, finite by [2.4(c)]. For $y \neq x$, $\kappa(x,y) = 1 + \sum_z P(x,z)\kappa(z,y)$ [2.3(b)], and $1 + \sum_z P(x,z)\kappa(z,x) = \mathbb{E}_x \tau_x^+ = 1/\mu(x)$ by the Markov property and [1.14]. Hence

$$\begin{aligned} Ph(x) &= \sum_{y \neq x} \mu(y)\big(\kappa(x,y) - 1\big) + \mu(x)\Big(\frac{1}{\mu(x)} - 1\Big) \\ &= h(x) - (1 - \mu(x)) + 1 - \mu(x) = h(x). \end{aligned}$$

So $h = Ph$, hence $h = (1-\gamma)U_\gamma h \to P^* h$, and every row of P^* is μ [2.8(d)]: $h \equiv \langle \mu, h \rangle$. ∎

2.4 The Running Examples

Example 2.12 (Potential theory of the three-state goal chain)**.** For the chain of [1.24], take the target $B = \{g\}$, $C = \{1,2\}$:

$$P_C = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \qquad U_C = (I - P_C)^{-1} = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}, \qquad \kappa_g = U_C \mathbf{1} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}.$$

$\varrho(P_C) = \frac{1+\sqrt{5}}{4} \approx 0.809$, so $\mathbb{P}_x\{\tau_g > t\}$ decays like 0.809^t [2.4(a),(b)]. The Abel limit has every row equal to δ_g, and

$$D = \begin{pmatrix} 4 & 2 & -6 \\ 2 & 2 & -4 \\ 0 & 0 & 0 \end{pmatrix}.$$

For the unit cost $r := \mathbf{1}_C$ (one unit per step before the goal), the gain is $\bar{r} = P^* \mathbf{1}_C = 0$ and the bias is $w = D\mathbf{1}_C = (6,4,0)$: the bias of the unit cost *is* the mean hitting time, and the Poisson equation [2.10(b)] reduces to the hitting-time equation [2.3(b)]. With $\gamma = 0.9$ the discounted reach value is

$$q_g^\gamma = (I - \gamma P_C)^{-1} \gamma P(\cdot, g) \approx (0.583, 0.712)$$

on C [2.3(c)].

Example 2.13 (Asymmetry of temporal distance)**.**

(a) In the three-state chain, $\kappa(1,2) = 2$ (from 1 the chain moves to 2 with probability $\frac{1}{2}$ per step), but $\kappa(2,1) = \infty$, since from 2 the chain is absorbed at g with probability $\frac{1}{2}$ before returning to 1. Also $\kappa(1,g) = 6 = \kappa(1,2) + \kappa(2,g)$: the triangle inequality [2.5(b)] is tight because every path from 1 to g passes through 2.

(b) In the 5×5 gridworld of [1.25], which is reversible, mean hitting times are still asymmetric: from a corner to the centre $\kappa = 36$, from the centre to a corner $\kappa \approx 90.74$. Kemeny's constant [2.11] is $\bar{\kappa} \approx 54.08$.

Reversibility makes the *commute time* symmetric, not the hitting time. A representation of temporal distance that is symmetric by construction cannot express either effect.

Assumed without proof in this chapter

Nothing beyond Chapter 0 and Cramer's rule for the inverse of a matrix.

3

Controlled Markov Processes and the Trajectory Measure

A controlled Markov process together with a policy determines one probability measure on the space of infinite trajectories, the *trajectory measure*. Every quantity studied in this book (values, successor measures, return laws, hitting times) is a functional of it. This chapter constructs the measure, identifies the Markov chains it induces, and shows that, as far as the law of a single time step is concerned, history-dependent policies can be replaced by Markov ones. The model follows Puterman and Chan (2025, Ch. 2) and Mannor et al. (2025, Ch. 6); the construction by the Ionescu-Tulcea theorem is the one indicated by Bellemare et al. (2023, Remark 2.1) and carried out by Akgül et al. (2026, App. A.7).

3.1 Controlled Markov Processes and Policies

Definition 3.1 (Controlled Markov process)**.** A *controlled Markov process* (CMP) is a triple $(\mathcal{X}, \mathcal{A}, P)$ of standard measurable spaces $\mathcal{X}$ (*states*) and $\mathcal{A}$ (*actions*) and a Markov kernel P from $\mathcal{Z} := \mathcal{X} \times \mathcal{A}$ to $\mathcal{X}$ (the *transition kernel*), written $P(dx'|x,a)$. A *Markov decision process* (MDP) $(\mathcal{X}, \mathcal{A}, P, r, \gamma)$ adds a measurable *reward* $r : \mathcal{Z} \to [0,1]$ and a *discount factor* $\gamma \in [0,1)$ (Bellman, 1957). Points of $\mathcal{Z}$ are written $z = (x,a)$. A Markov chain is a CMP with one action. Unless declared otherwise, $\mathcal{X}$ and $\mathcal{A}$ are countable; $|\mathcal{X}|$ and $|\mathcal{A}|$ denote their sizes when finite.

Definition 3.2 (Histories and policies)**.**

(a) The *history space* at time t is $\mathcal{H}_t := \mathcal{Z}^t \times \mathcal{X}$; a history is $\xi_t = (x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t)$.

(b) A *policy* is a sequence $\pi = (\pi_t)_{t \geq 0}$ of Markov kernels π_t from $\mathcal{H}_t$ to $\mathcal{A}$; Π^{HR} denotes the set of all policies (history-dependent, randomised).

(c) π is *Markov* if $\pi_t(da|\xi_t)$ depends on ξ_t only through x_t, written $\pi_t(da|x_t)$; Π^{MR}. A Markov policy is *stationary* if $\pi_t = \pi_0 =: \pi$ for all t, so π is a Markov kernel from $\mathcal{X}$ to $\mathcal{A}$; Π^{SR}. A stationary policy is *deterministic* if $\pi(\cdot|x) = \delta_{f(x)}$ for a measurable $f : \mathcal{X} \to \mathcal{A}$; then $\pi(x) := f(x)$; Π^{SD}. Thus

$$\Pi^{\mathrm{SD}} \subseteq \Pi^{\mathrm{SR}} \subseteq \Pi^{\mathrm{MR}} \subseteq \Pi^{\mathrm{HR}}.$$

(d) For a stationary π, the *state kernel* and *state-action kernel* are

$$P_\pi(dx'|x) := \int \pi(da|x)P(dx'|x,a),$$
$$(P \otimes \pi)(dx'\,da'|x,a) := P(dx'|x,a)\,\pi(da'|x')$$

Markov kernels on $\mathcal{X}$ and on $\mathcal{E}$, and the *policy reward* is $r_\pi(x) := \int \pi(da|x)r(x,a)$.

3.2 The Trajectory Measure

Theorem 3.3 (The trajectory measure; Ionescu Tulcea, 1949, Akgül et al., 2026)**.** *Let $(\mathcal{X},\mathcal{A},P)$ be a CMP, $\Omega := \mathcal{E}^{\mathbb{N}_0}$ with the product σ-algebra $\mathcal{F}$, coordinates $Z_t = (X_t,A_t)$, histories $\xi_t := (Z_0,\dots,Z_{t-1},X_t)$ and filtration $\mathcal{F}_t := \sigma(Z_0,\dots,Z_t)$. For every $\nu \in \mathcal{P}(\mathcal{X})$ and $\pi \in \Pi^{\mathrm{HR}}$ there is a unique $\mathbb{P}_\nu^\pi \in \mathcal{P}(\Omega)$ such that*

(a) $\mathbb{P}_\nu^\pi\{X_0 \in B\} = \nu(B)$*;*

(b) $\mathbb{P}_\nu^\pi\{A_t \in C \mid \sigma(\xi_t)\} = \pi_t(C|\xi_t)$ *a.s.;*

(c)

$$\mathbb{P}_\nu^\pi\{X_{t+1} \in B \mid \mathcal{F}_t\} = P(B|X_t,A_t)$$

a.s.

For countable spaces this means

$$\begin{aligned}\mathbb{P}_\nu^\pi\{\xi_t &= (x_0,a_0,\dots,x_t),A_t \\ &= a_t\} \\ &= \nu(x_0)\prod_{s=0}^{t}\pi_s(a_s|x_0,a_0,\dots,x_s)\prod_{s=0}^{t-1}P(x_{s+1}|x_s,a_s).\end{aligned}$$

$\mathbb{P}_x^\pi := \mathbb{P}_{\delta_x}^\pi$*; for $z = (x,a)$, $\mathbb{P}_z^\pi$ is the measure for the initial law δ_x and the policy that plays a at time* 0 *and π_t afterwards. $\mathbb{E}_\nu^\pi$, $\mathbb{E}_x^\pi$, $\mathbb{E}_z^\pi$ are the expectations.*

Proof. Interleave the coordinates: $Y_{2t} := X_t \in \mathcal{X}$, $Y_{2t+1} := A_t \in \mathcal{A}$. Apply Ionescu-Tulcea [0.5] with initial law ν and kernels $K_{2t+1}(y_{0:2t};\cdot) := \pi_t(\cdot|\xi_t)$, $K_{2t+2}(y_{0:2t+1};\cdot) := P(\cdot|x_t,a_t)$, which are Markov kernels from the past coordinates. The resulting measure on $\prod_t(\mathcal{X}\times\mathcal{A}) = \Omega$ has the finite-dimensional laws $\nu\otimes K_1\otimes\cdots\otimes K_n$; these determine it [0.5], and they are equivalent to (a)–(c): for $B' \in \sigma(\xi_t)$,

$$\mathbb{P}\{\xi_t \in B', A_t \in C\} = \int_{B'} \mathbb{P}\{\xi_t \in d\xi\}\pi_t(C|\xi)$$

is the defining property of $\pi_{2t+1} = \pi_{2t} \otimes K_{2t+1}$, and likewise for (c) (Çınlar, 2011, Remark IV.4.8(a)). For countable spaces the product formula is the iterated product of the kernels. ∎

Intuition. The policy and the environment take turns in drawing the next coordinate of an infinite sequence. Ionescu-Tulcea guarantees that these turn-by-turn rules define one probability measure on whole trajectories, with no topological assumption on the spaces. Everything an agent can ever observe or optimise is a function of a trajectory, so its distribution under this measure is all there is.

Proposition 3.4 (Induced chains and the Markov property)**.** *Let $\pi \in \Pi^{\mathrm{SR}}$ and $\nu \in \mathscr{P}(\mathcal{X})$; $\mathcal{X}$, $\mathcal{A}$ countable.*

(a) Under $\mathbb{P}_\nu^\pi$, (X_t) is a Markov chain on $\mathcal{X}$ with transition matrix P_π and initial law ν.

(b) Under $\mathbb{P}_\nu^\pi$, (Z_t) is a Markov chain on $\mathcal{Z}$ with transition matrix $P \otimes \pi$ and initial law $\nu \otimes \pi$; under $\mathbb{P}_z^\pi$, with initial law δ_z. In particular $\mathbb{P}_z^\pi$ is the law of the chain $P \otimes \pi$ started at z, and $\mathbb{P}_\nu^\pi = \int \nu(dx)\pi(da|x)\mathbb{P}_{(x,a)}^\pi$.

(c) (Markov property for the trajectory measure.) Let θ_s be the shift $(\theta_s\omega)_t := \omega_{s+t}$ on Ω. For bounded measurable $F : \Omega \to \mathbb{R}$,

$$\mathbb{E}_\nu^\pi[F \circ \theta_s \mid \mathscr{F}_s] = \mathbb{E}_{Z_s}^\pi F$$

and

$$\mathbb{E}_\nu^\pi[F \circ \theta_{s+1} \mid \mathscr{F}_s] = \int P(dx'|Z_s)\, \mathbb{E}_{x'}^\pi F$$

a.s.; the same holds at stopping times on $\{\tau < \infty\}$.

Proof.

(b) By the product formula of [3.3] with $\pi_s = \pi$,

$$\mathbb{P}_\nu^\pi\{Z_0 = z_0, \dots, Z_t = z_t\} = \nu(x_0)\pi(a_0|x_0) \prod_{s<t} P(x_{s+1}|z_s)\pi(a_{s+1}|x_{s+1}),$$

which is

$$(\nu \otimes \pi)(z_0) \prod_{s<t} (P \otimes \pi)(z_s, z_{s+1}).$$

By [1.2(a)] and the uniqueness in [0.5], $\mathbb{P}_\nu^\pi$ is the path measure of the chain $P \otimes \pi$ on $\mathcal{Z}$. The same computation with $\pi_0 = \delta_a$ gives $\mathbb{P}_z^\pi$, and the mixture formula follows by summing over z_0.

(a) Sum the formula of (b) over $a_0, \dots, a_t$:

$$\mathbb{P}_\nu^\pi\{X_0 = x_0, \dots, X_t = x_t\} = \nu(x_0) \prod_{s<t} P_\pi(x_{s+1}|x_s).$$

(c) By (b), this is the Markov property [1.2(b),(c)] of the chain $P \otimes \pi$ on the countable space $\mathscr{Z}$. For the second identity, condition the first identity at time $s+1$ on $\mathscr{F}_s$ and use the tower property [0.6(a)], (b) and [3.3(c)]:

$$\begin{aligned}\mathbb{E}_\nu^\pi[\mathbb{E}_{Z_{s+1}}^\pi F \mid \mathscr{F}_s] &= \int P(dx'|Z_s) \int \pi(da'|x') \mathbb{E}_{(x',a')}^\pi F \\ &= \int P(dx'|Z_s) \mathbb{E}_{x'}^\pi F.\end{aligned}$$

■

3.3 Returns and Values

Definition 3.5 (Return and value)**.** Fix an MDP $(\mathscr{X}, \mathscr{A}, P, r, \gamma)$.

(a) The *return functional* is $G : \Omega \to \mathbb{R}$, $G(\omega) := \sum_{t\geq 0} \gamma^t r(z_t)$ for $\omega = (z_0, z_1, \dots)$; the *H-step return* is $G_H(\omega) := \sum_{t<H} r(z_t)$. $R_t := r(Z_t)$ is the reward at time t.

(b) The *value* of $\pi \in \Pi^{\mathrm{HR}}$ is $V^\pi(\nu) := \mathbb{E}_\nu^\pi G$, with $V^\pi(x) := V^\pi(\delta_x)$, and its *action value* is $Q^\pi(x,a) := \mathbb{E}_{(x,a)}^\pi G$. The *optimal values* are $V^\star(x) := \sup_{\pi\in\Pi^{\mathrm{HR}}} V^\pi(x)$ and $Q^\star(x,a) := \sup_{\pi\in\Pi^{\mathrm{HR}}} Q^\pi(x,a)$.

(c) The *discounted occupancy measure* of π from ν is $d_\nu^\pi \in \mathscr{P}(\mathscr{Z})$,

$$d_\nu^\pi(C) := (1-\gamma) \sum_{t\geq 0} \gamma^t \mathbb{P}_\nu^\pi\{Z_t \in C\};$$

its state marginal is $d_{\nu,\mathscr{X}}^\pi(B) := d_\nu^\pi(B \times \mathscr{A})$.

Proposition 3.6 (Properties of the return)**.**

(a) G is measurable with $0 \leq G \leq \frac{1}{1-\gamma}$, and

$$\Big|G - \sum_{t<H} \gamma^t r(Z_t)\Big| \leq \frac{\gamma^H}{1-\gamma}.$$

(b) (Shift recursion.) $G = r(Z_0) + \gamma\, G \circ \theta_1$ pointwise on Ω.

(c) $V^\pi(\nu) = \frac{1}{1-\gamma}\langle d_\nu^\pi, r\rangle$, and $0 \leq V^\pi, Q^\pi, V^\star, Q^\star \leq \frac{1}{1-\gamma}$.

Proof.

(a) Each partial sum is measurable (coordinates are measurable and r is), and the partial sums increase to G; the tail is at most $\sum_{t\ge H}\gamma^t$.

(b)
$$G(\omega) = r(z_0) + \gamma\sum_{t\ge 0}\gamma^t r(z_{t+1}) = r(z_0) + \gamma G(\theta_1\omega).$$

(c) By monotone convergence [0.2(b)],
$$\mathbb{E}_\nu^\pi G = \sum_t \gamma^t \mathbb{E}_\nu^\pi r(Z_t) = \frac{1}{1-\gamma}\int r\,d d_\nu^\pi.$$

The bounds follow from (a). ∎

3.4 Markov Policies Suffice for Marginals

Theorem 3.7 (Marginal equivalence; Strauch, 1966, Derman, 1970)**.** *Let $\mathcal{X}, \mathcal{A}$ be countable, $\pi \in \Pi^{\mathrm{HR}}$ and $\nu \in \mathcal{P}(\mathcal{X})$. Define $\pi' \in \Pi^{\mathrm{MR}}$ by*
$$\pi'_t(a|x) := \mathbb{P}_\nu^\pi\{A_t = a \mid X_t = x\} \quad \text{if } \mathbb{P}_\nu^\pi\{X_t = x\} > 0,$$
and arbitrarily otherwise. Then $\mathbb{P}_\nu^{\pi'}\{Z_t = z\} = \mathbb{P}_\nu^\pi\{Z_t = z\}$ for all t and z. Consequently $d_\nu^{\pi'} = d_\nu^\pi$, and $V^{\pi'}(\nu) = V^\pi(\nu)$ for every reward r and every γ, and the same holds for H-step values.

Proof. Write $p_t(z) := \mathbb{P}_\nu^\pi\{Z_t = z\}$, p'_t likewise for π', and m_t, m'_t for the laws of X_t. Under any policy $\tilde\pi$, [3.3(b)] and the tower property [0.6(a)] give
$$\mathbb{P}_\nu^{\tilde\pi}\{Z_t = (x,a)\} = \mathbb{E}_\nu^{\tilde\pi}[\mathbf{1}\{X_t = x\}\tilde\pi_t(a|\xi_t)],$$
and [3.3(c)] gives
$$\mathbb{P}_\nu^{\tilde\pi}\{X_{t+1} = y\} = \sum_z \mathbb{P}_\nu^{\tilde\pi}\{Z_t = z\}P(y|z).$$
Induction on t: $m_0 = m'_0 = \nu$. If $m_t = m'_t$, then
$$p'_t(x,a) = m'_t(x)\pi'_t(a|x) = m_t(x)\mathbb{P}_\nu^\pi\{A_t = a \mid X_t = x\} = p_t(x,a)$$
(both sides vanish when $m_t(x) = 0$), and then $m_{t+1} = m'_{t+1}$ by the second identity. The consequences follow from [3.6(c)] and the analogous formula $\mathbb{E}_\nu^{\tilde\pi} G_H = \sum_{t<H}\sum_z p_t(z)r(z)$. ∎

Intuition. An expected sum of rewards only sees the law of each Z_t separately, never their joint law across times. A Markov policy can reproduce these one-time marginals by randomising with the right conditional frequencies. Return *laws* are different: they depend on the joint law of the whole trajectory, and for them memory can matter. This is the first sign that the expected and the distributional theory part ways.

3.5 The Running Examples

Example 3.8 (The goal CMP)**.** Let $\mathcal{X} = \{1, 2, g\}$ and $\mathcal{A} = \{\mathsf{f}, \mathsf{b}\}$ (forward, back), with deterministic transitions

$$1 \xrightarrow{\mathsf{f}} 2, \quad 2 \xrightarrow{\mathsf{f}} g, \quad 1 \xrightarrow{\mathsf{b}} 1, \quad 2 \xrightarrow{\mathsf{b}} 1, \quad g \xrightarrow{\mathsf{f},\mathsf{b}} g,$$

and the goal reward $r(x, a) := \mathbf{1}\{x = g\}$. Under the uniform policy

$$\pi_{\mathrm{u}}(\cdot|x) = \frac{1}{2}(\delta_{\mathsf{f}} + \delta_{\mathsf{b}}),$$

the state kernel $P_{\pi_{\mathrm{u}}}$ is exactly the chain of [1.24]. Since g is absorbing, $G = \sum_{t \geq \tau_g} \gamma^t = \gamma^{\tau_g}/(1-\gamma)$, so $V^{\pi_{\mathrm{u}}} = q_g^{\gamma}/(1-\gamma)$ with the discounted reach value of [2.3(c)]; for $\gamma = 0.9$, $V^{\pi_{\mathrm{u}}} \approx (5.83, 7.12, 10)$ by [2.12]. The deterministic policy $\pi_{\mathsf{f}} \equiv \mathsf{f}$ reaches g in 2 steps from 1 and 1 step from 2, so $V^{\pi_{\mathsf{f}}} = (\gamma^2, \gamma, 1)/(1-\gamma)$, which is $(8.1, 9, 10)$ for $\gamma = 0.9$. Under π_{u} the return G is a nondegenerate random variable; under π_{f} it is deterministic.

Example 3.9 (The gridworld CMP)**.** Let $\mathcal{X} = \{1, \dots, 5\}^2$, $\mathcal{A} = \{\mathsf{N}, \mathsf{S}, \mathsf{E}, \mathsf{W}\}$, with deterministic moves that leave the state unchanged at the boundary, as in [1.25], except that the goal $g = (5, 5)$ is absorbing; $r(x, a) := \mathbf{1}\{x = g\}$. Under the uniform policy the state kernel is that of [1.25] with the row of g replaced by δ_g, so the chain is absorbed at g a.s. [1.9(c)], and $\mathbb{E}_x^{\pi_{\mathrm{u}}} \tau_g < \infty$ is the solution of the Poisson equation [2.4(c)].

Assumed without proof in this chapter

Nothing beyond Chapters 0–2.

4

Bellman Operators and Dynamic Programming

Dynamic programming replaces optimisation over trajectories by a one-step equation. This chapter proves the principle of optimality in its two classical forms, finite horizon and discounted infinite horizon (Bellman, 1957; Blackwell, 1965), the convergence of value iteration and policy iteration (Howard, 1960), and three tools on which the rest of the book relies: the greedy-policy loss bound (Singh and Yee, 1994), the performance-difference lemma (Kakade and Langford, 2002), and the simulation lemma (Kearns and Singh, 2002). Optimality over *all* history-dependent policies is proved by a supermartingale argument on the trajectory measure, without first reducing to Markov policies. Throughout, $\mathcal{X}$ is countable, $\mathcal{A}$ is finite, and $(\mathcal{X}, \mathcal{A}, P, r, \gamma)$ is an MDP [3.1]; functions on $\mathcal{X}$ carry the sup norm.

4.1 Bellman Operators

Definition 4.1 (Bellman operators)**.** For a stationary π and $Q \in B_b(\mathcal{X})$, $V \in B_b(\mathcal{X})$ let

$$
\begin{aligned}
T^{\pi} Q(x,a) &:= r(x,a) + \gamma \int P(dx' \mid x,a) \int \pi(da' \mid x')\, Q(x',a'),\\
T^{\star} Q(x,a) &:= r(x,a) + \gamma \int P(dx' \mid x,a)\, \max_{a' \in \mathcal{A}} Q(x',a'),\\
T^{\pi} V(x) &:= r_{\pi}(x) + \gamma P_{\pi} V(x),\\
T^{\star} V(x) &:= \max_{a \in \mathcal{A}} \Big[r(x,a) + \gamma \int P(dx' \mid x,a) V(x') \Big].
\end{aligned}
$$

A stationary deterministic π is *greedy* for Q, written $\pi = \pi_Q$, if $\pi(x) \in \arg\max_a Q(x,a)$ for all x (ties broken by a fixed order of $\mathcal{A}$). The *advantage* of π is $\mathbb{A}^{\pi}(x,a) := Q^{\pi}(x,a) - V^{\pi}(x)$.

Proposition 4.2 (Monotonicity and contraction)**.** *Each of the four operators $\mathsf{T} \in \{T^{\pi}, T^{\star}\}$ (on $B_b(\mathcal{X})$ or $B_b(\mathcal{X})$) satisfies:*

(a) (Monotonicity.) $Q \le Q'$ implies $\mathsf{T}Q \le \mathsf{T}Q'$.

(b) (Constant shift.) $\mathsf{T}(Q + c\mathbf{1}) = \mathsf{T}Q + \gamma c \mathbf{1}$ for $c \in \mathbb{R}$.

(c) (Contraction.)

$$\|\mathsf{T}Q - \mathsf{T}Q'\|_\infty \le \gamma \|Q - Q'\|_\infty.$$

Hence T *has a unique fixed point in* B_b*, and its iterates converge to it at rate* γ^k *[0.8].*

Proof. (a) and (b): integration against Markov kernels and $\max_a$ are monotone and commute with adding constants [0.4(c)]. (c): $Q \le Q' + \|Q - Q'\|_\infty \mathbf{1}$, so by (a) and (b) $\mathsf{T}Q \le \mathsf{T}Q' + \gamma\|Q - Q'\|_\infty \mathbf{1}$; exchange Q, Q'. B_b with the sup norm is a Banach space. ∎

Theorem 4.3 (Bellman consistency)**.** *Let* π *be stationary. Then* Q^π *is the unique fixed point of* T^π *on* $B_b(\mathcal{Z})$ *and* V^π *the unique fixed point of* T^π *on* $B_b(\mathcal{X})$*; moreover*

$$V^\pi(x) = \int \pi(da|x) Q^\pi(x,a),$$

$$Q^\pi(x,a) = r(x,a) + \gamma \int P(dx'|x,a) V^\pi(x').$$

For finite $\mathcal{X}$*,* $V^\pi = (I - \gamma P_\pi)^{-1} r_\pi$*, i.e.* V^π *is the resolvent [2.7] of the state chain applied to the policy reward.*

Proof. $V^\pi(x) = \int \pi(da|x) Q^\pi(x,a)$ is the mixture formula [3.4(b)]. By the shift recursion [3.6(b)] and the Markov property [3.4(c)] with $s = 0$,

$$\begin{aligned} Q^\pi(z) &= \mathbb{E}_z^\pi\big[r(Z_0) + \gamma\, G \circ \theta_1\big] \\ &= r(z) + \gamma \int P(dx'|z)\, \mathbb{E}_{x'}^\pi G \\ &= r(z) + \gamma \int P(dx'|z) V^\pi(x'). \end{aligned}$$

Substituting the first identity into the second gives $Q^\pi = T^\pi Q^\pi$, and $V^\pi = T^\pi V^\pi$ by integrating against π. Uniqueness: [4.2(c)]. The matrix form is [2.7(a)] for P_π. ∎

4.2 Finite Horizon: Backward Induction

Theorem 4.4 (Principle of optimality, finite horizon; Bellman, 1957)**.** *Fix* $H \in \mathbb{N}$ *and define* $W_H \coloneqq 0$ *and, for* $h = H-1, \dots, 0$*,*

$$W_h(x) \coloneqq \max_{a \in \mathcal{A}} \Big[r(x,a) + \int P(dx'|x,a) W_{h+1}(x') \Big],$$

with maximisers $f_h(x)$. Let $\pi_H^\star$ be the Markov deterministic policy $\pi_h(\cdot|x) = \delta_{f_h(x)}$ for $h < H$. Then for every $\pi \in \Pi^{\mathrm{HR}}$ and x, $\mathbb{E}_x^\pi G_H \le W_0(x) = \mathbb{E}_x^{\pi_H^\star} G_H$. So Markov deterministic policies are optimal for the H-step return among all history-dependent randomised policies.

Proof. Let $\pi \in \Pi^{\mathrm{HR}}$ and $M_h := \sum_{s<h} r(Z_s) + W_h(X_h)$ for $h \le H$. By [3.3(b),(c)] and the tower property [0.6(a)],

$$\mathbb{E}_x^\pi\big[r(Z_h) + W_{h+1}(X_{h+1}) \,\big|\, \sigma(\xi_h)\big] = \int \pi_h(da|\xi_h)\Big[r(X_h,a) + \int P(dx'|X_h,a) W_{h+1}(x')\Big] \le W_h(X_h),$$

so $\mathbb{E}_x^\pi[M_{h+1} \mid \sigma(\xi_h)] \le M_h$. Taking expectations repeatedly, $\mathbb{E}_x^\pi G_H = \mathbb{E}_x^\pi M_H \le \mathbb{E}_x^\pi M_0 = W_0(x)$. For $\pi_H^\star$ the inequality is an equality at every step, because π_h puts its mass on a maximiser. ∎

4.3 Discounted Infinite Horizon

Theorem 4.5 (Bellman optimality, discounted; Blackwell, 1965)**.** *Let $\bar{Q}$ be the unique fixed point of $T^\star$ on $B_b(\mathcal{X})$ [4.2(c)] and $\bar{V}(x) := \max_a \bar{Q}(x,a)$.*

(a) For every $\pi \in \Pi^{\mathrm{HR}}$: $V^\pi \le \bar{V}$ and $Q^\pi \le \bar{Q}$.

(b) Every greedy policy $\pi_{\bar{Q}}$ satisfies $Q^{\pi_{\bar{Q}}} = \bar{Q}$ and $V^{\pi_{\bar{Q}}} = \bar{V}$.

(c) Hence $Q^\star = \bar{Q}$, $V^\star = \bar{V}$ is the unique fixed point of $T^\star$ on $B_b(\mathcal{X})$, and there is an optimal policy in Π^{SD}, optimal simultaneously for all initial states: $\pi^\star := \pi_{\bar{Q}}$..

Proof.

(a) $\bar{V}$ is bounded and $\bar{Q}(x,a) = r(x,a) + \gamma \int P(dx'|x,a)\bar{V}(x')$. Let $M_t := \sum_{s<t} \gamma^s r(Z_s) + \gamma^t \bar{V}(X_t)$. As in the proof of [4.4],

$$\mathbb{E}_x^\pi[M_{t+1} \mid \sigma(\xi_t)] = \sum_{s<t} \gamma^s r(Z_s) + \gamma^t \int \pi_t(da|\xi_t)\bar{Q}(X_t,a) \le M_t.$$

Hence

$$\mathbb{E}_x^\pi \sum_{s<t} \gamma^s r(Z_s) \le \bar{V}(x) - \gamma^t \mathbb{E}_x^\pi \bar{V}(X_t);$$

let $t \to \infty$ [3.6(a)]: $V^\pi(x) \le \bar{V}(x)$. For Q^π, run the same argument under $\mathbb{P}_z^\pi$ with

$$M_t' := \sum_{1 \le s < t} \gamma^{s-1} r(Z_s) + \gamma^{t-1} \bar{V}(X_t),$$

$t \geq 1$, which is again a supermartingale:

$$\mathbb{E}_z^\pi[G \circ \theta_1] \leq \mathbb{E}_z^\pi \bar{V}(X_1) = \int P(dx'|z)\bar{V}(x').$$

Hence

$$Q^\pi(z) = r(z) + \gamma \mathbb{E}_z^\pi[G \circ \theta_1] \leq \bar{Q}(z)$$

by [3.6(b)].

(b) For $\bar{\pi} := \pi_{\bar{Q}}$, $T^{\bar{\pi}}\bar{Q} = T^\star \bar{Q} = \bar{Q}$, so $\bar{Q} = Q^{\bar{\pi}}$ by uniqueness [4.3], and $V^{\bar{\pi}}(x) = \bar{Q}(x, \bar{\pi}(x)) = \bar{V}(x)$.

(c) (a) and (b) give $V^\star = \bar{V}$, $Q^\star = \bar{Q}$. Finally $T^\star \bar{V}(x) = \max_a \bar{Q}(x,a) = \bar{V}(x)$.

■

Theorem 4.6 (Value iteration and the greedy-policy loss; Singh and Yee, 1994)**.**

(a) For every $Q_0 \in B_b(\mathcal{Z})$, $Q_k := (T^\star)^k Q_0$ satisfies

$$\|Q_k - Q^\star\|_\infty \leq \gamma^k \|Q_0 - Q^\star\|_\infty.$$

(b) For every $Q \in B_b(\mathcal{Z})$,

$$\|V^\star - V^{\pi_Q}\|_\infty \leq \frac{2\|Q - Q^\star\|_\infty}{1-\gamma}.$$

(c) Starting from $Q_0 = 0$, the greedy policy π_{Q_k} is ε-optimal ($V^\star - V^{\pi_{Q_k}} \leq \varepsilon$) as soon as

$$k \geq \frac{1}{1-\gamma} \ln \frac{2}{\varepsilon(1-\gamma)^2}.$$

Proof.

4.2(c) and [4.5(c)].

(a) Let $\epsilon := \|Q - Q^\star\|_\infty$, $\pi := \pi_Q$, $x \in \mathcal{X}$, $a^\star := \pi^\star(x)$, $a := \pi(x)$. Since $Q(x,a) \geq Q(x,a^\star)$,

$$Q^\star(x,a^\star) - Q^\star(x,a) \leq Q^\star(x,a^\star) - Q(x,a^\star) + Q(x,a) - Q^\star(x,a) \leq 2\epsilon.$$

Then

$$\begin{aligned} V^\star(x) - V^\pi(x) &= [Q^\star(x,a^\star) - Q^\star(x,a)] + [Q^\star(x,a) - Q^\pi(x,a)] \\ &\leq 2\epsilon + \gamma \int P(dx'|x,a)(V^\star - V^\pi)(x') \end{aligned}$$

by [4.3] and [4.5]. Take the supremum over x.

(b) $\|Q_0 - Q^\star\|_\infty \le \frac{1}{1-\gamma}$ [3.6(c)]; combine (a), (b) with $\gamma^k \le e^{-(1-\gamma)k}$.

■

Theorem 4.7 (Policy iteration; Howard, 1960)**.** *Let $\pi_0 \in \Pi^{\mathrm{SD}}$ and $\pi_{k+1} := \pi_{Q^{\pi_k}}$. Then*

(a) $Q^{\pi_{k+1}} \ge T^\star Q^{\pi_k} \ge Q^{\pi_k}$ (monotone improvement);

(b)

$$\|Q^\star - Q^{\pi_k}\|_\infty \le \gamma^k \|Q^\star - Q^{\pi_0}\|_\infty;$$

(c) if $\mathcal{X}$ is finite, π_k is optimal for some $k \le |\mathcal{A}|^{|\mathcal{X}|}$, and $Q^{\pi_{k+1}} = Q^{\pi_k}$ only if π_k is optimal.

Proof.

(a)

$$T^{\pi_{k+1}} Q^{\pi_k} = T^\star Q^{\pi_k} \ge T^{\pi_k} Q^{\pi_k} = Q^{\pi_k}.$$

Applying the monotone [4.2(a)] map $T^{\pi_{k+1}}$ repeatedly, $(T^{\pi_{k+1}})^n Q^{\pi_k} \ge T^\star Q^{\pi_k}$ for all $n \ge 1$, and the left side tends to $Q^{\pi_{k+1}}$ [4.2(c)], [4.3].

(b)

$$0 \le Q^\star - Q^{\pi_{k+1}} \le T^\star Q^\star - T^\star Q^{\pi_k} \le \gamma \|Q^\star - Q^{\pi_k}\|_\infty$$

by (a), [4.5(a)] and [4.2(c)].

(c) If $Q^{\pi_{k+1}} = Q^{\pi_k}$, then $T^\star Q^{\pi_k} = Q^{\pi_k}$ by (a), so $Q^{\pi_k} = Q^\star$. Otherwise the improvement is strict somewhere, no policy repeats, and there are $|\mathcal{A}|^{|\mathcal{X}|}$ deterministic stationary policies.

■

4.4 The Performance-Difference Lemma and the Linear Program

Lemma 4.8 (Performance difference; Kakade and Langford, 2002)**.** *For a stationary π, any $\pi' \in \Pi^{\mathrm{HR}}$ and $\nu \in \mathcal{P}(\mathcal{X})$,*

$$\begin{aligned} V^{\pi'}(\nu) - V^{\pi}(\nu) &= \frac{1}{1-\gamma} \int \mathbb{A}^{\pi}\, d d_\nu^{\pi'} \\ &= \frac{1}{1-\gamma} \int d_\nu^{\pi'}(dx, da) \left[Q^{\pi}(x,a) - V^{\pi}(x) \right]. \end{aligned}$$

Proof. Pointwise on Ω the bounded telescoping sum gives

$$G - V^{\pi}(X_0) = \sum_{t \ge 0} \gamma^t \left[r(Z_t) + \gamma V^{\pi}(X_{t+1}) - V^{\pi}(X_t) \right].$$

Under $\mathbb{P}_\nu^{\pi'}$, by [3.3(c)] and [4.3],

$$\mathbb{E}_\nu^{\pi'}[r(Z_t) + \gamma V^\pi(X_{t+1}) \mid \mathscr{F}_t] = Q^\pi(Z_t).$$

Take expectations, exchange sum and expectation by dominated convergence [0.2(c)], and use the definition [3.5(c)] of $d_\nu^{\pi'}$. ∎

Intuition. The difference between two policies is the average, along the trajectories of the *new* policy, of how much better its actions are than the old policy's average action, judged by the old policy's own values. Policy iteration [4.7] uses this with the greedy improvement; policy-gradient methods (Part II) use it with small improvements.

Theorem 4.9 (Occupancy measures and the linear program; Derman, 1970)**.** *Let $\mathcal{X}$ be finite and $\nu \in \mathcal{P}(\mathcal{X})$. Let K_ν be the set of $d : \mathcal{Z} \to [0,\infty)$ satisfying the* Bellman flow equations

$$\sum_a d(x',a) = (1-\gamma)\nu(x') + \gamma \sum_{(x,a)\in\mathcal{Z}} P(x'|x,a)\, d(x,a) \qquad (x' \in \mathcal{X}).$$

(a) $\mathsf{K}_\nu = \{d_\nu^\pi : \pi \in \Pi^{\mathrm{SR}}\}$; for $d \in \mathsf{K}_\nu$ with state marginal $m(x) := \sum_a d(x,a)$, the policy $\pi(a|x) := d(x,a)/m(x)$ (arbitrary if $m(x)=0$) satisfies $d_\nu^\pi = d$. Also $d_\nu^{\pi'} \in \mathsf{K}_\nu$ for every $\pi' \in \Pi^{\mathrm{HR}}$.

(b) (Duality.) For every $V \in B_b(\mathcal{X})$ with $V \ge T^*V$ (primal feasibility) and every $d \in \mathsf{K}_\nu$ (dual feasibility), $\frac{1}{1-\gamma}\langle d, r\rangle \le \langle \nu, V\rangle$. Equality holds for $V = V^*$ and $d = d_\nu^{\pi^*}$; so

$$V^*(\nu) = \min\{\langle \nu, V\rangle : V \ge T^*V\} = \max\left\{\tfrac{1}{1-\gamma}\langle d, r\rangle : d \in \mathsf{K}_\nu\right\}.$$

Proof.

(a) For any π',

$$\mathbb{P}_\nu^{\pi'}\{X_{t+1} = x'\} = \sum_z \mathbb{P}_\nu^{\pi'}\{Z_t = z\} P(x'|z)$$

[3.3(c)]; multiply by $(1-\gamma)\gamma^{t+1}$ and sum over t, adding the term $t=0$, $(1-\gamma)\nu(x')$: this is the flow equation for $d_\nu^{\pi'}$. Conversely let $d \in \mathsf{K}_\nu$ and π as stated, so $d(x,a) = m(x)\pi(a|x)$ (when $m(x)=0$, $d(x,\cdot)=0$). Then $\sum_z P(x'|z)d(z) = \sum_x m(x)P_\pi(x'|x)$, and the flow equation reads $m(I-\gamma P_\pi) = (1-\gamma)\nu$. $I-\gamma P_\pi$ is invertible [2.7(a)], and $d_{\nu,\mathcal{X}}^\pi$ solves the same equation by the first part, so $m = d_{\nu,\mathcal{X}}^\pi$ and $d(x,a) = d_{\nu,\mathcal{X}}^\pi(x)\pi(a|x) = d_\nu^\pi(x,a)$ [3.4(b)].

(b) From $V(x) \ge r(x,a) + \gamma \sum_{x'} P(x'|x,a)V(x')$ and $d \ge 0$,

$$\begin{aligned}\langle d, r\rangle &\le \sum_{z} d(z)\big[V(x) - \gamma \sum_{x'} P(x'|z)V(x')\big] \\ &= \sum_{x'} V(x')\big[m(x') - \gamma \sum_{z} d(z)P(x'|z)\big] \\ &= (1-\gamma)\langle \nu, V\rangle,\end{aligned}$$

by the flow equation. $V^\star$ is primal feasible [4.5(c)], $d_\nu^{\pi^\star}$ is dual feasible by (a), and

$$\frac{1}{1-\gamma}\langle d_\nu^{\pi^\star}, r\rangle = V^{\pi^\star}(\nu) = V^\star(\nu)$$

[3.6(c)].

■

4.5 The Simulation Lemma

Theorem 4.10 (Simulation lemma; Kearns and Singh, 2002)**.** *Let P' be a second transition kernel on the same $(\mathcal{X}, \mathcal{A})$, with the same r and γ; primes denote the corresponding values. Let*

$$\epsilon_P := \sup_{z \in \mathcal{Z}} \| P'(\cdot|z) - P(\cdot|z)\|_{\mathrm{TV}}.$$

Then for every stationary π,

$$\|Q'^{\pi} - Q^{\pi}\|_\infty \le \frac{\gamma\,\epsilon_P}{(1-\gamma)^2}, \qquad \|Q^{\star\prime} - Q^\star\|_\infty \le \frac{\gamma\,\epsilon_P}{(1-\gamma)^2}.$$

More precisely,

$$\|Q'^{\pi} - Q^{\pi}\|_\infty \le \frac{\gamma}{1-\gamma} \sup_z \Big| \int (P' - P)(dx'|z) V^\pi(x') \Big|.$$

Proof. By [3.4(b)] and [4.3], Q^π is the resolvent of the chain $P \otimes \pi$ on $\mathcal{Z}$ applied to r: $Q^\pi(z) = \mathbb{E}_z^\pi \sum_t \gamma^t r(Z_t)$ [2.7(b)]; likewise Q'^π with $P' \otimes \pi$. Since both chains use the same π,

$$\|(P' \otimes \pi)(z,\cdot) - (P \otimes \pi)(z,\cdot)\|_{\mathrm{TV}} = \|P'(\cdot|z) - P(\cdot|z)\|_{\mathrm{TV}} \le \epsilon_P,$$

and

$$((P' \otimes \pi) - (P \otimes \pi))Q^\pi(z) = \int (P' - P)(dx'|z) V^\pi(x')$$

by [4.3]. Apply [2.7(e)] with $\mathrm{sp}(r) \le 1$. For the optimal values, $T^{\star\prime}$ and $T^\star$ differ by $\gamma \int (P' - P)(dx'|z) \max_{a'} Q(x', a')$, so

$$\|T^{\star\prime}Q^\star - T^\star Q^\star\|_\infty \le \gamma\epsilon_p \,\mathrm{sp}(V^\star) \le \frac{\gamma\epsilon_p}{1-\gamma}$$

[0.11(c)]; then [0.8(c)] for the contraction $T^{\star\prime}$ gives

$$\|Q^\star - Q^{\star\prime}\|_\infty \le \frac{\gamma\epsilon_p}{(1-\gamma)^2}.$$

■

Intuition. A model error of ϵ_p per step is felt for an effective horizon of $\frac{1}{1-\gamma}$ steps, and each error moves a value whose range is itself $\frac{1}{1-\gamma}$. The two factors of the effective horizon in the simulation lemma are the starting point of the sample-complexity theory of Part III, which removes one of them by exploiting the variance of V^π rather than its range.

4.6 The Running Examples

Example 4.11 (Dynamic programming on the goal CMP)**.** For the goal CMP of [3.8],

$$V^\star = V^{\pi_{\mathsf{f}}} = (\gamma^2, \gamma, 1)/(1-\gamma)$$

and

$$Q^\star(1,\mathsf{f}) = \tfrac{\gamma^2}{1-\gamma}, \quad Q^\star(1,\mathsf{b}) = \tfrac{\gamma^3}{1-\gamma}, \quad Q^\star(2,\mathsf{f}) = \tfrac{\gamma}{1-\gamma}, \quad Q^\star(2,\mathsf{b}) = \tfrac{\gamma^3}{1-\gamma},$$

which is the fixed point of $T^\star$ [4.5]. Policy iteration [4.7] from $\pi_0 = \pi_{\mathsf{b}}$ (always back, value 0 at 1, 2) gives $\pi_1(2) = \mathsf{f}$ and $\pi_1(1) = \mathsf{b}$ (ties at 1 broken towards b under the order $\mathsf{b} < \mathsf{f}$), then $\pi_2 = \pi_{\mathsf{f}}$, which is optimal: two improvement steps, the length of the path to the goal. Value iteration from $Q_0 = 0$ discovers the goal one step further away per iteration: $Q_1 = r$ is positive only at g, and Q_2 is positive exactly at the pairs at g and at $(2,\mathsf{f})$, the only pair that reaches g in one step. The advantage of the uniform policy at $\gamma = 0.9$ is

$$\mathbb{A}^{\pi_{\mathrm{u}}}(1,\mathsf{f}) = \gamma V^{\pi_{\mathrm{u}}}(2) - V^{\pi_{\mathrm{u}}}(1) \approx 0.58 > 0,$$

so the performance-difference lemma [4.8] predicts the gain of moving probability towards f.

Assumed without proof in this chapter

Nothing beyond Chapters 0–3.

5

Stochastic Shortest Paths and Average Reward

Discounting is one of three ways to make an infinite sum of rewards finite. The other two are stopping, in which the process is absorbed in a terminal state (the goal-reaching problems of Parts V and VI), and averaging, in which the long-run reward rate is optimised. This chapter develops dynamic programming for both: the stochastic shortest path problem, whose optimal value is the optimal expected cost of reaching a goal (Puterman and Chan, 2025, Ch. 7), and the average-reward problem, reached from the discounted one by Blackwell's vanishing-discount argument (Blackwell, 1962). The potential theory of Chapter 2 supplies the operators: the killed potential for the first, the Laurent expansion of the resolvent for the second. $\mathcal{X}$ and $\mathcal{A}$ are finite throughout.

5.1 Proper Policies and the Weighted Contraction

Definition 5.1 (Stochastic shortest path problems)**.** Let $g \in \mathcal{X}$ be a *goal*, absorbing under every action, $C := \mathcal{X} \setminus \{g\}$, and $c : \mathcal{X} \to [0, \infty)$ a *cost* with $c(g, \cdot) = 0$ and $c \geq c_{\min} > 0$ on $C \times \mathcal{A}$. A stationary policy π is *proper* if $\mathbb{P}_x^\pi\{\tau_g < \infty\} = 1$ for all x, and *improper* otherwise. The *cost-to-go* of $\pi \in \Pi^{\mathrm{HR}}$ is $J^\pi(x) := \mathbb{E}_x^\pi \sum_{t<\tau_g} c(Z_t) \in [0, \infty]$ and the optimal cost is $J^\star(x) := \inf_\pi J^\pi(x)$. The problem is an *SSP* if some stationary policy is proper (Puterman and Chan, 2025, Def. 6.4), and a *transient model* if every stationary policy is proper (Puterman and Chan, 2025, Def. 6.2). For $V \in \mathbb{R}^{\mathcal{X}}$ with $V(g) = 0$ define the *SSP Bellman operators*

$$T_c^\pi V(x) := c_\pi(x) + P_\pi V(x),$$
$$T_c^\star V(x) := \min_a \Big[c(x,a) + \sum_{x'} P(x' \mid x, a) V(x') \Big],$$
$$(x \in C)$$

and $T_c^\pi V(g) = T_c^\star V(g) = 0$.

Proposition 5.2 (Proper policies)**.** *For a stationary π, with $P_{\pi,C}$ the killed kernel [2.1]:*

(a) π is proper iff g is accessible under P_π from every state, iff $\varrho(P_{\pi,C}) < 1$;

(b) if π is proper, $J^\pi = (I - P_{\pi,C})^{-1} c_\pi$ on C, the unique fixed point of T_c^π, and $J^\pi \le \|c\|_\infty \kappa^\pi(\cdot, g)$;
(c) if π is improper, $J^\pi(x) = \infty$ for some x.

Proof.

(a) If g is accessible from every state, [2.4(a),(b)] give a geometric tail and $\varrho(P_{\pi,C}) < 1$, so π is proper. If g is not accessible from some x, then $\mathbb{P}_x^\pi\{\tau_g < \infty\} = 0$. If $\varrho(P_{\pi,C}) < 1$, then $P_{\pi,C}^t(x, C) = \mathbb{P}_x^\pi\{\tau_g > t\} \to 0$ [0.9(c)], [2.2(a)].
(b) By [2.2(b)], $J^\pi(x) = \sum_{y \in C} U_C(x, y) c_\pi(y)$, and $U_C = (I - P_{\pi,C})^{-1}$ [2.4(b)]; uniqueness is invertibility, and $c_\pi \le \|c\|_\infty$ gives the bound by [2.4(c)].
(c) If $\mathbb{P}_x^\pi\{\tau_g = \infty\} > 0$, then on that event the cost is at least $c_{\min} \cdot \infty$.

■

Theorem 5.3 (Weighted-norm contraction in transient models)**.** *In a transient model let* b *be the solution on C of*

$$\mathsf{b} = \mathbf{1} + \max_{\pi \in \Pi^{\mathrm{SD}}} P_{\pi,C} \mathsf{b},$$

i.e. the maximal expected time to the goal (Puterman and Chan, 2025, (6.54) and Prop. 6.7), $\bar{\mathsf{b}} := \max_x \mathsf{b}(x)$*, and* $\|V\|_{\mathsf{b}} := \max_{x \in C} |V(x)|/\mathsf{b}(x)$*. Then for every stationary π,*

$$\|P_{\pi,C} V\|_{\mathsf{b}} \le \Big(1 - \frac{1}{\bar{\mathsf{b}}}\Big) \|V\|_{\mathsf{b}},$$

so T_c^π and $T_c^\star$ are contractions with modulus $1 - 1/\bar{\mathsf{b}}$ in $\|\cdot\|_{\mathsf{b}}$, and value iteration for $J^\star$ converges geometrically from any start.

Proof. For every deterministic π, $P_{\pi,C}\mathsf{b} \le \mathsf{b} - \mathbf{1}$ by the defining equation, and for randomised π the same follows by averaging over actions. Hence

$$|P_{\pi,C} V(x)| \le \|V\|_{\mathsf{b}} P_{\pi,C}\mathsf{b}(x) \le \|V\|_{\mathsf{b}} (\mathsf{b}(x) - 1) \le \|V\|_{\mathsf{b}} \mathsf{b}(x)(1 - 1/\bar{\mathsf{b}}).$$

For $T_c^\star$: $|T_c^\star V - T_c^\star V'| \le \max_a |P(\cdot|\cdot, a)(V - V')|$ pointwise, and the same estimate applies to each action; use [0.8]. ■

Intuition. Without discounting, a Bellman operator contracts because the process stops. The weight b measures how long the process may take to stop, and a step of the chain shrinks the weighted error by the fraction $1/\bar{\mathsf{b}}$ of the worst remaining time. The effective horizon of a goal-reaching problem is the worst expected time to the goal.

Fact 5.4 (Optimality in SSP problems; Puterman and Chan, 2025, Thm. 6.10)**.** *In an SSP in which every improper stationary policy has infinite cost from some state (guaranteed by $c \geq c_{\min} > 0$ on C), $J^\star$ is finite, it is the unique solution of $V = T_c^\star V$ with $V(g) = 0$, there is an optimal proper stationary deterministic policy (greedy for $J^\star$), and value and policy iteration started at a proper policy converge to $J^\star$.*

Corollary 5.5 (The optimal mean hitting time)**.** *With unit costs $c = \mathbf{1}_{C\times\mathcal{A}}$, $J^\star(x) = \kappa^\star(x,g) := \inf_\pi \mathbb{E}_x^\pi \tau_g$ is the minimal expected time to the goal. It satisfies*

$$\kappa^\star(x,g) = 1 + \min_a \sum_{x'} P(x'\,|\,x,a)\kappa^\star(x',g)$$

for $x \neq g$, and $\kappa^\star(x,z) \leq \kappa^\star(x,y) + \kappa^\star(y,z)$ whenever every state can reach y and z: the optimal expected hitting time is a quasimetric.

Proof. The Bellman equation is [5.4] for unit costs. For the triangle inequality, follow an optimal policy for y until τ_y and then one for z: by the strong Markov property [1.2(c)] this history-dependent policy reaches z in expected time $\kappa^\star(x,y) + \kappa^\star(y,z)$, and $\kappa^\star$ is the infimum over all policies. ∎

5.2 Average Reward and Blackwell Optimality

Definition 5.6 (Gain and bias of a policy)**.** For a stationary π with state kernel P_π, the *gain* and *bias* are $\bar{r}^\pi := P_\pi^* r_\pi$ and $w^\pi := D_\pi r_\pi$ [2.9]. A policy is *gain optimal* if $\bar{r}^\pi \geq \bar{r}^{\pi'}$ for all stationary π', and *Blackwell optimal* if there is $\gamma_0 < 1$ such that π is γ-discount optimal for every $\gamma \in (\gamma_0, 1)$.

Proposition 5.7 (Gain as a long-run average)**.** *For stationary π,*

$$\bar{r}^\pi(x) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_x^\pi \sum_{t<T} r(Z_t) = \lim_{\gamma\uparrow 1}(1-\gamma)V_\gamma^\pi(x),$$

and

$$V_\gamma^\pi = \frac{\bar{r}^\pi}{1-\gamma} + w^\pi + O(1-\gamma),$$

where V_γ^π is the γ-discounted value.

Proof. $\mathbb{E}_x^\pi \sum_{t<T} r(Z_t) = \sum_{t<T} P_\pi^t r_\pi(x)$ [3.4(a)]; apply the Cesàro limit [2.8(c)], the Abel limit [2.8] and the Laurent expansion [2.10(c)] with $V_\gamma^\pi = U_\gamma r_\pi$ [4.3]. ∎

Theorem 5.8 (Blackwell optimality; Blackwell, 1962)**.**

(a) There is a Blackwell-optimal policy $\pi^\star \in \Pi^{\mathrm{SD}}$.

(b) Every Blackwell-optimal policy is gain optimal among stationary policies, and among gain-optimal policies it maximises the bias.

Proof.

(a) For $\pi, \pi' \in \Pi^{\mathrm{SD}}$ and each x, $\gamma \mapsto V_\gamma^{\pi}(x) - V_\gamma^{\pi'}(x)$ is a rational function on $[0, 1)$ (Cramer's rule, as in [2.8]); a nonzero rational function has finitely many zeros, so its sign is constant on some interval $(\gamma_{\pi\pi' x}, 1)$. With finitely many policies and states, there is $\gamma_0 < 1$ such that the order of all values $V_\gamma^{\pi}(x)$ is the same for all $\gamma \in (\gamma_0, 1)$. For each γ some deterministic stationary policy is γ-optimal at every state [4.5]; pick one for a fixed $\gamma_1 \in (\gamma_0, 1)$; since the order does not change, it is optimal for every $\gamma \in (\gamma_0, 1)$.

(b) For any stationary π, $V_\gamma^{\pi^\star} \geq V_\gamma^{\pi}$ near 1; multiply by $1-\gamma$ and let $\gamma \uparrow 1$ [5.7]: $\bar{r}^{\pi^\star} \geq \bar{r}^{\pi}$. If $\bar{r}^{\pi} = \bar{r}^{\pi^\star}$, then

$$V_\gamma^{\pi^\star} - V_\gamma^{\pi} = w^{\pi^\star} - w^{\pi} + O(1-\gamma) \geq 0$$

near 1, so $w^{\pi^\star} \geq w^{\pi}$.

■

Fact 5.9 (Average-reward optimality equation; Puterman and Chan, 2025, Ch. 8)**.** *If every stationary deterministic policy induces a unichain P_π (one closed class plus transient states), the optimal gain is a constant $\bar{r}^\star$, and there is $w \in \mathbb{R}^{\mathcal{X}}$ with*

$$\bar{r}^\star + w(x) = \max_a \Big[r(x,a) + \sum_{x'} P(x' \mid x, a) w(x') \Big];$$

a policy greedy for w in this equation is gain optimal.

5.3 The Running Examples

Example 5.10 (The goal CMP as a shortest path problem)**.** In the goal CMP [3.8] with unit costs, π_{f} and π_{u} are proper and π_{b} is improper ($\kappa^{\pi_{\mathrm{b}}}(1, g) = \infty$). The optimal costs are $\kappa^\star = (2, 1, 0)$, attained by π_{f}, while $\kappa^{\pi_{\mathrm{u}}} = (6, 4, 0)$ [2.12]. The model is an SSP but not a transient model, so [5.3] does not apply to all policies; value iteration from $V_0 = 0$ still converges, to $(2, 1, 0)$ after two steps. For the goal reward and the average criterion every proper policy has gain 1 and π_{b} has gain 0 at 1 and 2; among the proper policies the bias ranks them by speed, and π_{f} is Blackwell optimal [5.8].

Assumed without proof in this chapter

[5.4] optimality in SSP problems; [5.9] the average-reward optimality equation.

6

One Measure, Four Functionals

This chapter states the organising theorem of the book. For a stationary policy, the kernel of trajectory measures $z \mapsto \mathbb{P}_z^\pi$ is the unique fixed point of one operator on path space, the *trajectory Bellman operator* Φ^π, which prepends a step to a trajectory, and Φ^π contracts with modulus γ in a Wasserstein distance for the discounted path metric. Every quantity that reinforcement learning computes from a policy is the image of this fixed point under a functional of trajectories, and every Bellman equation of the book is the image of the single equation $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$. Four functionals recur: the return (values and return laws, Bellemare et al., 2017), the discounted occupancy (successor measures, Dayan, 1993), the hitting time and the discounted reach value (goal-conditioned values). The construction generalises the push-forward formulation of distributional reinforcement learning of Akgül et al. (2026, Sec. 3, Lemma 3.1) from return functionals to arbitrary functionals that intertwine with the prepending map. Throughout, $\mathcal{X}$ and $\mathcal{A}$ are countable, π is stationary, and $\Omega = \mathcal{Z}^{\mathbb{N}_0}$ is the path space of [3.3].

6.1 Path Metrics and Couplings

Definition 6.1 (Discounted path metric and transport cost)**.**

(a) The *discounted path metric* on Ω is

$$d_\gamma(\omega, \omega') := \sum_{t \geq 0} \gamma^t \mathbf{1}\{z_t \neq z'_t\} \leq \frac{1}{1-\gamma}.$$

(b) For $\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\Omega)$, a *coupling* is a $\varsigma \in \mathcal{P}(\Omega \times \Omega)$ with marginals $\mathbb{P}, \mathbb{P}'$, and the *transport cost* is

$$\mathsf{W}(\mathbb{P}, \mathbb{P}') := \inf_{\varsigma} \int d_\gamma \, d\varsigma.$$

(c) For a metric space (E, d_E) and $\eta, \eta' \in \mathcal{P}(E)$ with bounded support, $W_1(\eta, \eta') := \inf_\varsigma \int d_E \, d\varsigma$ over couplings ς of η, η' (the *Wasserstein-1 distance*).

(d) A *path kernel* is a family $\mathbb{Q} = (\mathbb{Q}_z)_{z\in\mathcal{Z}}$ of elements of $\mathscr{P}(\Omega)$;

$$\bar{\mathsf{W}}(\mathbb{Q},\mathbb{Q}') := \sup_z \mathsf{W}(\mathbb{Q}_z,\mathbb{Q}'_z).$$

Likewise, for families $\eta = (\eta_z)_z$ in $\mathscr{P}(E)$, $\bar{W}_1(\eta,\eta') := \sup_z W_1(\eta_z,\eta'_z)$.

Lemma 6.2 (Transport calculus)**.**

(a) (Push-forward.) If $f : (E,d_E) \to (E',d_{E'})$ is L-Lipschitz, then $W_1(f_\#\eta, f_\#\eta') \le L W_1(\eta,\eta')$; the same holds with W for $f : \Omega \to E'$ that is L-Lipschitz from d_γ.

(b) (Mixtures.) For a countable family and weights $p_i \ge 0$ with $\sum_i p_i = 1$,

$$W_1\Big(\sum_i p_i\eta_i, \sum_i p_i\eta'_i\Big) \le \sum_i p_i W_1(\eta_i,\eta'_i),$$

and likewise for W.

(c) (Separation.) $\mathsf{W}(\mathbb{P},\mathbb{P}') = 0$ implies $\mathbb{P} = \mathbb{P}'$; $W_1(\eta,\eta') = 0$ implies $\eta = \eta'$ for $E = \mathbb{R}$.

(d) (Means.) For $E = \mathbb{R}$,

$$\Big|\int g\,\eta(dg) - \int g\,\eta'(dg)\Big| \le W_1(\eta,\eta').$$

Proof.

(a) If ς couples η,η', then $(f\times f)_\#\varsigma$ couples $f_\#\eta, f_\#\eta'$ and costs

$$\int d_{E'}(f(u),f(u'))\varsigma(du,du') \le L\int d_E\,d\varsigma$$

[0.2(d)]. Take the infimum.

(b) Given $\epsilon > 0$ choose couplings ς_i of η_i,η'_i with cost $\le W_1(\eta_i,\eta'_i)+\epsilon$; $\sum_i p_i\varsigma_i$ couples the mixtures and costs $\le \sum_i p_i W_1(\eta_i,\eta'_i) + \epsilon$. Only countably many couplings are chosen, so no measurable selection is needed.

(c) Let ς couple $\mathbb{P},\mathbb{P}'$ with $\int d_\gamma\,d\varsigma \le \epsilon$. For every t,

$$\varsigma\{z_{0:t} \ne z'_{0:t}\} \le \sum_{s\le t}\varsigma\{z_s \ne z'_s\} \le \gamma^{-t}\epsilon,$$

so the laws of $Z_{0:t}$ under $\mathbb{P}$ and $\mathbb{P}'$ are within $\gamma^{-t}\epsilon$ in total variation [0.11(b)]. As $\epsilon \to 0$ they agree for every t, hence $\mathbb{P} = \mathbb{P}'$ [0.5]. For $E = \mathbb{R}$: for $c \in \mathbb{R}$ and $n \in \mathbb{N}$, $f_n(g) := (1 - n(g-c)^+)^+$ is n-Lipschitz, so $|\int f_n\,d\eta - \int f_n\,d\eta'| \le n\epsilon$ along couplings of cost ϵ; thus $\int f_n\,d\eta = \int f_n\,d\eta'$, and $f_n \downarrow \mathbf{1}_{(-\infty,c]}$ gives equal distribution functions [0.2(c)].

(d) For a coupling ς,

$$\left|\int g\,d\eta - \int g'\,d\eta'\right| = \left|\int (g-g')\,d\varsigma\right| \le \int |g-g'|\,d\varsigma.$$

■

Definition 6.3 (Transport costs)**.** Let (E,d) be a countable metric space with $\sup d < \infty$ and $p \in [1,\infty)$. For $\mu,\mu' \in \mathcal{P}(E)$,

$$W_p(\mu,\mu') := \inf_{\varsigma}\Big(\int d^p\,d\varsigma\Big)^{1/p}$$

over couplings ς [0.10]. On path space, W_p is defined in the same way with the discounted path metric d_γ [6.1], and $\bar{W}_p$, $\bar{\mathsf{W}}_p$ are the suprema over the starting pair. W_1 and W_1 are those of [6.1].

Lemma 6.4 (Gluing)**.** *W_p is a metric on $\mathcal{P}(E)$. More precisely, for couplings ς_1 of (μ,μ') and ς_2 of (μ',μ''),*

$$\varsigma(x,y,w) := \varsigma_1(x,y)\varsigma_2(y,w)/\mu'(y)$$

($:= 0$ if $\mu'(y) = 0$) is a probability on E^3 with pair marginals ς_1,ς_2, and its (x,w)-marginal couples μ,μ'' with

$$\Big(\int d(x,w)^p\Big)^{1/p} \le \Big(\int d(x,y)^p\Big)^{1/p} + \Big(\int d(y,w)^p\Big)^{1/p}.$$

Proof. Summing ς over w gives

$$\varsigma_1(x,y)\cdot\sum_w \varsigma_2(y,w)/\mu'(y) = \varsigma_1(x,y),$$

and over x gives ς_2. Then $d(x,w) \le d(x,y)+d(y,w)$ and Minkowski's inequality in $L^p(\varsigma)$ give the triangle inequality; taking infima over ς_1,ς_2 proves it for W_p. Symmetry is clear, and $W_p = 0$ forces equality as in [6.2(c)], since $W_1 \le W_p$ by Jensen and W_1 separates by the argument there (with d in place of $|\cdot|$, Lipschitz functions $(1 - n\,d(\cdot,x_0))^+$ separate points of a countable space). ■

6.2 The Trajectory Bellman Operator

Definition 6.5 (Prepending and the trajectory Bellman operator)**.**

(a) For $z \in \mathcal{Z}$, the *prepending map* $\mathrm{cons}_z : \Omega \to \Omega$ is $\mathrm{cons}_z(\omega) := (z,\omega_0,\omega_1,\dots)$. It is measurable, $\theta_1 \circ \mathrm{cons}_z = \mathrm{id}$, and $\omega = \mathrm{cons}_{z_0}(\theta_1\omega)$.

(b) The *trajectory Bellman operator* of π maps a path kernel $\mathbb{Q}$ to

$$(\Phi^\pi \mathbb{Q})_z := \sum_{z' \in \mathcal{Z}} (P \otimes \pi)(z, z') \, (\mathrm{cons}_z)_\# \mathbb{Q}_{z'}.$$

In words: draw the next pair z' from the chain, draw a trajectory from $\mathbb{Q}_{z'}$, and prepend z.

Theorem 6.6 (The trajectory measure is the fixed point of a contraction)**.** *Let $\mathbb{P}^\pi$ denote the path kernel $z \mapsto \mathbb{P}^\pi_z$ of [3.3].*

(a) (Fixed point.) $\Phi^\pi \mathbb{P}^\pi = \mathbb{P}^\pi$.

(b) (Contraction.)

$$\bar{\mathsf{W}}(\Phi^\pi \mathbb{Q}, \Phi^\pi \mathbb{Q}') \le \gamma \, \bar{\mathsf{W}}(\mathbb{Q}, \mathbb{Q}')$$

for all path kernels.

(c) (Uniqueness and convergence.) $\mathbb{P}^\pi$ is the only fixed point of Φ^π, and for every path kernel $\mathbb{Q}$,

$$\bar{\mathsf{W}}((\Phi^\pi)^k \mathbb{Q}, \mathbb{P}^\pi) \le \frac{\gamma^k}{1-\gamma}.$$

(d) (Exactness on finite windows.) For every $\mathbb{Q}$ and $k \in \mathbb{N}$, the law of $Z_{0:k-1}$ under $((\Phi^\pi)^k \mathbb{Q})_z$ equals its law under $\mathbb{P}^\pi_z$.

Proof.

(a) Let $F : \Omega \to \mathbb{R}$ be bounded measurable. Under $\mathbb{P}^\pi_z$, $Z_0 = z$ a.s., so $F = F \circ \mathrm{cons}_z \circ \theta_1$ a.s. By the Markov property [3.4(c)] at $s = 1$ and $Z_1 \sim (P \otimes \pi)(z, \cdot)$ [3.4(b)],

$$\begin{aligned} \mathbb{E}^\pi_z F &= \mathbb{E}^\pi_z \big[\mathbb{E}^\pi_{Z_1} (F \circ \mathrm{cons}_z) \big] \\ &= \sum_{z'} (P \otimes \pi)(z, z') \int F \circ \mathrm{cons}_z \, d\mathbb{P}^\pi_{z'} \\ &= \int F \, d(\Phi^\pi \mathbb{P}^\pi)_z. \end{aligned}$$

(b)

$$d_\gamma(\mathrm{cons}_z \omega, \mathrm{cons}_z \omega') = \gamma \, d_\gamma(\omega, \omega'),$$

so cons_z is γ-Lipschitz. By [6.2(b)] and [6.2(a)],

$$\begin{aligned} \mathsf{W}((\Phi^\pi \mathbb{Q})_z, (\Phi^\pi \mathbb{Q}')_z) &\le \sum_{z'} (P \otimes \pi)(z, z') \, \gamma \, \mathsf{W}(\mathbb{Q}_{z'}, \mathbb{Q}'_{z'}) \\ &\le \gamma \bar{\mathsf{W}}(\mathbb{Q}, \mathbb{Q}'). \end{aligned}$$

(c) $\bar{\mathsf{W}} \le \frac{1}{1-\gamma}$ because $d_\gamma \le \frac{1}{1-\gamma}$. If $\Phi^\pi \mathbb{Q} = \mathbb{Q}$, then by (a) and (b)

$$\bar{\mathsf{W}}(\mathbb{Q}, \mathbb{P}^\pi) \le \gamma \bar{\mathsf{W}}(\mathbb{Q}, \mathbb{P}^\pi),$$

so it is 0 and $\mathbb{Q} = \mathbb{P}^\pi$ [6.2(c)]. The rate follows by iterating (b) from (a). Neither completeness of the space of path kernels nor the triangle inequality for W is used; existence of the fixed point is supplied by Ionescu-Tulcea [3.3].

(d) Induction on k. Under $(\Phi^\pi \mathbb{Q})_z$, $Z_0 = z$ and $Z_{1:k}$ is $Z_{0:k-1}$ of a draw from $\mathbb{Q}_{Z_1}$ with $Z_1 \sim (P \otimes \pi)(z, \cdot)$; under $\mathbb{P}^\pi_z$ the same holds with $\mathbb{P}^\pi$ in place of $\mathbb{Q}$, by (a). Apply the induction hypothesis to $(\Phi^\pi)^{k-1}\mathbb{Q}$.

■

Intuition. The equation $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$ says: a trajectory from z is z followed by a trajectory from the next pair. It is the Bellman equation *before* any reward has been chosen and *before* any statistic of the trajectory has been taken. The contraction comes from the metric, which counts a disagreement at time t with weight γ^t: prepending a common step pushes every disagreement one step later.

6.3 Intertwining Functionals

Definition 6.7 (Intertwining functionals)**.** Let E be a measurable space. A measurable $F : \Omega \to E$ *intertwines* with a family of measurable maps $\beta_z : E \to E$, $z \in \mathcal{Z}$, if

$$F \circ \mathrm{cons}_z = \beta_z \circ F \qquad \text{for every } z \in \mathcal{Z}.$$

The associated *Bellman operator on laws* maps a family $\mathfrak{L} = (\mathfrak{L}_z)_z$ in $\mathcal{P}(E)$ to

$$(\mathfrak{B}_F \mathfrak{L})_z := \sum_{z'} (P \otimes \pi)(z, z') \, (\beta_z)_\# \mathfrak{L}_{z'},$$

and the *law of F under π* is $\mathfrak{L}^F_z := F_\# \mathbb{P}^\pi_z$.

Theorem 6.8 (Push-forward of the fixed-point equation)**.** *Let F intertwine with (β_z).*

(a)

$$F_\#(\Phi^\pi \mathbb{Q})_z = (\mathfrak{B}_F(F_\# \mathbb{Q}))_z$$

for every path kernel $\mathbb{Q}$; that is, $F_\# \circ \Phi^\pi = \mathfrak{B}_F \circ F_\#$.

(b) The law of F solves its Bellman equation: $\mathfrak{L}^F = \mathfrak{B}_F \mathfrak{L}^F$.

(c) *If E carries a metric d_E, every β_z is c-Lipschitz, and laws have bounded support, then*

$$\bar{W}_1(\mathfrak{B}_F\mathfrak{L}, \mathfrak{B}_F\mathfrak{L}') \le c\,\bar{W}_1(\mathfrak{L}, \mathfrak{L}').$$

If $c < 1$, $\mathfrak{L}^F$ is the unique fixed point of $\mathfrak{B}_F$ (for $E = \mathbb{R}$) and

$$\bar{W}_1(\mathfrak{B}_F^k\mathfrak{L}, \mathfrak{L}^F) \le c^k\,\bar{W}_1(\mathfrak{L}, \mathfrak{L}^F).$$

(d) *If F is L-Lipschitz from (Ω, d_γ) to (E, d_E), then*

$$\bar{W}_1(F_\#\mathbb{Q}, \mathfrak{L}^F) \le L\,\bar{\mathsf{W}}(\mathbb{Q}, \mathbb{P}^\pi)\,:$$

an error in the trajectory kernel bounds the error in every Lipschitz statistic.

Proof.

(a)
$$F_\#(\mathrm{cons}_z)_\#\mathbb{Q}_{z'} = (F \circ \mathrm{cons}_z)_\#\mathbb{Q}_{z'} = (\beta_z)_\#F_\#\mathbb{Q}_{z'},$$
and push-forward commutes with countable mixtures.

(b) Apply (a) to $\mathbb{Q} = \mathbb{P}^\pi$ and use [6.6(a)].

6.2(b) and [6.2(a)] as in [6.6(b)]. Uniqueness and the rate as in [6.6(c)], with [6.2(c)] for $E = \mathbb{R}$.

6.2(a) for each z.

■

6.4 The Four Functionals

Theorem 6.9 (First functional: the return; Bellemare et al., 2017)**.** *The return G [3.5] intertwines with the affine maps $\beta_z(g) := r(z) + \gamma g$ [3.6(b)]. Let $\eta_z^\pi := G_\#\mathbb{P}_z^\pi \in \mathcal{P}([0, \frac{1}{1-\gamma}])$ be the* return law *and $\mathcal{T}^\pi := \mathfrak{B}_G$ the* distributional Bellman operator,

$$(\mathcal{T}^\pi\eta)_z = \sum_{z'}(P \otimes \pi)(z, z')\,(\beta_z)_\#\eta_{z'}.$$

(a) *$\eta^\pi = \mathcal{T}^\pi\eta^\pi$, and η^π is the unique fixed point of $\mathcal{T}^\pi$ among families supported in $[0, \frac{1}{1-\gamma}]$.*

(b)
$$\bar{W}_1(\mathcal{T}^\pi\eta, \mathcal{T}^\pi\eta') \le \gamma\bar{W}_1(\eta, \eta').$$

(c) *G is $\mathrm{sp}(r)$-Lipschitz from d_γ, so*

$$\bar{W}_1(G_\#\mathbb{Q}, \eta^\pi) \le \mathrm{sp}(r)\,\bar{\mathsf{W}}(\mathbb{Q}, \mathbb{P}^\pi).$$

(d) (Expected values.) The mean of η_z^π is $Q^\pi(z)$, and taking means maps $\mathcal{T}^\pi$ to T^π: if $\bar\eta(z)$ is the mean of η_z, the mean of $(\mathcal{T}^\pi\eta)_z$ is $T^\pi\bar\eta(z)$.

Proof. (a)–(b): [6.8(b),(c)] with $c=\gamma$ (β_z has slope γ); $\mathcal{T}^\pi$ preserves support in $[0,\frac{1}{1-\gamma}]$ since $r\in[0,1]$. (c):

$$|G(\omega)-G(\omega')|\le\sum_t\gamma^t|r(z_t)-r(z_t')|\le \mathrm{sp}(r)d_\gamma(\omega,\omega');$$

apply [6.8(d)]. (d): $\int g\,\eta_z^\pi(dg)=\mathbb{E}_z^\pi G=Q^\pi(z)$ [0.2(d)], and the mean of $(\beta_z)_\#\eta_{z'}$ is $r(z)+\gamma\bar\eta(z')$. ∎

Theorem 6.10 (Second functional: the discounted occupancy; Dayan, 1993)**.** *Let $\mathsf{O}:\Omega\to\mathcal{M}(\mathcal{Z})$, $\mathsf{O}(\omega):=\sum_{t\ge0}\gamma^t\delta_{z_t}$, the* discounted occupancy *of a trajectory. It intertwines with $\beta_z(m):=\delta_z+\gamma m$. Its mean, the* successor measure

$$M^\pi(z,C):=\mathbb{E}_z^\pi\mathsf{O}(C)=\sum_t\gamma^t\mathbb{P}_z^\pi\{Z_t\in C\},$$

satisfies:

(a) $M^\pi=I+\gamma(P\otimes\pi)M^\pi$, and $M^\pi=\sum_t\gamma^t(P\otimes\pi)^t$ is the resolvent [2.6] of the state-action chain;

(b) for every reward r, $Q^\pi(z)=\int M^\pi(z,dz')r(z')$: the successor measure computes the action values of all *rewards at once;*

(c) the state successor measure $M_{\mathcal{X}}^\pi(x,B):=\sum_t\gamma^t\mathbb{P}_x^\pi\{X_t\in B\}$ is the resolvent of P_π, and $d_{\nu,\mathcal{X}}^\pi=(1-\gamma)\,\nu M_{\mathcal{X}}^\pi$, $d_\nu^\pi=(1-\gamma)(\nu\otimes\pi)M^\pi$;

(d) with the total-variation norm $\|m\|:=\sup_{\|f\|_\infty\le1}|mf|$ on $\mathcal{M}(\mathcal{Z})$,

$$\|\mathsf{O}(\omega)-\mathsf{O}(\omega')\|\le 2d_\gamma(\omega,\omega'),$$

and the map $M\mapsto I+\gamma(P\otimes\pi)M$ is a γ-contraction for $\sup_z\|M(z,\cdot)-M'(z,\cdot)\|$.

Proof.

$$\mathsf{O}(\mathrm{cons}_z\omega)=\delta_z+\gamma\mathsf{O}(\omega).$$

Taking expectations in [6.8(b)] (means are linear in the law, and $\mathsf{O}(C)\le\frac{1}{1-\gamma}$) gives

$$M^\pi(z,\cdot)=\delta_z+\gamma\sum_{z'}(P\otimes\pi)(z,z')M^\pi(z',\cdot),$$

which is (a); the series form is [3.4(b)] with [1.2(a)], and it is the resolvent by [2.7(a)]. (b): Fubini [0.2(e)] and [3.6(c)]. (c): the same computation for the state chain [3.4(a)], and [3.5(c)]. (d):

$$|\mathsf{O}(\omega)f - \mathsf{O}(\omega')f| \le \sum_t \gamma^t |f(z_t) - f(z'_t)| \le 2\|f\|_\infty d_\gamma(\omega, \omega');$$

the contraction follows from [0.4(d)]. ∎

Theorem 6.11 (Third and fourth functionals: hitting time and reach value)**.** *Fix a goal $g \in \mathcal{X}$ and let $\tau_g(\omega) := \inf\{t \ge 0 : x_t = g\}$.*

(a) τ_g intertwines with $\beta_z(k) := 0$ if $x = g$ and $\beta_z(k) := 1 + k$ otherwise. Its mean $\kappa^\pi(x, g) := \mathbb{E}^\pi_x \tau_g$ is the minimal solution of the Poisson equation $\kappa = 0$ at g, $\kappa = \mathbf{1} + P_\pi \kappa$ off g [2.3(b)], and $\kappa^\pi(\cdot, \cdot)$ is a quasimetric on $\mathcal{X}$ [2.5].

(b) γ^{τ_g} intertwines with $\beta_z(u) := 1$ if $x = g$ and $\beta_z(u) := \gamma u$ otherwise, which is γ-Lipschitz. Its mean $q_g^{\gamma,\pi}(x) := \mathbb{E}^\pi_x \gamma^{\tau_g}$ is the discounted reach value of the chain P_π [2.3(c)]; if g is absorbing and $r = \mathbf{1}\{x = g\}$, then $V^\pi = q_g^{\gamma,\pi}/(1 - \gamma)$.

(c) The maps β_z of (a) are 1-Lipschitz, not contractive: the hitting-time Bellman operator contracts only through the killed chain, with modulus governed by $\varrho(P_{\pi,C}) < 1$, $C = \mathcal{X} \setminus \{g\}$, when $\mathcal{X}$ is finite and g is accessible from every state [2.4(b)].

Proof. (a) If $x_0 = g$, $\tau_g(\mathrm{cons}_z \omega) = 0$; otherwise $\tau_g(\mathrm{cons}_z \omega) = 1 + \tau_g(\omega)$. The mean statements concern the state chain P_π [3.4(a)] and are [2.3(b)], [2.5]. (b) Similarly, with $\gamma^{1+k} = \gamma\gamma^k$; if g is absorbing, $G = \gamma^{\tau_g}/(1 - \gamma)$ as in [3.8]. (c) By inspection and [2.4(b)]. ∎

Corollary 6.12 (One measure, four functionals)**.** *For a stationary policy π:*

(a) the return law η^π, the successor measure M^π, the mean hitting times $\kappa^\pi(\cdot, g)$ and the reach values $q_g^{\gamma,\pi}$ are functionals of the single path kernel $\mathbb{P}^\pi$, and their Bellman equations [6.9(a)], [6.10(a)], [6.11(a),(b)] are the images of the one equation $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$ under push-forward [6.8(a)];

(b) values sit two levels below the trajectory measure: Q^π is the mean of η^π [6.9(d)] and a linear functional of M^π [6.10(b)];

(c) (Observation processes.) For a measurable f from $\mathcal{Z}$ to a countable $\mathcal{Y}$, the sequence map $F(\omega) := (f(z_t))_{t \ge 0}$ intertwines with $\beta_z := \mathrm{cons}_{f(z)}$ on $\mathcal{Y}^{\mathbb{N}_0}$, which is γ-Lipschitz for the discounted path metric on $\mathcal{Y}^{\mathbb{N}_0}$; so the law of the observation process is the fixed point of a γ-contraction on observation

path space. This is the push-forward process functional of Akgül et al. (2026, Sec. 3).

Proof. (a) and (b) collect [6.8]–[6.11]. (c): $F(\mathrm{cons}_z\omega) = (f(z), f(z_0), f(z_1), \dots)$, and prepending a common symbol multiplies the discounted path metric by γ; apply [6.8(b),(c)] with the separation argument of [6.2(c)] on $\mathcal{Y}^{\mathbb{N}_0}$. ■

Remark 6.13 (On the encoder–decoder modulus)**.** Akgül et al. (2026, Thm. 3.5) compose the transition kernel with an L_E-Lipschitz encoder and an L_D-Lipschitz decoder kernel and obtain the modulus $\gamma L_E L_D$ for the distributional Bellman operator on the transformed processes. Two points of the appendix proof are repaired in Part IV. In Lemma A.1 the bound $W_p^p(KP, KP') \le L\,\mathbb{E}|s - s'|^p$ should read $L^p\,\mathbb{E}|s - s'|^p$; after taking p-th roots the stated conclusion $W_p(KP, KP') \le L W_p(P, P')$ is correct. Lemma A.2 couples two path measures through a single path-space coupling and then applies the kernel bound step by step; this requires gluing per-step couplings along the path, which [6.2(b)] supplies on countable spaces and which needs a measurable selection of optimal couplings in general.

6.5 The Running Examples

Example 6.14 (All four functionals of the goal chain)**.** Consider the goal CMP [3.8] under the uniform policy π_{u}, started at $x = 1$, with $\gamma = 0.9$. Let (F_n) be the Fibonacci numbers, $F_0 = 0, F_1 = 1, F_{n+1} = F_n + F_{n-1}$.

(a) (Hitting time.) $\mathbb{P}_1^{\pi_{\mathrm{u}}}\{\tau_g = t\} = F_{t-1}/2^t$ for $t \ge 1$: the paths from 1 to g of length t that avoid g before t correspond to compositions of $t - 2$ into parts 1 and 2 (a loop $1 \to 1$, or an excursion $1 \to 2 \to 1$), and each has probability 2^{-t}. Its mean is $\kappa^{\pi_{\mathrm{u}}}(1, g) = 6$ [2.12].

(b) (Reach value and return law.)

$$q_g^{\gamma,\pi_{\mathrm{u}}}(1) = \sum_t F_{t-1}\gamma^t/2^t = \gamma^2/(4 - 2\gamma - \gamma^2) \approx 0.583,$$

and, since $G = \gamma^{\tau_g}/(1 - \gamma)$, the law of G under $\mathbb{P}_1^{\pi_{\mathrm{u}}}$ (the mixture of $\eta_{(1,a)}^{\pi_{\mathrm{u}}}$ over the first action) is

$$\sum_{t \ge 1} \frac{F_{t-1}}{2^t}\,\delta_{\gamma^t/(1-\gamma)},$$

an atomic law with mean $V^{\pi_{\mathrm{u}}}(1) \approx 5.83$ [3.8].

(c) (Successor measure.)

$$M_{\mathcal{X}}^{\pi_{\mathrm{u}}}(1,\cdot) \approx (2.878, 1.295, 5.827)$$

on $(1, 2, g)$, total mass $\frac{1}{1-\gamma} = 10$; its last entry is $V^{\pi_{\mathrm{u}}}(1)$ for the goal reward [6.10(b),(c)].

The four objects are different statistics of one law on trajectories: the hitting-time law determines the return law here because the reward is a goal indicator, but neither the mean hitting time nor the successor measure determines the return law, and the return law does not determine the successor measure.

Assumed without proof in this chapter

Nothing beyond Chapters 0–4.

Part II

Learning to Control: The Classical Axis

7

Stochastic Approximation

Dynamic programming needs the transition kernel. Learning replaces each application of a Bellman operator by a noisy sample of it and averages the samples with decreasing step sizes. The theory of such recursions is *stochastic approximation*, founded by Robbins and Monro (1951). This chapter proves the two convergence theorems that the classical theory of temporal-difference learning and Q-learning rests on: convergence of a recursion with a strongly monotone mean field (Euclidean), and convergence of an asynchronous recursion whose mean field is a sup-norm contraction (Jaakkola et al., 1994; Tsitsiklis, 1994). The general ODE method of Borkar (2023) is stated as a Fact.

7.1 Robbins–Monro Recursions

Definition 7.1 (Stochastic approximation)**.** Let $(\mathcal{F}_k)_{k\ge 0}$ be a filtration. A *stochastic-approximation recursion* in $\mathbb{R}^d$ is

$$\vartheta_{k+1} = \vartheta_k + \alpha_k\big(h(\vartheta_k) + \zeta_{k+1}\big), \qquad k \ge 0,$$

with ϑ_0 $\mathcal{F}_0$-measurable, a measurable *mean field* $h : \mathbb{R}^d \to \mathbb{R}^d$, $\mathcal{F}_k$-measurable *step sizes* $\alpha_k \ge 0$, and *martingale noise*: ζ_{k+1} is $\mathcal{F}_{k+1}$-measurable with $\mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0$. The step sizes satisfy the *Robbins–Monro conditions* if $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$ almost surely.

Fact 7.2 (Almost supermartingales; Robbins and Siegmund, 1971)**.** *Let $Y_k, a_k, b_k, c_k \ge 0$ be $\mathcal{F}_k$-measurable with $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] \le (1 + a_k)Y_k - b_k + c_k$ for all k, and $\sum_k a_k < \infty$, $\sum_k c_k < \infty$ almost surely. Then Y_k converges almost surely to a finite limit, and $\sum_k b_k < \infty$ almost surely. (For $a_k = b_k = c_k = 0$ this is the convergence of nonnegative supermartingales, Çınlar, 2011, Ch. V.4.)*

Theorem 7.3 (Robbins–Monro under strong monotonicity; Robbins and Monro, 1951)**.** *In [7.1], suppose there are $\vartheta^\star \in \mathbb{R}^d$ and constants $c, K > 0$ with*

$$\begin{aligned}\langle \vartheta - \vartheta^\star, h(\vartheta)\rangle &\le -c\|\vartheta - \vartheta^\star\|^2, \quad \|h(\vartheta)\|^2 \\ &\le K(1 + \|\vartheta - \vartheta^\star\|^2), \quad \mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k] \\ &\le K(1 + \|\vartheta_k - \vartheta^\star\|^2),\end{aligned}$$

and the Robbins–Monro conditions. Then $\vartheta_k \to \vartheta^\star$ almost surely.

Proof. Let $Y_k := \|\vartheta_k - \vartheta^\star\|^2$. Expanding the square and using $\mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0$,

$$\begin{aligned}\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] &= Y_k + 2\alpha_k \langle \vartheta_k - \vartheta^\star, h(\vartheta_k)\rangle + \alpha_k^2 \mathbb{E}[\|h(\vartheta_k) + \zeta_{k+1}\|^2 \mid \mathcal{F}_k] \\ &\le (1 + 2K\alpha_k^2) Y_k - 2c\alpha_k Y_k + 2K\alpha_k^2.\end{aligned}$$

By [7.2] with $a_k = 2K\alpha_k^2$, $b_k = 2c\alpha_k Y_k$, $c_k = 2K\alpha_k^2$, $Y_k \to Y_\infty$ and $\sum_k \alpha_k Y_k < \infty$ a.s. Since $\sum_k \alpha_k = \infty$, $\liminf_k Y_k = 0$, so $Y_\infty = 0$. $\blacksquare$

7.2 Asynchronous Contractions

Lemma 7.4 (Averaging of martingale noise)**.** *Let $W_0 := 0$ and $W_{k+1} := (1 - \alpha_k) W_k + \alpha_k w_k$ in $\mathbb{R}$, with $\mathcal{F}_k$-measurable $\alpha_k \in [0, 1]$ satisfying the Robbins–Monro conditions, $\mathcal{F}_{k+1}$-measurable w_k, $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[w_k^2 \mid \mathcal{F}_k] \le C$. Then $W_k \to 0$ almost surely. The same holds if the recursion is started at a fixed time k_0 instead of 0.*

Proof.

$$\mathbb{E}[W_{k+1}^2 \mid \mathcal{F}_k] = (1 - \alpha_k)^2 W_k^2 + \alpha_k^2 \mathbb{E}[w_k^2 \mid \mathcal{F}_k] \le (1 + \alpha_k^2) W_k^2 - 2\alpha_k W_k^2 + C\alpha_k^2.$$

By [7.2], W_k^2 converges and $\sum_k \alpha_k W_k^2 < \infty$; as $\sum_k \alpha_k = \infty$ the limit is 0. $\blacksquare$

Theorem 7.5 (Asynchronous stochastic approximation; Jaakkola et al., 1994; Tsitsiklis, 1994)**.** *Let I be a finite index set and, for $i \in \mathsf{I}$,*

$$\Delta_{k+1}(i) = \big(1 - \alpha_k(i)\big)\Delta_k(i) + \alpha_k(i)\, \Upsilon_k(i),$$

where $\alpha_k(i) \in [0, 1]$ are $\mathcal{F}_k$-measurable and satisfy the Robbins–Monro conditions for each i, Υ_k is $\mathcal{F}_{k+1}$-measurable, and, for some $c \in [0, 1)$ and $B, C < \infty$:

(a) (Contraction in mean.)

$$\max_i \big|\mathbb{E}[\Upsilon_k(i) \mid \mathcal{F}_k]\big| \le c \max_i |\Delta_k(i)|;$$

(b) (Bounded noise.)

$$\mathbb{E}\big[(\Upsilon_k(i) - \mathbb{E}[\Upsilon_k(i) \mid \mathcal{F}_k])^2 \mid \mathcal{F}_k\big] \le C;$$

(c) (Bounded iterates.) $\max_i |\Delta_k(i)| \le B$ *for all k, almost surely.*

Then $\Delta_k \to 0$ almost surely.

Proof. Write $\|\Delta\| := \max_i |\Delta(i)|$, $\bar{\Upsilon}_k(i) := \mathbb{E}[\Upsilon_k(i) \mid \mathcal{F}_k]$, $w_k(i) := \Upsilon_k(i) - \bar{\Upsilon}_k(i)$. For each fixed start time k_0 and each i, let $W_k^{k_0}(i)$ be the noise recursion of [7.4] driven by $w_k(i)$ and $\alpha_k(i)$, started at $W_{k_0}^{k_0}(i) = 0$. By [7.4], almost surely $W_k^{k_0}(i) \to 0$ as $k \to \infty$, simultaneously for the countably many pairs (k_0, i). Fix such an outcome.

Let $D_0 := B$ and $D_{j+1} := \frac{1+c}{2} D_j$, so $D_j \downarrow 0$. *Claim:* if $\|\Delta_k\| \le D_j$ for all $k \ge k_0$, then $\|\Delta_k\| \le D_{j+1}$ for all large k. Define $Y_{k_0}(i) := D_j$ and

$$Y_{k+1}(i) := (1 - \alpha_k(i)) Y_k(i) + \alpha_k(i) c D_j.$$

By induction on $k \ge k_0$, using (a),

$$-Y_k(i) + W_k^{k_0}(i) \le \Delta_k(i) \le Y_k(i) + W_k^{k_0}(i) :$$

for instance

$$\Delta_{k+1}(i) \le (1-\alpha_k(i))(Y_k(i) + W_k^{k_0}(i)) + \alpha_k(i)(cD_j + w_k(i)) = Y_{k+1}(i) + W_{k+1}^{k_0}(i).$$

Now

$$Y_k(i) - cD_j = (1-c) D_j \prod_{l=k_0}^{k-1} (1 - \alpha_l(i)) \to 0,$$

because $\sum_l \alpha_l(i) = \infty$ and $1 - u \le e^{-u}$. Hence $\limsup_k \|\Delta_k\| \le cD_j < D_{j+1}$, which proves the claim. Starting from (c) with $j = 0$, $k_0 = 0$, induction on j gives $\limsup_k \|\Delta_k\| \le D_j$ for every j. ∎

Intuition. The mean of the update pulls every coordinate into a box shrunk by the factor c; the noise is averaged away by the decreasing steps. Coordinates may be updated at different times and rates (asynchrony), as long as each is updated infinitely often with Robbins–Monro steps. This is exactly the situation of Q-learning, where only the visited state-action pair is updated.

7.3 The ODE Method

Fact 7.6 (The ODE method; Borkar, 2023, Thm. 2.1 and Cor. 2.1)**.** *In [7.1], let the step sizes be deterministic and satisfy the Robbins–Monro conditions, h be Lipschitz,*

$$\mathbb{E}[\|\zeta_{k+1}\|^2 \mid \mathcal{F}_k] \le K(1 + \|\vartheta_k\|^2),$$

and $\sup_k \|\vartheta_k\| < \infty$ almost surely. Suppose the ODE $\dot{\vartheta} = h(\vartheta)$ has a continuously differentiable Lyapunov function $\mathsf{L} \ge 0$ with $\mathsf{L}(\vartheta) \to \infty$ as $\|\vartheta\| \to \infty$ and

$\langle \nabla L(\vartheta), h(\vartheta) \rangle < 0$ *for* $\vartheta \neq \vartheta^*$. *Then* $\vartheta_k \to \vartheta^*$ *almost surely. The iterates track the ODE solutions on time intervals of length* $\sum \alpha_k$, *and the noise is asymptotically negligible.*

7.4 The Running Examples

Example 7.7 (Sample means and a reach probability)**.**

(a) With $h(\vartheta) = m - \vartheta$, $\zeta_{k+1} = Y_{k+1} - m$ for i.i.d. Y_k with mean m and finite variance, and $\alpha_k = 1/(k+1)$, the recursion [7.1] started at $\vartheta_0 = 0$ gives $\vartheta_k = \frac{1}{k} \sum_{j \le k} Y_j$, the sample mean, and [7.3] reproves the strong law in this case ($c = 1$).

(b) In the goal chain [1.24], let $Y_k := \gamma^{\tau_g}$ for independent trajectories from state 1. The recursion of (a) converges almost surely to $q_g^{\gamma}(1) \approx 0.583$ [2.12]. This estimator needs complete trajectories. The temporal-difference estimators of the next chapter replace Y_k by a one-step bootstrap and update every visited state, which is the setting of [7.5].

Assumed without proof in this chapter

[7.2] the Robbins–Siegmund theorem on almost supermartingales; [7.6] the ODE method.

8

Temporal-Difference Learning and Q-Learning

Temporal-difference learning (Sutton, 1988) and Q-learning (Watkins and Dayan, 1992) are the two algorithms with which reinforcement learning became a theory of its own. Each replaces a Bellman operator by a one-transition sample of it and updates only the visited coordinate. With the asynchronous stochastic-approximation theorem [7.5] both convergence proofs reduce to one line each: the mean of the sampled update is a γ-contraction of the error (Jaakkola et al., 1994; Tsitsiklis, 1994). The chapter also treats multi-step returns and the λ-operator, and it isolates the role of exploration. $\mathcal{X}$ and $\mathcal{A}$ are finite.

8.1 Temporal-Difference Learning

Definition 8.1 (TD(0))**.** Let π be stationary and $(Z_k)_{k\ge 0} = (X_k, A_k)_k$ a trajectory under $\mathbb{P}_\nu^\pi$. With step sizes $\alpha_k(x) \in [0,1]$, $\alpha_k(x) = 0$ unless $X_k = x$, and $\mathcal{F}_k := \sigma(Z_0, \dots, Z_{k-1}, X_k)$-measurable, *TD(0)* updates $V_k \in \mathbb{R}^{\mathcal{X}}$ by

$$V_{k+1}(x) = V_k(x) + \alpha_k(x)\,\delta_k, \qquad \delta_k := R_k + \gamma V_k(X_{k+1}) - V_k(X_k),$$

where δ_k is the *temporal-difference error*. For instance $\alpha_k(x) := \mathbf{1}\{X_k = x\}/N_k(x)$ with $N_k(x)$ the number of visits to x up to time k.

Theorem 8.2 (Convergence of TD(0); Sutton, 1988; Jaakkola et al., 1994)**.** *Suppose $V_0 \in [0, \frac{1}{1-\gamma}]^{\mathcal{X}}$ and, almost surely, $\sum_k \alpha_k(x) = \infty$ and $\sum_k \alpha_k(x)^2 < \infty$ for every x (this requires each state to be visited infinitely often, for example P_π irreducible [1.8]). Then $V_k \to V^\pi$ almost surely.*

Proof. Bounded iterates. $V_{k+1}(X_k)$ is a convex combination of $V_k(X_k)$ and $R_k + \gamma V_k(X_{k+1})$, so by induction $V_k \in [0, \frac{1}{1-\gamma}]^{\mathcal{X}}$. *Reduction to [7.5].* Let $\Delta_k := V_k - V^\pi$ and

$$\Upsilon_k(x) := \mathbf{1}\{X_k = x\}\big(R_k + \gamma V_k(X_{k+1}) - V^\pi(x)\big);$$

then

$$\Delta_{k+1}(x) = (1 - \alpha_k(x))\Delta_k(x) + \alpha_k(x)\Upsilon_k(x)$$

for every x, because $\alpha_k(x) = 0$ off $\{X_k = x\}$. Given $\mathcal{F}_k$, $A_k \sim \pi(\cdot|X_k)$ and $X_{k+1} \sim P(\cdot|Z_k)$ [3.3(b),(c)], so by [4.3]

$$\mathbb{E}[\Upsilon_k(x) \mid \mathcal{F}_k] = \mathbf{1}\{X_k = x\}\big(r_\pi(x) + \gamma P_\pi V_k(x) - V^\pi(x)\big) = \mathbf{1}\{X_k = x\}\,\gamma P_\pi \Delta_k(x),$$

whose modulus is at most $\gamma \max_y |\Delta_k(y)|$. The noise is bounded since all terms lie in $[-\frac{2}{1-\gamma}, \frac{2}{1-\gamma}]$. Apply [7.5] with $c = \gamma$. ∎

Proposition 8.3 (Multi-step returns and the λ-operator)**.** *Let π be stationary and $\lambda \in [0,1]$.*

(a) For $n \geq 1$,

$$(T^\pi)^n V(x) = \mathbb{E}_x^\pi\Big[\sum_{t<n} \gamma^t R_t + \gamma^n V(X_n)\Big]$$

(the n-step return).

(b) The λ-operator

$$T_\lambda^\pi V := (1-\lambda)\sum_{n\geq 0} \lambda^n (T^\pi)^{n+1} V$$

(for $\lambda < 1$; $T_1^\pi V := V^\pi$) has V^π as its unique fixed point and is a contraction with modulus $\frac{\gamma(1-\lambda)}{1-\gamma\lambda} \leq \gamma$.

(c)

$$T_\lambda^\pi V(x) = V(x) + \mathbb{E}_x^\pi \sum_{t\geq 0} (\gamma\lambda)^t \delta_t(V),$$

with $\delta_t(V) := R_t + \gamma V(X_{t+1}) - V(X_t)$: the λ-return is the current estimate plus the $\gamma\lambda$-discounted sum of future TD errors.

Proof.

(a) Induction on n with $T^\pi V = r_\pi + \gamma P_\pi V$ and $\mathbb{E}_x^\pi f(X_t) = P_\pi^t f(x)$ [3.4(a)], [1.2(a)].

(b) By (a),

$$(T^\pi)^{n+1} V - (T^\pi)^{n+1} V' = \gamma^{n+1} P_\pi^{n+1}(V - V').$$

Hence

$$\|T_\lambda^\pi V - T_\lambda^\pi V'\|_\infty \leq (1-\lambda)\sum_n \lambda^n \gamma^{n+1} \|V - V'\|_\infty = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|V - V'\|_\infty;$$

V^π is a fixed point of each $(T^\pi)^{n+1}$ [4.3]. The modulus is $\leq \gamma$ because $1-\lambda \leq 1-\gamma\lambda$.

(c) Expand $\sum_t (\gamma\lambda)^t \delta_t(V)$, telescope the V-terms with weights $(\gamma\lambda)^t$, and compare coefficients with

$$(1-\lambda)\sum_n \lambda^n\Big[\sum_{t\leq n} \gamma^t R_t + \gamma^{n+1} V(X_{n+1})\Big]:$$

both equal

$$\sum_t (\gamma\lambda)^t R_t + \sum_{t\geq 1} (\gamma\lambda)^{t-1}\gamma(1-\lambda) V(X_t) - V(X_0) + V(X_0),$$

after exchanging sums by dominated convergence [0.2(c)].

■

Intuition. $\lambda = 0$ bootstraps after one step (TD(0)); $\lambda = 1$ uses the whole return (Monte Carlo). Intermediate λ trades the bias of bootstrapping against the variance of long returns. The contraction modulus improves as $\lambda \to 1$, but each sample becomes noisier.

8.2 Q-Learning

Definition 8.4 (Q-learning)**.** Let (Z_k) be generated by an arbitrary behaviour policy $\pi_{\mathrm{b}} \in \Pi^{\mathrm{HR}}$ under $\mathbb{P}_\nu^{\pi_{\mathrm{b}}}$, and $\mathscr{F}_k := \sigma(Z_0, \dots, Z_k)$. With $\mathscr{F}_k$-measurable step sizes $\alpha_k(z) \in [0,1]$, $\alpha_k(z) = 0$ unless $Z_k = z$, *Q-learning* updates

$$Q_{k+1}(z) = Q_k(z) + \alpha_k(z)\Big(R_k + \gamma \max_{a'} Q_k(X_{k+1}, a') - Q_k(z)\Big).$$

Theorem 8.5 (Convergence of Q-learning; Watkins and Dayan, 1992)**.** *Suppose $Q_0 \in [0, \frac{1}{1-\gamma}]^{\mathscr{Z}}$ and, almost surely, $\sum_k \alpha_k(z) = \infty$, $\sum_k \alpha_k(z)^2 < \infty$ for every $z \in \mathscr{Z}$. Then $Q_k \to Q^\star$ almost surely, whatever the behaviour policy.*

Proof. As in [8.2], $Q_k \in [0, \frac{1}{1-\gamma}]^{\mathscr{Z}}$ for all k. Put $\Delta_k := Q_k - Q^\star$ and

$$\Upsilon_k(z) := \mathbf{1}\{Z_k = z\}\big(R_k + \gamma \max_{a'} Q_k(X_{k+1}, a') - Q^\star(z)\big).$$

Given $\mathscr{F}_k$, $X_{k+1} \sim P(\cdot|Z_k)$ [3.3(c)], so by $Q^\star = T^\star Q^\star$ [4.5],

$$\begin{aligned}\big|\mathbb{E}[\Upsilon_k(z) \mid \mathscr{F}_k]\big| &= \mathbf{1}\{Z_k \\ &= z\}\Big|\gamma \int P(dx'|z)\Big(\max_{a'} Q_k(x', a') - \max_{a'} Q^\star(x', a')\Big)\Big| \\ &\le \gamma \max_{z'} |\Delta_k(z')|.\end{aligned}$$

The noise is bounded. Apply [7.5] with $c = \gamma$. ■

Fact 8.6 (On-policy control; Singh et al., 2000)**.** SARSA *replaces* $\max_{a'} Q_k(X_{k+1}, a')$ *in [8.4] by* $Q_k(X_{k+1}, A_{k+1})$*, where* A_{k+1} *is drawn from a policy that depends on* Q_k*. If that policy is* greedy in the limit with infinite exploration *(every pair is tried infinitely often, and the probability of a non-greedy action tends to* 0*), and the step sizes satisfy the conditions of [8.5], then* $Q_k \to Q^\star$ *almost surely.*

Remark 8.7 (What the asymptotic theory does not say)**.** Theorems [8.2] and [8.5] are asymptotic and require that every coordinate be visited infinitely often. They say nothing about how many samples are needed, and the visitation condition is an assumption about the behaviour policy, not a property the algorithm ensures. Both gaps are the subject of Part III: finite-sample rates in the generative model, and exploration that the learner must create itself.

8.3 Overestimation and Double Estimation

Lemma 8.8 (The maximum of noisy estimates; Thrun and Schwartz, 1993)**.** *Let $\mu \in \mathbb{R}^n$ and $Y_a = \mu_a + \xi_a$, $a \le n$, with $\mathbb{E}\xi_a = 0$.*

(a) $\mathbb{E}\max_a Y_a \ge \max_a \mu_a$, with equality iff there is $a^\star \in \arg\max \mu$ with $\max_a Y_a = Y_{a^\star}$ almost surely.

(b) If $\mu_a = m$ for all a and the ξ_a are i.i.d. uniform on $[-\epsilon, \epsilon]$, then $\mathbb{E}\max_a Y_a = m + \epsilon\frac{n-1}{n+1}$.

(c) If $n = 2$, $\mu_1 = \mu_2 = m$ and ξ_1, ξ_2 are i.i.d. $\mathrm{N}(0, \sigma^2)$, then $\mathbb{E}\max_a Y_a = m + \sigma/\sqrt{\pi}$.

Proof.

(a) $\max_a Y_a \ge Y_{a^\star}$ for any fixed $a^\star \in \arg\max \mu$, and $\mathbb{E}Y_{a^\star} = \max \mu$. Equality forces $\max_a Y_a = Y_{a^\star}$ a.s.

(b) With $U_a := (\xi_a + \epsilon)/(2\epsilon)$ i.i.d. uniform on $[0,1]$, $\mathbb{P}\{\max U_a \le u\} = u^n$, so $\mathbb{E}\max U_a = \int_0^1 (1 - u^n)\,du = \frac{n}{n+1}$, and

$$\mathbb{E}\max \xi_a = 2\epsilon\frac{n}{n+1} - \epsilon = \epsilon\frac{n-1}{n+1}.$$

(c)

$$\max(\xi_1, \xi_2) = \frac{\xi_1 + \xi_2}{2} + \frac{|\xi_1 - \xi_2|}{2}$$

and $\xi_1 - \xi_2 \sim \mathrm{N}(0, 2\sigma^2)$, whose absolute value has mean $\sqrt{2\sigma^2}\sqrt{2/\pi} = 2\sigma/\sqrt{\pi}$.

∎

Theorem 8.9 (Compounded overestimation)**.** *Let $(\xi(x'))_{x' \in \mathcal{X}}$ be random vectors in $\mathbb{R}^{\mathcal{A}}$ with $\mathbb{E}\xi_a(x') = 0$ and $|\xi_a(x')| \le \epsilon$, and define the* noisy optimality operator

$$\tilde{T}Q(z) := r(z) + \gamma \sum_{x'} P(x'|z)\, \mathbb{E}\max_a \big(Q(x', a) + \xi_a(x')\big).$$

(a) *$\tilde{T}$ is monotone and a γ-contraction; let $\tilde{Q}$ be its fixed point.*

(b) *$Q^\star \le \tilde{Q} \le Q^\star + \frac{\gamma \bar{c}}{1-\gamma}$, where $\bar{c} := \max_{x'} \mathbb{E} \max_a \xi_a(x')$ $\le \epsilon$.*

(c) *If every $Q^\star(x', \cdot)$ is constant in the action and $\mathbb{E} \max_a \xi_a(x') = c$ for all x', then $\tilde{Q} = Q^\star + \frac{\gamma c}{1-\gamma}$; with i.i.d. uniform noise, $c = \epsilon \frac{|\mathcal{A}|-1}{|\mathcal{A}|+1}$ [8.8(b)].*

Proof.

(a) $Q \mapsto \mathbb{E} \max_a (Q(x', a) + \xi_a)$ is monotone and commutes with adding constants; proceed as in [4.2].

(b) By [8.8(a)] pointwise, $\tilde{T}Q \ge T^\star Q$; also $\mathbb{E} \max_a (Q + \xi) \le \max_a Q + \mathbb{E} \max_a \xi$, so $\tilde{T}Q \le T^\star Q + \gamma \bar{c}$. From $\tilde{Q} = \tilde{T}\tilde{Q} \ge T^\star \tilde{Q}$ and monotonicity of $T^\star$, $\tilde{Q} \ge (T^\star)^k \tilde{Q} \to Q^\star$ [4.6(a)]. From $\tilde{Q} \le T^\star \tilde{Q} + \gamma \bar{c}$,

$$\tilde{Q} - Q^\star \le T^\star \tilde{Q} - T^\star Q^\star + \gamma \bar{c} \le \gamma \|\tilde{Q} - Q^\star\|_\infty + \gamma \bar{c},$$

and the left side is nonnegative.

(c) For $Q = Q^\star + k\mathbf{1}$ with $Q(x', \cdot)$ constant, $\tilde{T}Q = T^\star Q^\star + \gamma(k + c) = Q^\star + \gamma(k + c)$; this is a fixed point iff $k = \gamma(k + c)$.

■

Intuition. Q-learning bootstraps from $\max_a Q_k(X_{k+1}, a)$, a maximum of noisy estimates. Each backup adds a nonnegative bias of the size of the noise, and the bias is propagated with discount γ, so it accumulates to $\frac{\gamma}{1-\gamma}$ times the per-step bias. In a goal-reaching task the values along an optimal path differ by factors of γ, and a bias of this size can reverse their order, which is the failure mode identified by Thrun and Schwartz (1993).

Lemma 8.10 (The double estimator; van Hasselt, 2010, Lemma 1)**.** *Let $(Y_a^{\mathsf{A}})_a$ and $(Y_a^{\mathsf{B}})_a$ be independent families with $\mathbb{E}Y_a^{\mathsf{A}} = \mathbb{E}Y_a^{\mathsf{B}} = \mu_a$, and $\hat{a} \in \arg\max_a Y_a^{\mathsf{A}}$ (measurably chosen). Then $\mathbb{E}Y_{\hat{a}}^{\mathsf{B}} = \mathbb{E}\mu_{\hat{a}} \le \max_a \mu_a$, with strict inequality iff $\mathbb{P}\{\hat{a} \notin \arg\max \mu\} > 0$.*

Proof. $\hat{a}$ is a function of Y^{A}, independent of Y^{B}, so

$$\mathbb{E}[Y_{\hat{a}}^{\mathsf{B}} \mid Y^{\mathsf{A}}] = \mu_{\hat{a}}$$

[0.6(a)], and $\mu_{\hat{a}} \le \max \mu$ with equality exactly on $\{\hat{a} \in \arg\max \mu\}$. ■

Theorem 8.11 (Double estimation never overestimates)**.** *With $\xi^{\mathsf{A}}(x'), \xi^{\mathsf{B}}(x')$ independent zero-mean noise vectors and*

$$\hat{a}(x') \in \arg\max_a \big(Q(x', a) + \xi_a^{\mathsf{A}}(x')\big),$$

define

$$\tilde{T}_2 Q(z) := r(z) + \gamma \sum_{x'} P(x'|z)\, \mathbb{E}\big[Q(x', \hat{a}(x')) + \xi^{\mathsf{B}}_{\hat{a}(x')}(x')\big].$$

Then $\tilde{T}_2 Q \le T^* Q$ for every Q, and every bounded fixed point Q of $\tilde{T}_2$ satisfies $Q \le Q^*$. If moreover $Q^*(x', \cdot)$ has a unique maximiser with gap larger than 2ϵ at every x' and $|\xi^{\mathsf{A}}| \le \epsilon$, then $\tilde{T}_2 Q^* = Q^*$.

Proof. By [8.10] applied to $Y^{\mathsf{A}} = Q(x', \cdot) + \xi^{\mathsf{A}}$, $Y^{\mathsf{B}} = Q(x', \cdot) + \xi^{\mathsf{B}}$, the expectation is at most $\max_a Q(x', a)$, so $\tilde{T}_2 Q \le T^* Q$. If $Q = \tilde{T}_2 Q \le T^* Q$, then $Q \le (T^*)^k Q \to Q^*$ by monotonicity [4.2(a)] and [4.6(a)]. Under the gap condition $\hat{a}(x')$ is the optimal action a.s., and $\tilde{T}_2 Q^* = T^* Q^*$. ∎

Fact 8.12 (Double Q-learning; van Hasselt, 2010, Thm. 1)**.** Double Q-learning *keeps two tables $Q^{\mathsf{A}}, Q^{\mathsf{B}}$ and, at each step, updates one of them, chosen at random, by [8.4] with the target $R_k + \gamma Q^{\mathsf{B}}(X_{k+1}, \hat{a})$, $\hat{a} \in \arg\max_a Q^{\mathsf{A}}(X_{k+1}, a)$ (roles exchanged for Q^{B}). Under the conditions of [8.5] and bounded reward variance, both tables converge to Q^* almost surely. In actor–critic methods with function approximation, Fujimoto et al. (2018) use the minimum of two critics as target (clipped double Q), which by [8.8(a)] in reverse trades overestimation for a bias towards underestimation.*

8.4 The Running Examples

Example 8.13 (Exploration in the goal CMP)**.** In the goal CMP [3.8] the goal g is absorbing. Along a single trajectory under any policy, the pairs at 1 and 2 are visited only finitely often once g has been reached, so the conditions of [8.5] fail and Q-learning does not converge at those pairs. With *restarts*, i.e. the behaviour returns to state 1 whenever g is reached, the uniform behaviour policy generates an irreducible chain on $\mathcal{Z}$, all pairs are visited infinitely often [1.8], and with $\alpha_k(z) = \mathbf{1}\{Z_k = z\}/N_k(z)$ Q-learning converges to the values of [4.11]. Restarts are a property of the data stream, not of the update: each update at z uses a transition drawn from $P(\cdot|z)$, so the limit is still Q^* of the original CMP, provided the update at g is fed the self-transition $g \to g$ and not the restart.

Assumed without proof in this chapter

[8.6] convergence of SARSA under GLIE; [8.12] convergence of double Q-learning.

9

Approximate Dynamic Programming

When $\mathcal{X}$ is large, values are represented by a parametric or nonparametric class, and each Bellman update is followed by a projection or a regression onto the class. Two questions arise: does the projected iteration converge, and how do the per-step approximation errors propagate to the quality of the final policy? For on-policy linear temporal-difference learning the projected Bellman operator is a contraction in the norm of the stationary law (Tsitsiklis and Van Roy, 1997); off-policy it need not be. For fitted value iteration the errors propagate with the factor $\frac{2\gamma}{(1-\gamma)^2}$, and when errors are measured under a data distribution the price of the change of distribution is a *concentrability coefficient* (Munos, 2003; Munos and Szepesvári, 2008). $\mathcal{X}$ is countable and $\mathcal{A}$ finite.

9.1 Linear Temporal-Difference Learning

Definition 9.1 (Linear architecture and weighted norm)**.** Let $\mathcal{X}$ be finite, π stationary with P_π irreducible and invariant law μ [1.14], and $\varphi : \mathcal{X} \to \mathbb{R}^d$ a *feature map* whose $|\mathcal{X}| \times d$ matrix Φ (rows $\varphi(x)^\top$) has rank d. Let $\|f\|_\mu^2 := \sum_x \mu(x) f(x)^2$ with inner product $\langle f, f' \rangle_\mu$, and let Π_μ be the $\|\cdot\|_\mu$-orthogonal projection onto $\{\Phi\vartheta : \vartheta \in \mathbb{R}^d\}$.

Theorem 9.2 (The projected Bellman equation; Tsitsiklis and Van Roy, 1997)**.** *In the setting of [9.1]:*

(a) $\|P_\pi f\|_\mu \le \|f\|_\mu$ *for all* f*;*

(b) $\Pi_\mu T^\pi$ *is a* γ*-contraction in* $\|\cdot\|_\mu$*, with a unique fixed point* $\Phi\vartheta^\star$*;*

(c)

$$\|V^\pi - \Phi\vartheta^\star\|_\mu \le \frac{1}{\sqrt{1-\gamma^2}} \|V^\pi - \Pi_\mu V^\pi\|_\mu;$$

(d) $\vartheta^\star$ *solves* $A\vartheta^\star = b$ *with* $A := \Phi^\top D_\mu (I - \gamma P_\pi)\Phi$, $b := \Phi^\top D_\mu r_\pi$, $D_\mu := \operatorname{diag}(\mu)$, *and*

$$\vartheta^\top A \vartheta \ge (1-\gamma)\|\Phi\vartheta\|_\mu^2 > 0$$

for $\vartheta \neq 0$.

Proof.

(a) By Jensen, $(P_\pi f(x))^2 \le P_\pi f^2(x)$, and $\mu P_\pi = \mu$: $\|P_\pi f\|_\mu^2 \le \mu P_\pi f^2 = \|f\|_\mu^2$.

(b) Projections are $\|\cdot\|_\mu$-nonexpansive, and

$$\|T^\pi f - T^\pi f'\|_\mu = \gamma \|P_\pi (f - f')\|_\mu \le \gamma \|f - f'\|_\mu$$

by (a). Apply [0.8] on the finite-dimensional range of Π_μ.

(c)

$$V^\pi - \Pi_\mu V^\pi \perp \Pi_\mu V^\pi - \Phi\vartheta^\star,$$

so

$$\begin{aligned}\|V^\pi - \Phi\vartheta^\star\|_\mu^2 &= \|V^\pi - \Pi_\mu V^\pi\|_\mu^2 + \|\Pi_\mu T^\pi V^\pi - \Pi_\mu T^\pi \Phi\vartheta^\star\|_\mu^2 \\ &\le \|V^\pi - \Pi_\mu V^\pi\|_\mu^2 + \gamma^2 \|V^\pi - \Phi\vartheta^\star\|_\mu^2.\end{aligned}$$

(d) $\Phi\vartheta = \Pi_\mu T^\pi \Phi\vartheta$ iff $T^\pi \Phi\vartheta - \Phi\vartheta \perp \Phi e_j$ in $\langle\cdot,\cdot\rangle_\mu$ for all j, i.e.

$$\Phi^\top D_\mu (r_\pi + \gamma P_\pi \Phi\vartheta - \Phi\vartheta) = 0.$$

By Cauchy–Schwarz and (a),

$$\vartheta^\top A \vartheta = \|\Phi\vartheta\|_\mu^2 - \gamma \langle \Phi\vartheta, P_\pi \Phi\vartheta\rangle_\mu \ge (1-\gamma)\|\Phi\vartheta\|_\mu^2,$$

and Φ has rank d.

■

Theorem 9.3 (Convergence of linear TD(0))**.** *In the setting of [9.1], let (Y_k, Y_k') be i.i.d. pairs with $Y_k \sim \mu$ and $Y_k' \sim P_\pi(\cdot|Y_k)$, rewards R_k with $\mathbb{E}[R_k \mid Y_k] = r_\pi(Y_k)$, $R_k \in [0,1]$, and deterministic Robbins–Monro step sizes. Then*

$$\vartheta_{k+1} = \vartheta_k + \alpha_k \big(R_k + \gamma \varphi(Y_k')^\top \vartheta_k - \varphi(Y_k)^\top \vartheta_k\big)\varphi(Y_k)$$

converges almost surely to $\vartheta^\star$ of [9.2].

(Fact, Tsitsiklis and Van Roy, 1997, Thm. 1.) The same holds when $(Y_k, Y_k') = (X_k, X_{k+1})$ is a single trajectory of the aperiodic irreducible chain P_π, and for TD(λ).

Proof. The mean field is $h(\vartheta) = b - A\vartheta$ [9.2(d)], so

$$\langle \vartheta - \vartheta^\star, h(\vartheta)\rangle = -(\vartheta - \vartheta^\star)^\top A (\vartheta - \vartheta^\star) \le -c\|\vartheta - \vartheta^\star\|^2$$

with $c := (1-\gamma)\chi_{\min}(\Phi^\top D_\mu \Phi) > 0$. The noise ζ_{k+1} is the update minus its mean; with bounded features its conditional second moment is at most $K(1 + \|\vartheta_k - \vartheta^\star\|^2)$, and

$$\|h(\vartheta)\|^2 \le K(1 + \|\vartheta - \vartheta^\star\|^2).$$

Apply [7.3]. ■

Example 9.4 (Off-policy divergence)**.** Let $\mathcal{X} = \{1, 2\}$, one feature $\varphi(1) = 1$, $\varphi(2) = 2$, $P_\pi(2|1) = 1$, $r \equiv 0$, so $V^\pi = 0$. If the TD(0) update of [9.3] is applied only at state 1 (an off-policy sampling law $\beta = \delta_1$ instead of μ), then

$$\vartheta_{k+1} = \vartheta_k + \alpha_k(2\gamma - 1)\vartheta_k,$$

which diverges for $\gamma > \frac{1}{2}$ and $\vartheta_0 \neq 0$ whenever $\sum_k \alpha_k = \infty$. The projected operator $\Pi_\beta T^\pi$ is not a contraction in $\|\cdot\|_\beta$ because [9.2(a)] fails for $\beta \neq \mu$. Function approximation, bootstrapping and off-policy sampling together can make temporal-difference learning diverge.

9.2 Error Propagation in Approximate Value Iteration

Definition 9.5 (Approximate value iteration and concentrability)**.**

(a) An *approximate value iteration* is a sequence $V_{k+1} = T^\star V_k - \varepsilon_k$ in $B_b(\mathcal{X})$, $k \geq 0$, with arbitrary errors ε_k; π_k is greedy for V_k, $T^{\pi_k} V_k = T^\star V_k$. Fitted value iteration (Munos and Szepesvári, 2008) and fitted Q-iteration (Ernst et al., 2005) produce V_{k+1} by regressing sampled targets $T^\star V_k$ onto a function class, so ε_k is the regression error.

(b) For $\nu, \beta \in \mathcal{P}(\mathcal{X})$, the *m-step concentrability* is

$$c(m) := \sup_{\pi_1, \dots, \pi_m} \sup_x \frac{(\nu P_{\pi_1} \cdots P_{\pi_m})(x)}{\beta(x)}$$

over stationary deterministic policies ($\frac{a}{0} := \infty$ for $a > 0$), and the *discounted-average concentrability* is

$$C_{\nu,\beta} := (1-\gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c(m)$$

(Munos and Szepesvári, 2008, Assumption A2).

Theorem 9.6 (Error propagation; Munos and Szepesvári, 2008)**.** *In [9.5], for every $K \geq 1$:*

(a) (Pointwise.) With

$$P^k_{\mathrm{mix}} := (P_{\pi^\star})^{K-k} + P_{\pi_K} P_{\pi_{K-1}} \cdots P_{\pi_{k+1}},$$

$$V^\star - V^{\pi_K} \leq (I - \gamma P_{\pi_K})^{-1} \Big[\sum_{k<K} \gamma^{K-k} P^k_{\mathrm{mix}} |\varepsilon_k| + \gamma^{K+1} \big((P_{\pi^\star})^{K+1} + P_{\pi_K} \cdots P_{\pi_0} \big) |V^\star - V_0| \Big].$$

(b) (Sup norm; Bertsekas and Tsitsiklis, 1996.) If $\|\varepsilon_k\|_\infty \le \epsilon$ for all k, then

$$\|V^\star - V^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}\epsilon + \frac{2\gamma^{K+1}}{1-\gamma}\|V^\star - V_0\|_\infty.$$

(c) (Concentrability.) If $\langle \beta, |\varepsilon_k| \rangle \le \epsilon$ for all k, then

$$\langle \nu, V^\star - V^{\pi_K} \rangle \le \frac{2\gamma\, C_{\nu,\beta}}{(1-\gamma)^2}\epsilon + \frac{2\gamma^{K+1}}{1-\gamma}\|V^\star - V_0\|_\infty.$$

Proof.

(a) Since $T^\star V_k \ge T^{\pi^\star} V_k$,

$$V^\star - V_{k+1} = T^{\pi^\star} V^\star - T^\star V_k + \varepsilon_k \le \gamma P_{\pi^\star}(V^\star - V_k) + \varepsilon_k,$$

and since $T^\star V^\star \ge T^{\pi_k} V^\star$,

$$V^\star - V_{k+1} \ge \gamma P_{\pi_k}(V^\star - V_k) + \varepsilon_k.$$

Iterating (kernels are monotone [0.4(c)]),

$$V^\star - V_K \ge \sum_{k<K} \gamma^{K-k-1} P_{\pi_{K-1}} \cdots P_{\pi_{k+1}} \varepsilon_k + \gamma^K P_{\pi_{K-1}} \cdots P_{\pi_0}(V^\star - V_0),$$
$$V^\star - V_K \le \sum_{k<K} \gamma^{K-k-1} (P_{\pi^\star})^{K-k-1} \varepsilon_k + \gamma^K (P_{\pi^\star})^K (V^\star - V_0).$$

From $T^{\pi_K} V_K = T^\star V_K \ge T^{\pi^\star} V_K$:

$$\begin{aligned} V^\star - V^{\pi_K} &= \left(T^{\pi^\star} V^\star - T^{\pi^\star} V_K\right) + \left(T^{\pi^\star} V_K - T^{\pi_K} V_K\right) \\ &\quad + \left(T^{\pi_K} V_K - T^{\pi_K} V^{\pi_K}\right) \\ &\le \gamma P_{\pi^\star}(V^\star - V_K) + \gamma P_{\pi_K}(V_K - V^\star + V^\star - V^{\pi_K}), \end{aligned}$$

i.e.

$$(I - \gamma P_{\pi_K})(V^\star - V^{\pi_K}) \le \gamma (P_{\pi^\star} - P_{\pi_K})(V^\star - V_K).$$

$(I - \gamma P_{\pi_K})^{-1} = \sum_m \gamma^m P_{\pi_K}^m$ is monotone [2.7(a)]. Insert the upper bound for the $P_{\pi^\star}$ term and the lower bound for the $-P_{\pi_K}$ term, and bound by absolute values.

(b) Every kernel product has sup-norm 1, $\|(I - \gamma P_{\pi_K})^{-1}\| = \frac{1}{1-\gamma}$ [2.7(a)], and $\sum_{k<K} 2\gamma^{K-k} \le \frac{2\gamma}{1-\gamma}$.

(c) Integrate (a) against ν. A term of

$$\nu(I-\gamma P_{\pi_K})^{-1}\gamma^{K-k}P_{\mathrm{mix}}^{k}|\varepsilon_k|$$

with $\gamma^m P_{\pi_K}^m$ from the resolvent is γ^{m+K-k} times ν pushed through $m+K-k$ kernels of stationary deterministic policies, applied to $|\varepsilon_k|$; by definition of c, this is at most $\gamma^{m+K-k}c(m+K-k)\langle\beta,|\varepsilon_k|\rangle$. Summing over $m\geq 0$ and $k<K$, with $j:=K-k\geq 1$ and $n:=m+j$,

$$\sum_{j\geq 1}\sum_{m\geq 0}2\gamma^{m+j}c(m+j)\epsilon\leq 2\epsilon\sum_{n\geq 1}n\gamma^{n}c(n)=\frac{2\gamma C_{\nu,\beta}}{(1-\gamma)^2}\epsilon.$$

The initial-error term is bounded in sup norm as in (b).

■

Intuition. An error made at iteration k is felt by the final policy after it has been transported by the dynamics of both the optimal and the learned policies. In sup norm nothing is lost in transport. When errors are controlled only on average under a data law β, transport to the evaluation law ν can magnify them, and $c(m)$ measures by how much after m steps. Offline reinforcement learning is the study of when $C_{\nu,\beta}$ is finite and what to do when it is not.

Fact 9.7 (Approximate policy iteration; Bertsekas and Tsitsiklis, 1996, Munos, 2003)**.** *If policy iteration [4.7] is run with approximate evaluations $\|V_k-V^{\pi_k}\|_\infty\leq\epsilon$ and π_{k+1} greedy for V_k, then*

$$\limsup_k\|V^\star-V^{\pi_k}\|_\infty\leq\frac{2\gamma\epsilon}{(1-\gamma)^2};$$

Munos (2003) gives the analogous L_p bounds with concentrability coefficients.

Fact 9.8 (Finite-sample fitted value iteration; Munos and Szepesvári, 2008, Thm. 2)**.** *If each regression in [9.5(a)] uses N states drawn from β and M next-state samples per action, over a function class $\mathcal{V}$ of functions bounded by $\frac{1}{1-\gamma}$, then for any $\varepsilon,\delta>0$ there are K (logarithmic in $1/\varepsilon$) and N,M (polynomial in $1/\varepsilon$, $\log(1/\delta)$, $\frac{1}{1-\gamma}$, $\log|\mathcal{A}|$ and the logarithm of a covering number of $\mathcal{V}$) such that with probability at least $1-\delta$,*

$$\|V^\star-V^{\pi_K}\|_{p,\nu}\leq\frac{2\gamma}{(1-\gamma)^2}C_{\nu,\beta}^{1/p}\,d_{p,\beta}(T^\star\mathcal{V},\mathcal{V})+\varepsilon,$$

where

$$d_{p,\beta}(T^\star\mathcal{V},\mathcal{V}):=\sup_{V\in\mathcal{V}}\inf_{V'\in\mathcal{V}}\|T^\star V-V'\|_{p,\beta}$$

is the inherent Bellman error *of the class.*

9.3 Target Networks and Polyak Averaging

Definition 9.9 (Target-based linear evaluation)**.** Let $\varphi : \mathcal{Z} \to \mathbb{R}^d$, $\beta \in \mathcal{P}(\mathcal{Z})$ a sampling law, π stationary, and

$$\mathsf{C}_0 := \mathbb{E}_{Z\sim\beta}\varphi(Z)\varphi(Z)^\top,$$
$$\mathsf{C}_1 := \mathbb{E}_{Z\sim\beta,\,Z'\sim(P\otimes\pi)(Z,\cdot)}\varphi(Z)\varphi(Z')^\top,$$
$$\mathsf{b} := \mathbb{E}_{\beta}\varphi(Z)r(Z)$$

with C_0 invertible. Given *target parameters* $\bar\vartheta$, the regression of the bootstrapped target $r(Z) + \gamma\varphi(Z')^\top\bar\vartheta$ on $\varphi(Z)$ has the solution

$$\mathsf{G}(\bar\vartheta) := \mathsf{C}_0^{-1}(\mathsf{b} + \gamma\mathsf{C}_1\bar\vartheta) = \mathsf{C}_0^{-1}\mathsf{b} + \mathsf{M}\bar\vartheta, \qquad \mathsf{M} := \gamma\mathsf{C}_0^{-1}\mathsf{C}_1.$$

Fitted evaluation iterates $\bar\vartheta_{l+1} = \mathsf{G}(\bar\vartheta_l)$ (a hard target update after each full fit); *Polyak averaging* iterates

$$\bar\vartheta_{l+1} = (1-\upsilon)\bar\vartheta_l + \upsilon\mathsf{G}(\bar\vartheta_l),$$

$\upsilon \in (0, 1]$ (Lillicrap et al., 2016). The fixed points of G are exactly the solutions of the TD equation $\mathsf{A}\vartheta = \mathsf{b}$, $\mathsf{A} := \mathsf{C}_0 - \gamma\mathsf{C}_1 = \mathsf{C}_0(I - \mathsf{M})$ [9.2(d)].

Theorem 9.10 (Fitted evaluation; Fellows et al., 2023, Thm. 1)**.** *The iteration $\bar\vartheta_{l+1} = \mathsf{G}(\bar\vartheta_l)$ converges for every start iff $\varrho(\mathsf{M}) < 1$; its limit is then the unique TD fixed point $\vartheta^\star = \mathsf{A}^{-1}\mathsf{b}$, and*

$$\|\bar\vartheta_l - \vartheta^\star\| \le c\,\varrho'^{\,l}$$

for every $\varrho' \in (\varrho(\mathsf{M}), 1)$.

Proof.

$$\bar\vartheta_l - \vartheta^\star = \mathsf{M}^l(\bar\vartheta_0 - \vartheta^\star)$$

whenever $\vartheta^\star$ is a fixed point; $\mathsf{M}^l \to 0$ iff $\varrho(\mathsf{M}) < 1$ [0.9(c)], and then $I - \mathsf{M}$ is invertible [0.9(a)], so the fixed point exists and is unique. If $\varrho(\mathsf{M}) \ge 1$, an eigenvector direction does not decay. The rate follows from $\|\mathsf{M}^l\|^{1/l} \to \varrho(\mathsf{M})$. ∎

Theorem 9.11 (Polyak averaging stabilises)**.** *Let* $\mathrm{spec}(\mathsf{M})$ *be the set of eigenvalues of* M*.*

(a) The Polyak iteration converges for every start iff

$$\max_{\chi\in\mathrm{spec}(\mathsf{M})} |1 + \upsilon(\chi - 1)| < 1.$$

(b) There is $\upsilon_0 > 0$ such that it converges for every $\upsilon \in (0, \upsilon_0)$ iff $\operatorname{Re}\chi < 1$ for every $\chi \in \operatorname{spec}(\mathsf{M})$; then

$$\upsilon_0 = \min_{\chi} \frac{2(1 - \operatorname{Re}\chi)}{|1 - \chi|^2}.$$

(c) (Lyapunov.) If $\operatorname{Re}\chi < 1$ for all $\chi \in \operatorname{spec}(\mathsf{M})$, there is a positive definite L such that

$$\mathscr{V}(\bar{\vartheta}) := (\bar{\vartheta} - \vartheta^\star)^\top \mathsf{L} (\bar{\vartheta} - \vartheta^\star)$$

strictly decreases along the Polyak iteration for all small υ and $\bar{\vartheta} \neq \vartheta^\star$.

In particular, Polyak averaging converges whenever fitted evaluation does ($|\chi| < 1$ implies $\operatorname{Re}\chi < 1$), and also in cases where it does not.

Proof.

(a) The iteration is affine with linear part $(1 - \upsilon)I + \upsilon\mathsf{M}$, whose eigenvalues are $1 + \upsilon(\chi - 1)$; apply the argument of [9.10].

(b)
$$|1 + \upsilon(\chi - 1)|^2 = 1 - 2\upsilon(1 - \operatorname{Re}\chi) + \upsilon^2 |1 - \chi|^2,$$
which is < 1 iff
$$0 < \upsilon < \frac{2(1 - \operatorname{Re}\chi)}{|1 - \chi|^2};$$
such υ exist iff $\operatorname{Re}\chi < 1$.

(c) $\mathsf{M} - I$ has all eigenvalues with negative real part, so the Lyapunov equation $(\mathsf{M} - I)^\top \mathsf{L} + \mathsf{L}(\mathsf{M} - I) = -I$ has a positive definite solution
$$\mathsf{L} = \int_0^\infty e^{s(\mathsf{M}-I)^\top} e^{s(\mathsf{M}-I)}\, ds.$$
With $e := \bar{\vartheta} - \vartheta^\star$, the next error is $(I + \upsilon(\mathsf{M} - I))e$ and
$$\begin{aligned} \mathscr{V}_{\text{next}} - \mathscr{V} &= -\upsilon\|e\|^2 + \upsilon^2 e^\top (\mathsf{M} - I)^\top \mathsf{L} (\mathsf{M} - I) e \\ &< 0 \quad \text{for } 0 \\ &< \upsilon \\ &< 1/\|(\mathsf{M} - I)^\top \mathsf{L} (\mathsf{M} - I)\|. \end{aligned}$$

∎

Example 9.12 (Polyak converges where TD and hard targets diverge)**.** Take $d = 2$, $\mathsf{C}_0 = \operatorname{diag}(8, 1)$ and
$$\mathsf{M} = \left(\begin{smallmatrix} 1.5 & 1.5 \\ -1 & 0 \end{smallmatrix}\right)$$

(an abstract instance of [9.9]). Then $\operatorname{spec}(\mathsf{M}) = \{0.75 \pm 0.968i\}$, with $|\chi| \approx 1.22 > 1$ and $\operatorname{Re}\chi = 0.75 < 1$, so fitted evaluation diverges [9.10] while Polyak averaging converges exactly for $\upsilon < \upsilon_0 = \frac{2 \cdot 0.25}{0.25^2 + 0.968^2} = 0.5$ [9.11(b)]. The eigenvalues of $\mathsf{A} = \mathsf{C}_0(I - \mathsf{M})$ are $-1.5 \pm 2.40i$; expected semi-gradient TD,

$$\vartheta_{k+1} = \vartheta_k + \alpha(\mathsf{b} - \mathsf{A}\vartheta_k),$$

has linear part $I - \alpha\mathsf{A}$ with eigenvalues of modulus $|1 + \alpha(1.5 \mp 2.40i)| > 1$ for every $\alpha > 0$, so it diverges. For the off-policy example [9.4], $\mathsf{M} = 2\gamma$ is a real scalar and every scheme needs $\gamma < \frac{1}{2}$: target networks do not remove the deadly triad in general.

Remark 9.13 (Regularisation towards the target)**.** Adding $\eta\|\vartheta - \bar{\vartheta}\|^2$ to the regression gives

$$\mathsf{G}_\eta(\bar{\vartheta}) = (\mathsf{C}_0 + \eta I)^{-1}(\mathsf{b} + (\gamma\mathsf{C}_1 + \eta I)\bar{\vartheta}),$$

whose fixed points are again the TD fixed points; its linear part $\mathsf{M}_\eta = I - (\mathsf{C}_0 + \eta I)^{-1}\mathsf{A}$ interpolates between hard targets ($\eta = 0$) and small-step semi-gradient TD ($\eta \to \infty$, where stability requires $\operatorname{Re}\operatorname{spec}(\mathsf{A}) > 0$) (Fellows et al., 2023). Ridge regularisation without the target term changes the fixed point but can make the iteration a contraction under off-policy sampling (Zhang et al., 2021).

9.4 The Running Examples

Example 9.14 (Concentrability in the goal CMP)**.** In the goal CMP [3.8] let $\nu = \delta_1$. The policy π_{f} moves ν to δ_2 in one step and to δ_g in two, the policy π_{b} keeps it at δ_1. Hence $c(m) < \infty$ for all m requires $\beta(x) > 0$ for $x \in \{1, 2, g\}$, and for the uniform data law $\beta \equiv \frac{1}{3}$, $c(m) = 3$ for all m and $C_{\nu,\beta} = 3$. If the data never contain state 2 ($\beta(2) = 0$), then $c(1) = \infty$: no error bound under β can control the loss at 1, since the first action from 1 leads to the unobserved state 2.

Assumed without proof in this chapter

[9.3] the Markov-sampling and TD(λ) case of linear TD; [9.7] approximate policy iteration; [9.8] finite-sample fitted value iteration; the solution of the Lyapunov equation by the integral of matrix exponentials in [9.11(c)].

10

Policy-Gradient Theory

Value-based methods improve a policy through its values. Policy-gradient methods parametrise the policy and ascend the value directly (Williams, 1992; Sutton et al., 1999a). The objective $\vartheta \mapsto V^{\pi_\vartheta}(\nu)$ is not concave, yet for natural parametrisations its stationary points are global optima, and the natural policy gradient (Kakade, 2001) converges at a rate free of any distribution-mismatch coefficient (Agarwal et al., 2021). All three facts are consequences of the performance-difference lemma [4.8]. $\mathcal{X}$ and $\mathcal{A}$ are finite; $\nu \in \mathcal{P}(\mathcal{X})$ is a start law.

10.1 The Policy-Gradient Theorem

Theorem 10.1 (Policy-gradient theorem; Sutton et al., 1999a)**.** *Let $\vartheta \mapsto \pi_\vartheta(a|x)$ be differentiable on an open $\Theta \subseteq \mathbb{R}^d$ for every (x,a), and write $V^\vartheta := V^{\pi_\vartheta}$ etc. Then $\vartheta \mapsto V^\vartheta(\nu)$ is differentiable and*

$$\begin{aligned}\nabla_\vartheta V^\vartheta(\nu) &= \frac{1}{1-\gamma}\sum_{x,a} d_\nu^{\pi_\vartheta}(x,a)\, Q^\vartheta(x,a)\, \nabla_\vartheta \log \pi_\vartheta(a|x)\\ &= \frac{1}{1-\gamma}\sum_{x,a} d_\nu^{\pi_\vartheta}(x,a)\, \mathrm{A}^\vartheta(x,a)\, \nabla_\vartheta \log \pi_\vartheta(a|x),\end{aligned}$$

where terms with $\pi_\vartheta(a|x) = 0$ are read as

$$d_{\nu,\mathcal{X}}^{\pi_\vartheta}(x) Q^\vartheta(x,a) \nabla \pi_\vartheta(a|x).$$

*Equivalently (*REINFORCE*, Williams, 1992),*

$$\nabla V^\vartheta(\nu) = \mathbb{E}_\nu^{\pi_\vartheta} \sum_{t\geq 0} \gamma^t \big(Q^\vartheta(Z_t) - b(X_t)\big) \nabla \log \pi_\vartheta(A_t|X_t)$$

for every baseline $b : \mathcal{X} \to \mathbb{R}$.

Proof.

$$V^\vartheta = (I - \gamma P_{\pi_\vartheta})^{-1} r_{\pi_\vartheta}$$

[4.3] is a composition of differentiable maps (matrix inversion is smooth on invertible matrices), hence differentiable. Differentiate $V^{\vartheta}(x) = \sum_a \pi_{\vartheta}(a|x) Q^{\vartheta}(x,a)$ and

$$Q^{\vartheta}(x,a) = r(x,a) + \gamma \sum_{x'} P(x'|x,a) V^{\vartheta}(x')$$

[4.3]:

$$\nabla V^{\vartheta}(x) = u(x) + \gamma\, P_{\pi_{\vartheta}} \nabla V^{\vartheta}(x), \qquad u(x) := \sum_a \nabla \pi_{\vartheta}(a|x)\, Q^{\vartheta}(x,a).$$

So

$$\nabla V^{\vartheta} = (I - \gamma P_{\pi_{\vartheta}})^{-1} u = M_{\mathcal{X}}^{\pi_{\vartheta}} u$$

(componentwise in $\mathbb{R}^d$), and

$$\nu M_{\mathcal{X}}^{\pi_{\vartheta}} = \frac{1}{1-\gamma} d_{\nu,\mathcal{X}}^{\pi_{\vartheta}}$$

[6.10(c)]. Writing

$$\nabla \pi_{\vartheta} = \pi_{\vartheta} \nabla \log \pi_{\vartheta}$$

and

$$d_{\nu}^{\pi_{\vartheta}}(x,a) = d_{\nu,\mathcal{X}}^{\pi_{\vartheta}}(x) \pi_{\vartheta}(a|x)$$

[3.4(b)] gives the first formula. The second holds because $\sum_a \nabla \pi_{\vartheta}(a|x) = \nabla 1 = 0$, so any $b(x)$ may be subtracted from $Q^{\vartheta}(x,a)$; take $b = V^{\vartheta}$. The trajectory form is the first formula with

$$d_{\nu}^{\pi_{\vartheta}} = (1-\gamma) \sum_t \gamma^t \mathbb{P}_{\nu}^{\pi_{\vartheta}}\{Z_t \in \cdot\}$$

[3.5(c)]. ∎

Definition 10.2 (Parametrisations)**.**

(a) The *direct* parametrisation is $\vartheta = \pi$, ranging over the product of simplices $\Delta_{\mathcal{A}}^{\mathcal{X}}$; then

$$\frac{\partial V^{\pi}(\nu)}{\partial \pi(a|x)} = \frac{1}{1-\gamma} d_{\nu,\mathcal{X}}^{\pi}(x)\, Q^{\pi}(x,a),$$

by the proof of [10.1].

(b) The *softmax* parametrisation is

$$\pi_\vartheta(a|x) = \exp(\vartheta_{x,a}) / \sum_{a'} \exp(\vartheta_{x,a'}),$$

$\vartheta \in \mathbb{R}^{\mathscr{X}\times\mathscr{A}}$; then

$$\frac{\partial V^\vartheta(\nu)}{\partial \vartheta_{x,a}} = \frac{1}{1-\gamma} d_\nu^{\pi_\vartheta}(x,a)\, \mathbb{A}^\vartheta(x,a).$$

(c) The *distribution-mismatch coefficient* of π relative to $\beta \in \mathscr{P}(\mathscr{X})$ is $\|d_{\nu,\mathscr{X}}^{\pi^*} / d_{\beta,\mathscr{X}}^{\pi}\|_\infty$.

10.2 Gradient Domination

Theorem 10.3 (Gradient domination; Agarwal et al., 2021)**.** *For the direct parametrisation, every π and every $\beta \in \mathscr{P}(\mathscr{X})$,*

$$V^*(\nu) - V^\pi(\nu) \le \left\| \frac{d_{\nu,\mathscr{X}}^{\pi^*}}{d_{\beta,\mathscr{X}}^{\pi}} \right\|_\infty \max_{\tilde\pi \in \Delta_{\mathscr{A}}^{\mathscr{X}}} \langle \tilde\pi - \pi, \nabla_\pi V^\pi(\beta) \rangle.$$

In particular, if $d_{\beta,\mathscr{X}}^{\pi} > 0$ everywhere, every first-order stationary point of $\pi \mapsto V^\pi(\beta)$ over the product of simplices is optimal.

Proof. By the performance-difference lemma [4.8] with $\pi' = \pi^*$ and $\nu = \nu$, and since $\sum_a \pi(a|x)\mathbb{A}^\pi(x,a) = 0$ forces $\max_a \mathbb{A}^\pi(x,a) \ge 0$,

$$\begin{aligned} V^*(\nu) - V^\pi(\nu) &= \frac{1}{1-\gamma} \sum_x d_{\nu,\mathscr{X}}^{\pi^*}(x) \sum_a \pi^*(a|x) \mathbb{A}^\pi(x,a) \\ &\le \frac{1}{1-\gamma} \left\| \frac{d_{\nu,\mathscr{X}}^{\pi^*}}{d_{\beta,\mathscr{X}}^{\pi}} \right\|_\infty \sum_x d_{\beta,\mathscr{X}}^{\pi}(x) \max_a \mathbb{A}^\pi(x,a). \end{aligned}$$

By [10.2(a)],

$$\begin{aligned} &\max_{\tilde\pi} \langle \tilde\pi - \pi, \nabla_\pi V^\pi(\beta) \rangle \\ &\quad = \frac{1}{1-\gamma} \sum_x d_{\beta,\mathscr{X}}^{\pi}(x) \max_{\tilde\pi(\cdot|x)} \sum_a (\tilde\pi(a|x) - \pi(a|x)) Q^\pi(x,a) \\ &\quad = \frac{1}{1-\gamma} \sum_x d_{\beta,\mathscr{X}}^{\pi}(x) \max_a \mathbb{A}^\pi(x,a). \end{aligned}$$

■

Intuition. The gradient can be small at a bad policy only where the policy rarely goes. The mismatch coefficient measures how much the optimal policy's state distribution concentrates where the current policy does not. Exploration enters policy-gradient theory through the start law β, which must cover the states the optimal policy visits.

10.3 The Natural Policy Gradient

Fact 10.4 (Softmax NPG is soft policy iteration; Kakade, 2001)**.** *For the softmax parametrisation, the natural-policy-gradient step*

$$\vartheta \leftarrow \vartheta + \eta F_\nu(\vartheta)^\dagger \nabla V^\vartheta(\nu),$$

with F_ν the Fisher information of π_ϑ under $d_\nu^{\pi_\vartheta}$, changes the policy by

$$\pi_{t+1}(a|x) = \frac{\pi_t(a|x)\exp\left(\eta\, \mathbb{A}^{\pi_t}(x,a)/(1-\gamma)\right)}{\mathsf{Z}_t(x)},$$
$$\mathsf{Z}_t(x) := \sum_a \pi_t(a|x)\exp\left(\eta\, \mathbb{A}^{\pi_t}(x,a)/(1-\gamma)\right)$$

(Agarwal et al., 2026, Lemma 10.6). We take this multiplicative update as the definition of the NPG iteration.

Theorem 10.5 (Global convergence of NPG; Agarwal et al., 2021)**.** *Let π_0 be uniform and π_{t+1} given by [10.4]. Then:*

(a) (Monotone improvement.)

$$V^{\pi_{t+1}}(\beta) - V^{\pi_t}(\beta) \geq \frac{1-\gamma}{\eta}\sum_x \beta(x)\log \mathsf{Z}_t(x) \geq 0$$

for every $\beta \in \mathscr{P}(\mathscr{X})$;

(b) for every $T \geq 1$,

$$V^\star(\nu) - V^{\pi_T}(\nu) \leq \frac{\log|\mathscr{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.$$

No mismatch coefficient and no dependence on $|\mathscr{X}|$ appear.

Proof. Write $\mathbb{A}_t := \mathbb{A}^{\pi_t}$, $d^{t+1} := d_{\beta,\mathscr{X}}^{\pi_{t+1}}$. By [10.4],

$$\frac{\eta}{1-\gamma}\mathbb{A}_t(x,a) = \log\frac{\pi_{t+1}(a|x)\mathsf{Z}_t(x)}{\pi_t(a|x)}.$$

(a) By [4.8] with $\pi' = \pi_{t+1}$,

$$V^{\pi_{t+1}}(\beta) - V^{\pi_t}(\beta) = \frac{1}{1-\gamma} \sum_x d^{t+1}(x) \sum_a \pi_{t+1}(a|x) \mathbb{A}_t(x,a)$$
$$= \frac{1}{\eta} \sum_x d^{t+1}(x) \big[\mathrm{KL}(\pi_{t+1}(\cdot|x) \| \pi_t(\cdot|x)) + \log \mathsf{Z}_t(x) \big].$$

KL divergences are nonnegative; by Jensen,

$$\log \mathsf{Z}_t(x) \geq \frac{\eta}{1-\gamma} \sum_a \pi_t(a|x) \mathbb{A}_t(x,a) = 0;$$

and $d^{t+1} \geq (1-\gamma)\beta$ (the term $t = 0$ of [3.5(c)]).

(b) Let $d^\star := d^{\pi^\star}_{\nu,\mathscr{X}}$. By [4.8] with $\pi' = \pi^\star$,

$$V^\star(\nu) - V^{\pi_t}(\nu) = \frac{1}{\eta} \sum_x d^\star(x) \sum_a \pi^\star(a|x) \log \frac{\pi_{t+1}(a|x) \mathsf{Z}_t(x)}{\pi_t(a|x)}$$
$$= \frac{1}{\eta} \mathbb{E}_{d^\star} \big[\mathrm{KL}_t - \mathrm{KL}_{t+1} + \log \mathsf{Z}_t \big],$$

with $\mathrm{KL}_t(x) := \mathrm{KL}(\pi^\star(\cdot|x) \| \pi_t(\cdot|x))$. By (a) with $\beta = d^\star$,

$$\mathbb{E}_{d^\star} \log \mathsf{Z}_t \leq \frac{\eta}{1-\gamma} (V^{\pi_{t+1}}(d^\star) - V^{\pi_t}(d^\star)).$$

Summing over $t < T$ and telescoping, with $\mathrm{KL}_0 \leq \log |\mathscr{A}|$ and $0 \leq V \leq \frac{1}{1-\gamma}$,

$$\sum_{t<T} \big(V^\star(\nu) - V^{\pi_t}(\nu) \big) \leq \frac{\log |\mathscr{A}|}{\eta} + \frac{1}{(1-\gamma)^2}.$$

By (a) with $\beta = \nu$, $V^{\pi_t}(\nu)$ increases in t, so the left side is at least $T(V^\star(\nu) - V^{\pi_T}(\nu))$.

■

Intuition. NPG is mirror ascent with the KL divergence, state by state, on the advantages of the current policy. The performance-difference lemma converts the per-state regret of mirror ascent into a value gap under the optimal policy's occupancy, and the optimal occupancy never has to be compared with the current one. As $\eta \to \infty$ the update becomes greedy, and NPG becomes policy iteration [4.7].

Fact 10.6 (Vanilla softmax policy gradient; Agarwal et al., 2021)**.**

(a) Gradient ascent on $V^{\vartheta}(\beta)$ with the softmax parametrisation, a start law $\beta > 0$ and step size $\eta \le (1-\gamma)^3/8$ converges to an optimal policy: $V^{\vartheta_t}(\nu) \to V^{\star}(\nu)$ (Agarwal et al., 2026, Thm. 10.3).

(b) Without further structure the convergence can be exponentially slow in the effective horizon: in a chain of length H with $\gamma = H/(H+1)$, the gradient at a suboptimal policy is exponentially small in H (Agarwal et al., 2026, Prop. 10.1).

10.4 The Running Examples

Example 10.7 (Policy gradient on the goal CMP)**.** In the goal CMP [3.8] with $\gamma = 0.9$, start at the uniform policy. Its advantages are positive for the forward action at both transient states: $\mathbb{A}^{\pi_u}(1, \mathsf{f}) \approx 0.58$ [4.11] and

$$\mathbb{A}^{\pi_u}(2, \mathsf{f}) = \frac{1}{2}\gamma(V^{\pi_u}(g) - V^{\pi_u}(1)) \approx 1.87.$$

The NPG update [10.4] multiplies the odds of f by

$$\exp(\eta(\mathbb{A}(x, \mathsf{f}) - \mathbb{A}(x, \mathsf{b}))/(1-\gamma))$$

at each state, so after one step with $\eta = 0.1$ the probability of f is ≈ 0.76 at 1 and ≈ 0.98 at 2; the iterates converge to π_{f} at the rate [10.5(b)] with $|\mathscr{A}| = 2$. The gradient is largest at state 2, the state closest to the reward: credit propagates backwards from the goal, as in value iteration [4.11].

Assumed without proof in this chapter

[10.4] the equivalence of softmax NPG with the multiplicative update; [10.6] convergence and slowness of vanilla softmax policy gradient.

Part III

The Statistical Theory of Reinforcement Learning

11

Lower Bounds and Bandits

The statistical theory of reinforcement learning asks how much data is necessary and sufficient. Upper bounds come from algorithms and concentration; lower bounds come from the impossibility of distinguishing two environments that differ where the learner has little data. This chapter builds the lower-bound toolbox on the trajectory measure: the Kullback–Leibler divergence between two trajectory measures decomposes into the expected per-step divergences along the trajectory, and the Bretagnolle–Huber inequality turns a small divergence into an unavoidable error. The multi-armed bandit, the one-state CMP, is the first application: UCB attains regret $O(\sqrt{kn \log n})$ (Auer et al., 2002), and no policy does better than $\Omega(\sqrt{kn})$ (Lattimore and Szepesvári, 2020). $\mathcal{X}$ and $\mathcal{A}$ are countable.

11.1 Divergences

Definition 11.1 (Kullback–Leibler divergence)**.** For $\mu, \mu' \in \mathcal{P}(E)$ on a countable E, $\mathrm{KL}(\mu\|\mu') := \sum_x \mu(x) \log \frac{\mu(x)}{\mu'(x)}$ if $\mu(x) > 0 \implies \mu'(x) > 0$, and $+\infty$ otherwise ($0 \log 0 := 0$). For Bernoulli laws, $\mathrm{kl}(p, q) := \mathrm{KL}(\mathrm{Ber}(p)\|\mathrm{Ber}(q))$. For kernels K, K', $\mathrm{KL}(K\|K')(x) := \mathrm{KL}(K(x,\cdot)\|K'(x,\cdot))$.

Proposition 11.2 (Properties of KL)**.**

(a) $\mathrm{KL}(\mu\|\mu') \geq 0$*, with equality iff* $\mu = \mu'$*.*

(b) (Chain rule.)

$$\mathrm{KL}(\mu \otimes K \,\|\, \mu' \otimes K') = \mathrm{KL}(\mu\|\mu') + \sum_x \mu(x)\, \mathrm{KL}(K\|K')(x).$$

(c) (Data processing.) $\mathrm{KL}(f_\# \mu \| f_\# \mu') \leq \mathrm{KL}(\mu\|\mu')$ *for every map* f *between countable spaces.*

(d) (χ^2 bound.)

$$\mathrm{KL}(\mu\|\mu') \leq \sum_x \frac{(\mu(x) - \mu'(x))^2}{\mu'(x)};$$

in particular $\mathrm{kl}(p, q) \leq \frac{(p-q)^2}{q(1-q)}$.

(e) *(Pinsker; Fact.)* $\|\mu - \mu'\|_{\mathrm{TV}} \le \sqrt{\mathrm{KL}(\mu\|\mu')/2}$ *(Lattimore and Szepesvári, 2020, Eq. (14.12)).*

Proof.

(a) By $\log u \le u - 1$,

$$-\mathrm{KL}(\mu\|\mu') = \sum_{\mu(x)>0} \mu(x) \log \frac{\mu'(x)}{\mu(x)} \le \sum_{\mu(x)>0} (\mu'(x) - \mu(x)) \le 0,$$

with equality iff $\mu'(x) = \mu(x)$ wherever $\mu(x) > 0$, hence everywhere.

(b) Expand $\log \frac{\mu(x)K(x,y)}{\mu'(x)K'(x,y)}$ and sum (the infinite case is checked separately: the left side is infinite iff one of the right terms is).

(c) By the log-sum inequality, which is (a) applied to normalised restrictions: for y in the range of f,

$$\sum_{x \in f^{-1}(y)} \mu(x) \log \frac{\mu(x)}{\mu'(x)} \ge f_{\#}\mu(y) \log \frac{f_{\#}\mu(y)}{f_{\#}\mu'(y)}.$$

Sum over y.

(d) $\log u \le u - 1$ with $u = \mu(x)/\mu'(x)$:

$$\mathrm{KL} \le \sum_x \mu(x)\Big(\frac{\mu(x)}{\mu'(x)} - 1\Big) = \sum_x \frac{(\mu(x) - \mu'(x))^2}{\mu'(x)}.$$

■

Theorem 11.3 (Bretagnolle–Huber inequality; Bretagnolle and Huber, 1979)**.** *For $\mu, \mu' \in \mathcal{P}(E)$ on a countable E and every $B \subseteq E$,*

$$\mu(B) + \mu'(E \setminus B) \ge \sum_x \mu(x) \wedge \mu'(x) \ge \tfrac{1}{2} \exp\big(-\mathrm{KL}(\mu\|\mu')\big).$$

Proof. The first inequality is

$$\mu(B) + \mu'(B^c) \ge \sum_{x \in B} \mu \wedge \mu' + \sum_{x \notin B} \mu \wedge \mu'.$$

For the second, assume $\mathrm{KL} < \infty$. By Cauchy–Schwarz,

$$\Big(\sum_x \sqrt{\mu\mu'}\Big)^2 = \Big(\sum_x \sqrt{(\mu \wedge \mu')(\mu \vee \mu')}\Big)^2 \le \sum \mu \wedge \mu' \cdot \sum \mu \vee \mu' \le 2 \sum_x \mu \wedge \mu'.$$

By Jensen,

$$\sum_x \sqrt{\mu\mu'} = \sum_{\mu(x)>0} \mu(x) \exp\big(\tfrac{1}{2} \log \tfrac{\mu'(x)}{\mu(x)}\big) \ge \exp(-\tfrac{1}{2}\,\mathrm{KL}(\mu\|\mu')).$$

■

11.2 Divergence of Trajectory Measures

Lemma 11.4 (Divergence decomposition)**.** *Let P, P' be two transition kernels on $(\mathcal{X}, \mathcal{A})$, $\pi \in \Pi^{\mathrm{HR}}$, $\nu \in \mathcal{P}(\mathcal{X})$, and $\mathbb{P}, \mathbb{P}'$ the trajectory measures $\mathbb{P}_\nu^\pi$ under P and under P' [3.3]. For $n \in \mathbb{N}$, let $\mathbb{P}_n, \mathbb{P}'_n$ be the laws of $(Z_0, \dots, Z_{n-1}, X_n)$. Then*

$$\mathrm{KL}(\mathbb{P}_n \| \mathbb{P}'_n) = \sum_{t<n} \mathbb{E}\big[\, \mathrm{KL}(P\|P')(Z_t)\big] = \sum_{z \in \mathcal{Z}} \mathbb{E}[N_n(z)]\ \mathrm{KL}(P\|P')(z),$$

with $N_n(z) := \sum_{t<n} \mathbf{1}\{Z_t = z\}$ the number of visits to z before n, and $\mathbb{E}$ the expectation under $\mathbb{P}$.

Proof. By the product formula of [3.3], the law of the history is built by alternately applying π_t (identical under both measures) and P or P'. Apply the chain rule [11.2(b)] $2n$ times: the policy steps contribute $\mathrm{KL}(\pi_t \| \pi_t) = 0$, and the t-th transition step contributes $\mathbb{E}[\mathrm{KL}(P\|P')(Z_t)]$. Exchange the sums [0.2(b)]. ∎

Intuition. The learner can tell two environments apart only through the transitions it actually samples, and each visit to a pair z yields $\mathrm{KL}(P\|P')(z)$ nats of evidence. To hide a difference, an adversary places it at pairs that the learner visits rarely. Every lower bound in Part III is an instance of this sentence.

Corollary 11.5 (Two-point method)**.** *Let $\mu_0, \mu_1 \in \mathcal{P}(E)$ and let $\hat{j} : E \to \{0, 1\}$ be any test. Then*

$$\max_j \mu_j\{\hat{j} \neq j\} \geq \frac{1}{4} \exp(-\mathrm{KL}(\mu_0 \| \mu_1)).$$

Consequently, if $\vartheta_0, \vartheta_1 \in \mathbb{R}$ with $|\vartheta_0 - \vartheta_1| \geq 2\varepsilon$ and $\hat{\vartheta}$ is any estimator, then

$$\max_j \mu_j\{|\hat{\vartheta} - \vartheta_j| \geq \varepsilon\} \geq \frac{1}{4} \exp(-\mathrm{KL}(\mu_0 \| \mu_1)).$$

For n i.i.d. samples,

$$\mathrm{KL}(\mu_0^{\otimes n} \| \mu_1^{\otimes n}) = n\, \mathrm{KL}(\mu_0 \| \mu_1)$$

[11.2(b)].

11.3. with $B = \{\hat{j} = 1\}$; the maximum of two numbers is at least their mean. For estimators use the test

$$\hat{j} := \mathbf{1}\{|\hat{\vartheta} - \vartheta_1| < |\hat{\vartheta} - \vartheta_0|\},$$

which errs only if $|\hat{\vartheta} - \vartheta_j| \geq \varepsilon$. ∎

11.3 Multi-Armed Bandits

Definition 11.6 (Bandits and regret)**.** A *k-armed bandit* is a CMP with one state, actions $\mathcal{A} = \{1, \dots, k\}$, and reward laws $\mathsf{P}_a \in \mathscr{P}([0,1])$ with means m_a; at round t the learner chooses A_t by a history-dependent policy and observes $Y_t \sim \mathsf{P}_{A_t}$. With $m^\star := \max_a m_a$, gaps $\Delta_a := m^\star - m_a$ and $T_a(n) := \sum_{t<n} \mathbf{1}\{A_t = a\}$, the *regret* after n rounds is

$$\mathsf{R}_n := nm^\star - \mathbb{E}\sum_{t<n} Y_t = \sum_a \Delta_a \mathbb{E}T_a(n).$$

UCB with confidence δ plays each arm once and then

$$A_t \in \arg\max_a \left(\hat{m}_a(t) + \sqrt{2\log(1/\delta)/T_a(t)}\right),$$

with $\hat{m}_a(t)$ the empirical mean of arm a.

Theorem 11.7 (UCB; Auer et al., 2002, Lattimore and Szepesvári, 2020, Thm. 7.1–7.2)**.** *With $\delta = 1/n^2$, UCB satisfies*

$$\mathsf{R}_n \le \sum_{a:\Delta_a>0} \left(\frac{16\log n}{\Delta_a} + 3\Delta_a\right) \qquad \text{and} \qquad \mathsf{R}_n \le 8\sqrt{kn\log n} + 3\sum_a \Delta_a.$$

Proof. Realise the rewards by an independent array $(Y_{a,s})_{s\ge1}$, $Y_{a,s} \sim \mathsf{P}_a$, the s-th reward of arm a; $\hat{m}_{a,s}$ is the mean of the first s. Let arm 1 be optimal and a suboptimal. Put $u := \lceil 16\log n/\Delta_a^2 \rceil$, so $\sqrt{2\log(1/\delta)/u} \le \Delta_a/2$. Let

$$\mathsf{G} := \{m_1 < \hat{m}_{1,s} + \sqrt{2\log(1/\delta)/s}\ \forall s \le n\} \cap \{\hat{m}_{a,u} + \sqrt{2\log(1/\delta)/u} < m_1\}.$$

On G, $T_a(n) \le u$*:* otherwise at some round a is chosen with $T_a = u$, so its index is at least that of arm 1, which exceeds m_1, while its index is below m_1. By Hoeffding [0.12(a)] (one-sided, $[0,1]$ rewards), the first event fails with probability at most $\sum_{s\le n}\delta = 1/n$, and the second fails only if $\hat{m}_{a,u} - m_a \ge \Delta_a/2$, with probability at most $e^{-u\Delta_a^2/2} \le n^{-8}$. Hence

$$\mathbb{E}T_a(n) \le u + n\cdot\frac{2}{n} \le \frac{16\log n}{\Delta_a^2} + 3;$$

multiply by Δ_a and sum [11.6]. For the second bound, split arms at a threshold $\bar{\Delta}$:

$$\mathsf{R}_n \le n\bar{\Delta} + \sum_{\Delta_a \ge \bar{\Delta}} (16\log n/\Delta_a + 3\Delta_a) \le n\bar{\Delta} + 16k\log n/\bar{\Delta} + 3\sum_a \Delta_a,$$

and take $\bar{\Delta} = \sqrt{16k\log n/n}$. ∎

Theorem 11.8 (Minimax lower bound for bandits; Lattimore and Szepesvári, 2020, Thm. 15.2)**.** *Let $k \geq 2$ and $n \geq 3(k-1)$. For every policy there is a Bernoulli bandit with means in $[\frac{1}{2}, \frac{3}{4}]$ on which $\mathsf{R}_n \geq \frac{1}{110}\sqrt{(k-1)n}$.*

Proof. Let $\Delta := \sqrt{(k-1)/(22n)} \leq \frac{1}{8}$. Environment ν: $m_1 = \frac{1}{2} + \Delta$, $m_a = \frac{1}{2}$ for $a \neq 1$. Pick $i \neq 1$ with $\mathbb{E}_\nu T_i(n) \leq \frac{n}{k-1}$. Environment ν': as ν but $m_i = \frac{1}{2} + 2\Delta$. By [11.4] and [11.2(d)],

$$\mathrm{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) = \mathbb{E}_\nu T_i(n)\, \mathrm{kl}(\frac{1}{2}, \frac{1}{2} + 2\Delta) \leq \frac{n}{k-1} \cdot \frac{16\Delta^2}{1 - 16\Delta^2} \leq \frac{22n\Delta^2}{k-1} = 1.$$

In ν, every round with $A_t \neq 1$ costs at least Δ; in ν', every round with $A_t \neq i$ costs at least Δ. With $B := \{T_1(n) \leq n/2\}$, $\mathsf{R}_n(\nu) \geq \frac{n\Delta}{2}\mathbb{P}_\nu(B)$ and $\mathsf{R}_n(\nu') \geq \frac{n\Delta}{2}\mathbb{P}_{\nu'}(B^c)$. By [11.3],

$$\mathsf{R}_n(\nu) + \mathsf{R}_n(\nu') \geq \frac{n\Delta}{4} e^{-1},$$

so one of them is at least

$$\frac{n\Delta}{8e} = \frac{\sqrt{(k-1)n}}{8e\sqrt{22}} \geq \frac{\sqrt{(k-1)n}}{110}.$$

■

Fact 11.9 (Asymptotic lower bound; Lai and Robbins, 1985)**.** *Call a policy* consistent *on Bernoulli bandits if $\mathsf{R}_n = o(n^p)$ for every $p > 0$ and every Bernoulli bandit. Then every consistent policy satisfies*

$$\liminf_n \mathsf{R}_n / \log n \geq \sum_{a : \Delta_a > 0} \Delta_a / \mathrm{kl}(m_a, m^*)$$

(Lattimore and Szepesvári, 2020, Thm. 16.2). Since

$$\mathrm{kl}(m_a, m^*) \leq \Delta_a^2 / (m^*(1 - m^*))$$

[11.2(d)], the $\log n / \Delta_a$ dependence of [11.7] is unimprovable up to constants.

11.4 The Running Examples

Example 11.10 (Estimating a reach probability)**.** In the goal CMP [3.8], let the forward action at state 2 succeed only with an unknown probability p (and

stay at 2 otherwise). To estimate p to accuracy $\varepsilon \le \frac{1}{8}$ from n samples of that transition, any estimator errs with probability at least

$$\frac{1}{4}\exp(-n\,\mathrm{kl}(\frac{1}{2}, \frac{1}{2}+2\varepsilon)) \ge \frac{1}{4}\exp(-22n\varepsilon^2)$$

for one of $p \in \{\frac{1}{2}, \frac{1}{2}+2\varepsilon\}$ [11.5], [11.2(d)]. So $n = \Omega(\varepsilon^{-2}\log(1/\delta))$ samples per pair are necessary for confidence $1-\delta$. Part III asks how this per-pair cost multiplies along an effective horizon of $\frac{1}{1-\gamma}$ steps.

Assumed without proof in this chapter

[11.2(e)] Pinsker's inequality; [11.9] the Lai–Robbins lower bound.

12

Sample Complexity with a Generative Model

The simplest statistical question in reinforcement learning removes exploration: the learner may sample a next state from $P(\cdot|z)$ at any pair z it chooses. How many samples are needed to compute ε-accurate optimal values, or an ε-optimal policy? The answer,

$$\tilde{\Theta}\big(|\mathcal{X}||\mathcal{A}|/((1-\gamma)^3\varepsilon^2)\big),$$

was obtained by Azar et al. (2013) for values and by Sidford et al. (2018) for policies. It is smaller by a factor $|\mathcal{X}|$ than the number of samples needed to estimate the model in total variation, and smaller by a factor $\frac{1}{1-\gamma}$ than what Hoeffding's inequality and the simulation lemma give. This chapter proves the elementary bound, isolates the two mechanisms behind the sharp one (a fixed comparison function, and the total variance of the return), and states the sharp bounds. $\mathcal{X}$, $\mathcal{A}$ are finite and $\mathcal{Z} = \mathcal{X} \times \mathcal{A}$.

12.1 The Plug-in Model

Definition 12.1 (Generative model and empirical MDP)**.** A *generative model* returns, on query $z \in \mathcal{Z}$, a sample $Y \sim P(\cdot|z)$, independent of all earlier samples. With n queries at every pair, the *empirical kernel* is

$$\hat{P}(y|z) := \frac{1}{n}\#\{\text{samples at } z \text{ equal to } y\},$$

and hats denote the objects of the *empirical MDP* $(\mathcal{X}, \mathcal{A}, \hat{P}, r, \gamma)$: $\hat{Q}^\pi$, $\hat{Q}^\star$, $\hat{\pi}^\star := \pi_{\hat{Q}^\star}$. The total number of samples is $N = n|\mathcal{X}||\mathcal{A}|$.

Theorem 12.2 (Elementary bound)**.** *Let* $\delta \in (0,1)$ *and* $\epsilon_n := \sqrt{\log(2|\mathcal{Z}|/\delta)/(2n)}$. *With probability at least* $1-\delta$*:*

(a)

$$\|\hat{Q}^\star - Q^\star\|_\infty \le \frac{\gamma}{(1-\gamma)^2}\epsilon_n;$$

(b) for every stationary π *fixed before sampling,*

$$\|\hat{Q}^\pi - Q^\pi\|_\infty \le \frac{\gamma}{(1-\gamma)^2}\epsilon_n$$

(with $|\mathcal{Z}|$ *in* ϵ_n *unchanged);*

(c)

$$\|V^\star - V^{\hat{\pi}^\star}\|_\infty \le \frac{2\gamma}{(1-\gamma)^3}\epsilon_n.$$

Hence

$$n \ge \frac{\log(2|\mathcal{Z}|/\delta)}{2(1-\gamma)^4\varepsilon^2}$$

samples per pair suffice for ε-accurate optimal values, i.e.

$$N = O\Big(\frac{|\mathcal{X}||\mathcal{A}|\log(|\mathcal{Z}|/\delta)}{(1-\gamma)^4\varepsilon^2}\Big),$$

and $\frac{1}{(1-\gamma)^2}$ times more for an ε-optimal greedy policy by this route.

Proof. $V^\star$ is fixed before sampling and takes values in $[0, \frac{1}{1-\gamma}]$. By Hoeffding [0.12(a)] and the union bound [0.12(e)] over the $|\mathcal{Z}|$ pairs, with probability $\ge 1-\delta$, $|(\hat{P}-P)V^\star(z)| \le \frac{\epsilon_n}{1-\gamma}$ for all z; on this event:

(a) By the Bellman optimality equations [4.5] in both MDPs,

$$\hat{Q}^\star - Q^\star = \gamma\hat{P}(\hat{V}^\star - V^\star) + \gamma(\hat{P}-P)V^\star,$$

where $\hat{V}^\star := \max_a \hat{Q}^\star(\cdot, a)$. Since

$$\|\hat{V}^\star - V^\star\|_\infty \le \|\hat{Q}^\star - Q^\star\|_\infty$$

and $\hat{P}$ is a Markov kernel,

$$\|\hat{Q}^\star - Q^\star\|_\infty \le \gamma\|\hat{Q}^\star - Q^\star\|_\infty + \frac{\gamma\epsilon_n}{1-\gamma}.$$

(b) The same argument with T^π and the fixed function V^π in place of $V^\star$ (a separate Hoeffding event, again over $|\mathcal{Z}|$ pairs).

(c) The greedy-policy loss bound [4.6(b)] applied to $Q = \hat{Q}^\star$ and (a).

■

Intuition. The elementary bound needs no estimate of the model in total variation, which would cost $|\mathcal{X}|$ times more samples: it compares the two Bellman equations along the one fixed function $V^\star$. The remaining waste is the range of $V^\star$, $\frac{1}{1-\gamma}$, used in Hoeffding's inequality. Along a trajectory, however, the one-step variances of $V^\star$ cannot all be large at once.

12.2 The Total Variance of the Return

Theorem 12.3 (Total variance; Azar et al., 2013)**.** *Let π be stationary, $\Sigma^\pi(z) := \mathrm{Var}_z^\pi(G)$ the variance of the return law η_z^π [6.9], and*

$$v^\pi(z) := \mathrm{Var}_{Z' \sim (P \otimes \pi)(z,\cdot)} \left(Q^\pi(Z') \right)$$

the one-step variance of the next action value.

(a) (Variance Bellman equation.)

$$\Sigma^\pi = \gamma^2 v^\pi + \gamma^2 (P \otimes \pi) \Sigma^\pi,$$

i.e.

$$\Sigma^\pi = \gamma^2 \sum_t \gamma^{2t} (P \otimes \pi)^t v^\pi.$$

(b) $\|\Sigma^\pi\|_\infty \le \frac{1}{(1-\gamma)^2}$, hence

$$\| \sum_t \gamma^{2t} (P \otimes \pi)^t v^\pi \|_\infty \le \frac{1}{\gamma^2 (1-\gamma)^2}.$$

(c) (Total variance.) With $U_\gamma^\pi := \sum_t \gamma^t (P \otimes \pi)^t$,

$$\| U_\gamma^\pi \sqrt{v^\pi} \|_\infty \le \sqrt{\frac{2}{\gamma^2 (1-\gamma)^3}}.$$

A crude bound $\sqrt{v^\pi} \le \frac{1}{1-\gamma}$ would give $\frac{1}{(1-\gamma)^2}$ in (c).

Proof.

(a) Under $\mathbb{P}_z^\pi$, $G = r(z) + \gamma G \circ \theta_1$ [3.6(b)], so $\Sigma^\pi(z) = \gamma^2 \mathrm{Var}(G \circ \theta_1)$. By the Markov property [3.4(c)], conditionally on $\mathscr{F}_1$ the variable $G \circ \theta_1$ has the law $\eta_{Z_1}^\pi$, with mean $Q^\pi(Z_1)$ and variance $\Sigma^\pi(Z_1)$. The law of total variance gives

$$\mathrm{Var}(G \circ \theta_1) = \mathbb{E} \Sigma^\pi(Z_1) + \mathrm{Var}(Q^\pi(Z_1)) = (P \otimes \pi) \Sigma^\pi(z) + v^\pi(z).$$

The series form is [0.9(a)] with $\|\gamma^2 (P \otimes \pi)\| = \gamma^2 < 1$.

(b) $G \in [0, \frac{1}{1-\gamma}]$, so its variance is at most $\frac{1}{4(1-\gamma)^2} \le \frac{1}{(1-\gamma)^2}$. Then use (a).

(c) $(1-\gamma) U_\gamma^\pi(z, \cdot)$ is a probability [2.7(a)]; by Jensen for $\sqrt{\cdot}$,

$$U_\gamma^\pi \sqrt{v^\pi} \le \sqrt{\frac{1}{1-\gamma} U_\gamma^\pi v^\pi}.$$

With $w := \sum_t \gamma^{2t}(P \otimes \pi)^t v^\pi$, so that $v^\pi = w - \gamma^2(P \otimes \pi)w$, we have

$$U_\gamma^\pi v^\pi = U_\gamma^\pi\big((1-\gamma)I + \gamma(I - \gamma(P \otimes \pi))\big)w = (1-\gamma)U_\gamma^\pi w + \gamma w,$$

whose sup norm is at most $2\|w\|_\infty \le \frac{2}{\gamma^2(1-\gamma)^2}$ by (b). Hence

$$U_\gamma^\pi \sqrt{v^\pi} \le \sqrt{\frac{2}{\gamma^2(1-\gamma)^3}}.$$

■

Intuition. The one-step variances of the action value along a trajectory add up, after discounting, to at most the variance of the return, which is at most the squared range $\frac{1}{(1-\gamma)^2}$, not to the effective horizon times the squared range. Bernstein's inequality [0.12(b)] pays for the one-step variance; the total-variance theorem pays for their sum. The return law of Part IV enters the statistical theory here, through its second moment.

12.3 The Sharp Bounds

Fact 12.4 (Minimax upper bounds; Azar et al., 2013; Sidford et al., 2018)**.** *There is an absolute constant c such that, with n samples per pair and*

$$N = n|\mathcal{Z}| \ge \frac{c|\mathcal{X}||\mathcal{A}|\log(c|\mathcal{Z}|/\delta)}{(1-\gamma)^3\varepsilon^2},$$

with probability at least $1-\delta$:

(a) if $\varepsilon \le 1$, then $\|\hat{Q}^\star - Q^\star\|_\infty \le \varepsilon$;

(b) if $\varepsilon \le 1/\sqrt{1-\gamma}$, then $\|Q^{\hat{\pi}^\star} - Q^\star\|_\infty \le \varepsilon$

(Agarwal et al., 2026, Thm. 2.6). The proof of (a) combines Bernstein's inequality for $(\hat{P} - P)V^\star$, a comparison of empirical and true one-step variances, and the total-variance theorem [12.3]; Sidford et al. (2018) achieve the same sample size, with near-linear running time, by variance-reduced value iteration.

Fact 12.5 (Minimax lower bound; Azar et al., 2013)**.** *There are absolute constants $\varepsilon_0, \delta_0, c > 0$ and a class of MDPs such that every algorithm that outputs $\hat{Q}$ with $\|\hat{Q} - Q^\star\|_\infty \le \varepsilon$ with probability at least $1-\delta$ on every MDP of the class, for $\varepsilon < \varepsilon_0$ and $\delta < \delta_0$, needs at least*

$$\frac{c|\mathcal{X}||\mathcal{A}|\log(c/\delta)}{(1-\gamma)^3\varepsilon^2}$$

samples from the generative model (Agarwal et al., 2026, Thm. 2.7). The hard instances place, at each pair, a transition to a state with a self-loop of probability $p \approx \gamma$, whose value $\frac{1}{1-\gamma p}$ is a $\frac{1}{(1-\gamma)^2}$-sensitive function of p; the two-point method [11.5] then applies pair by pair.

Remark 12.6 (One reward versus all rewards)**.** The bound $\tilde{\Theta}(|\mathcal{X}||\mathcal{A}|/((1-\gamma)^3\varepsilon^2))$ concerns one reward fixed in advance: the argument of [12.2] and [12.4] compares the empirical and true Bellman equations along one fixed function $V^\star$. Estimating P to accuracy ε' in total variation at every pair costs $\Theta(|\mathcal{X}|/\varepsilon'^2)$ samples per pair, $|\mathcal{X}|$ times more, and by the simulation lemma [4.10] it yields accurate values for *every* reward at once. Chapter 16 shows that, up to the horizon factors, this extra $|\mathcal{X}|$ is unavoidable when the reward is revealed only after the data are collected (Jin et al., 2020a). The thesis of Stojanović (2026) takes this pair of bounds as its starting point; Part V develops the representations (low-rank, successor measures) designed to recover the single-reward rate for many rewards.

12.4 The Running Examples

Example 12.7 (Sample sizes for the goal CMP)**.** For the goal CMP [3.8], $|\mathcal{Z}| = 6$. With $\gamma = 0.9$, $\varepsilon = 0.1$ and $\delta = 0.05$, the elementary bound [12.2(a)] asks for

$$n \geq \frac{\log(240)}{2 \cdot 10^{-4} \cdot 10^{-2}} \approx 2.7 \times 10^6$$

samples per pair. The transitions of this CMP are deterministic, so $v^\pi \equiv 0$ for deterministic π, and every $(\hat{P} - P)V^\star(z)$ vanishes after a single sample per pair: Bernstein's inequality, which pays for variance, captures this, and Hoeffding's does not. The gap between 10^6 and 1 is the content of [12.3] in an extreme case.

Assumed without proof in this chapter

[12.4] the minimax upper bounds; [12.5] the minimax lower bound.

13

Exploration by Optimism

Without a generative model the learner sees only the transitions its own policy produces, and the asymptotic theory of Chapter 8 simply assumes that every pair is visited often enough. *Optimism in the face of uncertainty* turns this assumption into a guarantee: act greedily with respect to an upper confidence bound on $Q^\star$; either the bound is accurate, or the policy visits the pairs where it is not and the uncertainty shrinks. The idea goes back to E^3 (Kearns and Singh, 2002) and R-max (Brafman and Tennenholtz, 2002), became a regret theory with UCRL2 (Jaksch et al., 2010), and reached minimax optimality with UCBVI (Azar et al., 2017). This chapter proves a regret bound for a Hoeffding-bonus value iteration in full and states the sharp results. Throughout, the setting is episodic.

13.1 Episodic Regret

Definition 13.1 (Episodic learning)**.** An *episodic MDP* is a finite MDP $(\mathcal{X}, \mathcal{A}, P, r)$ with $r : \mathcal{Z} \to [0,1]$ and a horizon $H \in \mathbb{N}$; episode $k = 1, \dots, K$ starts at an arbitrary $x_{k,0}$, and the learner chooses a Markov deterministic policy $\pi_k = (\pi_{k,h})_{h<H}$ based on all earlier episodes and observes $z_{k,h} = (x_{k,h}, a_{k,h})$, $x_{k,h+1} \sim P(\cdot|z_{k,h})$. $V_h^\star, Q_h^\star$ are the optimal values with $H-h$ steps to go [4.4], V_h^π those of π, and the *regret* is

$$\mathrm{Regret}(K) := \sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi_k}\right)(x_{k,0}).$$

$T := KH$ is the number of steps. $N_k(z)$ is the number of visits to z in episodes $1, \dots, k-1$, and $\hat{P}_k(\cdot|z)$ the empirical law of the next states observed at z in these episodes (an arbitrary fixed law if $N_k(z) = 0$).

Definition 13.2 (Optimistic value iteration)**.** Fix $\delta \in (0,1)$, $L := \log(6|\mathcal{Z}|HT/\delta)$, and the *bonus* $b_k(z) := H\sqrt{L/(N_k(z) \vee 1)}$. Before episode k compute $\bar{V}_{k,H} := 0$ and, for $h = H-1, \dots, 0$,

$$\bar{Q}_{k,h}(z) := \min\left\{H - h,\ r(z) + \hat{P}_k \bar{V}_{k,h+1}(z) + b_k(z)\right\},$$
$$\bar{V}_{k,h}(x) := \max_a \bar{Q}_{k,h}(x,a)$$

and play the greedy policy $\pi_{k,h}(x) \in \arg\max_a \bar{Q}_{k,h}(x,a)$.

Lemma 13.3 (Good event)**.** *With probability at least* $1 - \frac{2\delta}{3}$*, for all episodes* k*, steps* h *and pairs* z*,*

(a) $\left|(\hat{P}_k - P)V^\star_{h+1}(z)\right| \le b_k(z)$*;*

(b)

$$\|\hat{P}_k(\cdot|z) - P(\cdot|z)\|_{\ell^1} \le \sqrt{2(|\mathcal{X}| + L)/(N_k(z) \vee 1)}.$$

Proof. Realise the next states at z as an i.i.d. sequence $Y_{z,1}, Y_{z,2}, \dots \sim P(\cdot|z)$, used in order of the visits; then $\hat{P}_k(\cdot|z)$ is the empirical law of the first $N_k(z)$ terms, and the joint law of the interaction is unchanged (Lattimore and Szepesvári, 2020, Sec. 4.6). For a fixed z and a fixed sample size $n \in \{1, \dots, T\}$: (a) the function $V^\star_{h+1}$ is fixed and has range H, so by Hoeffding [0.12(a)] $|(\hat{P}^{(n)} - P)V^\star_{h+1}(z)| \le H\sqrt{L/(2n)}$ fails with probability at most $\frac{\delta}{3|\mathcal{X}|HT}$; (b)

$$\|\mu - \mu'\|_{\ell^1} = 2\max_{B \subseteq \mathcal{X}} (\mu(B) - \mu'(B))$$

[0.11(a)], and by Hoeffding for each of the $2^{|\mathcal{X}|}$ sets B and a union bound,

$$\|\hat{P}^{(n)} - P\|_{\ell^1} \le 2\sqrt{(|\mathcal{X}| \log 2 + L)/(2n)}$$

fails with probability at most $\frac{\delta}{3|\mathcal{X}|T}$. A union bound over z, h and n gives the claim; for $N_k(z) = 0$ both bounds hold trivially since $b_k(z) \ge H$ and $\|\cdot\|_{\ell^1} \le 2$. ∎

Lemma 13.4 (Optimism)**.** *On the event of [13.3],* $\bar{Q}_{k,h} \ge Q^\star_h$ *and* $\bar{V}_{k,h} \ge V^\star_h$ *for all* k, h*.*

Proof. Backward induction on h; $h = H$ is trivial. If $\bar{V}_{k,h+1} \ge V^\star_{h+1}$, then either $\bar{Q}_{k,h}(z) = H - h \ge Q^\star_h(z)$, or

$$\bar{Q}_{k,h}(z) \ge r(z) + \hat{P}_k V^\star_{h+1}(z) + b_k(z) \ge r(z) + PV^\star_{h+1}(z) = Q^\star_h(z)$$

by [13.3(a)] and [4.4]. Take maxima over a. ∎

Lemma 13.5 (Pigeonhole)**.** *For every realisation,*

$$\sum_{k \le K} \sum_{h < H} \frac{1}{\sqrt{N_k(z_{k,h}) \vee 1}} \le 2H|\mathcal{Z}| + 2\sqrt{2}\sqrt{|\mathcal{Z}|T}.$$

Proof. Fix z and let $n_k \le H$ be the number of visits to z in episode k, so $N_{k+1} = N_k + n_k$. The episodes with $N_k < H$ contribute at most $\sum n_k \le 2H$ (the counter is below H before each of them). For $N_k \ge H$, $N_k + n_k \le 2N_k$, and

$$\frac{n_k}{\sqrt{N_k}} \le \frac{\sqrt{2}n_k}{\sqrt{N_{k+1}}} \le 2\sqrt{2}\bigl(\sqrt{N_{k+1}} - \sqrt{N_k}\bigr),$$

since

$$\sqrt{N+n} - \sqrt{N} = \frac{n}{\sqrt{N+n} + \sqrt{N}} \ge \frac{n}{2\sqrt{N+n}}.$$

Telescoping, the contribution of z is at most $2H + 2\sqrt{2}\sqrt{N_{K+1}(z)}$. Sum over z and use Cauchy–Schwarz with $\sum_z N_{K+1}(z) = T$. ∎

Theorem 13.6 (Regret of optimistic value iteration)**.** *With probability at least* $1-\delta$*, the algorithm [13.2] satisfies*

$$\begin{aligned}\mathrm{Regret}(K) &\le \bigl(2\sqrt{L} + \sqrt{2(|\mathcal{X}|+L)}\bigr)H\bigl(2H|\mathcal{Z}| + 2\sqrt{2}\sqrt{|\mathcal{Z}|T}\bigr) + H\sqrt{2TL} \\ &= \tilde{O}\bigl(H|\mathcal{X}|\sqrt{|\mathcal{A}|T}\bigr).\end{aligned}$$

Proof. Work on the event of [13.3] intersected with the Azuma event below. By optimism [13.4], $(V_0^\star - V_0^{\pi_k})(x_{k,0}) \le \delta_{k,0}$, where $\delta_{k,h} := (\bar{V}_{k,h} - V_h^{\pi_k})(x_{k,h})$. Let $z = z_{k,h}$, $a = \pi_{k,h}(x_{k,h})$. Since

$$\bar{V}_{k,h}(x_{k,h}) = \bar{Q}_{k,h}(z) \le r(z) + \hat{P}_k\bar{V}_{k,h+1}(z) + b_k(z)$$

and $V_h^{\pi_k}(x_{k,h}) = r(z) + PV_{h+1}^{\pi_k}(z)$,

$$\delta_{k,h} \le b_k(z) + (\hat{P}_k - P)V_{h+1}^\star(z) + (\hat{P}_k - P)(\bar{V}_{k,h+1} - V_{h+1}^\star)(z) + P(\bar{V}_{k,h+1} - V_{h+1}^{\pi_k})(z).$$

The second term is at most $b_k(z)$ [13.3(a)]; the third at most $\|\hat{P}_k(\cdot|z) - P(\cdot|z)\|_{\ell^1} H$ [13.3(b)], since $0 \le \bar{V}_{k,h+1} - V_{h+1}^\star \le H$; the last equals $\delta_{k,h+1} + \xi_{k,h}$ with

$$\xi_{k,h} := P(\bar{V}_{k,h+1} - V_{h+1}^{\pi_k})(z_{k,h}) - (\bar{V}_{k,h+1} - V_{h+1}^{\pi_k})(x_{k,h+1}).$$

Unrolling over h ($\delta_{k,H} = 0$),

$$\delta_{k,0} \le \sum_{h<H}\Bigl[\bigl(2\sqrt{L} + \sqrt{2(|\mathcal{X}|+L)}\bigr)\frac{H}{\sqrt{N_k(z_{k,h}) \vee 1}} + \xi_{k,h}\Bigr].$$

The policy π_k and the functions $\bar{V}_{k,\cdot}$ are determined before episode k, so the $\xi_{k,h}$, ordered in time, are martingale differences with values in an interval of length $2H$; by Azuma–Hoeffding [0.12(c)], $\sum_{k,h}\xi_{k,h} \le H\sqrt{2TL}$ with probability at least $1 - \frac{\delta}{3}$. Sum over k and apply [13.5]. ∎

Intuition. The regret of an optimistic episode is paid at the pairs it visits, in proportion to their remaining uncertainty; visiting a pair reduces that uncertainty. The pigeonhole lemma counts how often one can be optimistic about something before having seen it enough times. The factor $\sqrt{|\mathcal{X}|}$ lost in the cross term $(\hat{P}_k - P)(\bar{V} - V^\star)$, whose second factor depends on the data, is what the sharper analyses remove.

13.2 The Sharp Results

Fact 13.7 (Minimax regret)**.**

(a) (UCBVI; Azar et al., 2017.) With Hoeffding bonuses,

$$\mathrm{Regret}(K) \le 20H^{3/2}L\sqrt{|\mathcal{X}||\mathcal{A}|K} + 250H^2|\mathcal{X}|^2|\mathcal{A}|L^2;$$

with Bernstein–Freedman bonuses built from empirical variances,

$$\mathrm{Regret}(K) \le 30HL\sqrt{|\mathcal{X}||\mathcal{A}|K} + 2500H^2|\mathcal{X}|^2|\mathcal{A}|L^2 + 4H^{3/2}\sqrt{KL},$$

i.e. $\tilde{O}(\sqrt{H|\mathcal{X}||\mathcal{A}|T})$ for large T, with probability $1-\delta$ and $L = \log(5H|\mathcal{X}||\mathcal{A}|T/\delta)$ (time-homogeneous P).

(b) (Lower bounds; Jaksch et al., 2010; Darwiche Domingues et al., 2021.) For time-homogeneous episodic MDPs the regret of any algorithm is

$$\Omega(\sqrt{H^2|\mathcal{X}||\mathcal{A}|K}) = \Omega(\sqrt{H|\mathcal{X}||\mathcal{A}|T})$$

in the worst case; for stage-dependent transitions it is $\Omega(\sqrt{H^3|\mathcal{X}||\mathcal{A}|K})$ (Darwiche Domingues et al., 2021, Thm. 9). The Bernstein version of UCBVI is therefore minimax optimal up to logarithmic factors for large T.

(c) (Optimistic Q-learning; Jin et al., 2018.) For stage-dependent transitions, Q-learning with step size $\alpha_t = \frac{H+1}{H+t}$ and Hoeffding bonus $c\sqrt{H^3\iota/t}$ has regret $O(\sqrt{H^4|\mathcal{X}||\mathcal{A}|T\iota})$, and with Bernstein bonuses $O(\sqrt{H^3|\mathcal{X}||\mathcal{A}|T\iota} + H^9|\mathcal{X}|^3|\mathcal{A}|^3\iota^2)$, $\iota = \log(|\mathcal{X}||\mathcal{A}|T/\delta)$: model-free learning matches the model-based rate up to $\sqrt{H}$.

Fact 13.8 (Linear MDPs; Jin et al., 2020b)**.** *Call an episodic MDP* linear *with features $\varphi : \mathcal{Z} \to \mathbb{R}^d$ if $P(\cdot|z) = \langle \varphi(z), \mathsf{m}(\cdot)\rangle$ and $r(z) = \langle \varphi(z), w\rangle$ for signed measures $\mathsf{m} = (\mathsf{m}_1, \dots, \mathsf{m}_d)$ and $w \in \mathbb{R}^d$ (with norm bounds). Least-squares value iteration with the elliptical bonus $\beta\|\varphi(z)\|_{\Lambda_k^{-1}}$, Λ_k the regularised feature*

covariance (LSVI-UCB), has regret $\tilde{O}(\sqrt{d^3H^3T})$ *with probability* $1-\delta$ *and running time polynomial in d, H,* $|\mathcal{A}|$ *and T; neither* $|\mathcal{X}|$ *nor* $|\mathcal{A}|$ *enters the regret. Tabular MDPs are the case* $d = |\mathcal{X}||\mathcal{A}|$ *with indicator features. The assumption is on the transition kernel, not on* $Q^\star$*: under linear realisability of* $Q^\star$ *alone, exponential lower bounds hold.*

13.3 The Running Examples

Example 13.9 (Optimism in the goal CMP)**.** Run [13.2] on the goal CMP [3.8] with horizon $H = 3$, start $x_{k,0} = 1$ and $\delta = 0.05$. The bonus satisfies $b_k(z) \geq H \geq H - h$ until $N_k(z) \geq L$, so at first every $\bar{Q}_{k,h}(z)$ equals its cap $H - h$: all actions look equally good, and the choice is made by tie-breaking. With ties broken towards the less-tried action, the agent alternates b and f and reaches g in the first episodes; with a fixed order it may repeat b for about L visits of $(1, \mathsf{b})$ before the cap is undercut and f is tried. Either way the bound [13.6] holds, since such episodes are counted by the term $2H|\mathcal{Z}|$ of [13.5]. Optimism supplies the exploration that Q-learning had to be given from outside in [8.13]; the price of a worst-case bonus is a burn-in of order L visits per pair.

Assumed without proof in this chapter

[13.7] the sharp UCBVI bounds, the regret lower bounds and optimistic Q-learning; [13.8] LSVI-UCB for linear MDPs.

14

State Abstraction, Bisimulation and Block MDPs

The sample complexities of Chapters 11–13 grow with $|\mathcal{X}|$. Every way around this assumes structure, and the structures form a succession. An *abstraction* groups states that can be treated alike, and the theory asks which groupings preserve optimality (Li et al., 2006). *Bisimulation* is the coarsest grouping that preserves the model, and its metric relaxation measures how far two states are from being interchangeable (Ferns et al., 2004). A *block MDP* is an MDP whose observations are generated from a small latent MDP through a grouping that the learner does not know and must discover (Du et al., 2019). The next chapter continues the succession with linear and low-rank MDPs, which replace groupings by feature maps; block MDPs are low-rank MDPs whose features are indicators of latent states, as shown below. $\mathcal{X}$, $\mathcal{A}$ are finite unless stated otherwise.

14.1 Abstractions and Their Hierarchy

Definition 14.1 (Abstractions; Li et al., 2006)**.** An *abstraction* is a surjection $\phi : \mathcal{X} \to \bar{\mathcal{X}}$. With a *weighting* $\mathsf{w}(\cdot|\bar{x}) \in \mathcal{P}(\phi^{-1}(\bar{x}))$, the *abstract MDP* has reward and kernel

$$\bar{r}(\bar{x}, a) := \sum_{x \in \phi^{-1}(\bar{x})} \mathsf{w}(x|\bar{x}) r(x, a),$$

$$\bar{P}(\bar{y}|\bar{x}, a) := \sum_{x \in \phi^{-1}(\bar{x})} \mathsf{w}(x|\bar{x}) P\big(\phi^{-1}(\bar{y})\big|x, a\big).$$

A policy $\bar{\pi}$ on $\bar{\mathcal{X}}$ is *lifted* to $\bar{\pi} \circ \phi$. A function f on $\mathcal{Z}$ is *ϕ-measurable* if $f(x, a)$ depends on x only through $\phi(x)$. For x_1, x_2 with $\phi(x_1) = \phi(x_2)$ and all a, ϕ is

(a) *model-irrelevant* if $r(x_1, a) = r(x_2, a)$ and $P(\phi^{-1}(\bar{y})|x_1, a) = P(\phi^{-1}(\bar{y})|x_2, a)$ for all $\bar{y}$;
(b) *Q^π-irrelevant* if $Q^{\bar{\pi}\circ\phi}(x_1, a) = Q^{\bar{\pi}\circ\phi}(x_2, a)$ for every abstract policy $\bar{\pi}$;
(c) *$Q^\star$-irrelevant* if $Q^\star(x_1, a) = Q^\star(x_2, a)$;
(d) *$a^\star$-irrelevant* if every abstract class has an action optimal at all its states with equal optimal value;
(e) *$\pi^\star$-irrelevant* if every abstract class has an action optimal at all its states.

Theorem 14.2 (The hierarchy of abstractions; Li et al., 2006, Thm. 2)**.** *Each class in [14.1] is contained in the next: model-irrelevant $\Rightarrow$ Q^π-irrelevant $\Rightarrow$ $Q^\star$-irrelevant $\Rightarrow$ $a^\star$-irrelevant $\Rightarrow$ $\pi^\star$-irrelevant.*

Proof. *Model $\Rightarrow$ Q^π.* Let f be ϕ-measurable on $\mathcal{Z}$ and $\bar\pi$ abstract. Then $V(x) := \sum_a \bar\pi(a|\phi(x)) f(x,a)$ is constant on classes, so

$$\sum_{x'} P(x'|x,a)V(x') = \sum_{\bar y} P(\phi^{-1}(\bar y)|x,a)V(\bar y)$$

is ϕ-measurable by model irrelevance, and so is $T^{\bar\pi\circ\phi} f$. The ϕ-measurable functions form a closed subspace invariant under the contraction $T^{\bar\pi\circ\phi}$, which therefore contains its fixed point $Q^{\bar\pi\circ\phi}$ [0.8]. $Q^\pi \Rightarrow Q^\star$. Run policy iteration [4.7] from an abstract deterministic policy. If Q^{π_k} is ϕ-measurable, a greedy policy can be chosen constant on classes (break ties by a fixed order), so π_{k+1} is again abstract and $Q^{\pi_{k+1}}$ is ϕ-measurable. Policy iteration reaches an optimal policy [4.7(c)], so $Q^\star$ is ϕ-measurable. $Q^\star \Rightarrow a^\star$: equal $Q^\star(x,\cdot)$ on a class give a common maximiser with equal value. $a^\star \Rightarrow \pi^\star$ is a weakening. ∎

Theorem 14.3 (Abstractions that preserve optimality; Li et al., 2006, Thm. 3)**.** *Let ϕ be $Q^\star$-irrelevant (in particular model-irrelevant [14.2]) and w any weighting. Then $\bar Q(\bar x,a) := Q^\star(x,a)$, $x \in \phi^{-1}(\bar x)$, is the optimal action value of the abstract MDP, and every abstract greedy policy, lifted, is optimal in the ground MDP. For $\pi^\star$-irrelevant abstractions the conclusion can fail (Li et al., 2006, Fig. 1(b)).*

Proof. $\bar Q$ is well defined by irrelevance. With $\bar V := \max_a \bar Q$,

$$\begin{aligned}\bar r(\bar x,a) + \gamma \sum_{\bar y} \bar P(\bar y|\bar x,a)\bar V(\bar y) &= \sum_x \mathsf{w}(x|\bar x)\Big[r(x,a) + \gamma \sum_{x'} P(x'|x,a)V^\star(x')\Big] \\ &= \sum_x \mathsf{w}(x|\bar x) Q^\star(x,a) \\ &= \bar Q(\bar x,a),\end{aligned}$$

because $\bar V(\phi(x')) = V^\star(x')$. So $\bar Q$ solves the abstract Bellman optimality equation and is its optimal value [4.5]. A greedy abstract action at $\bar x$ maximises $Q^\star(x,\cdot)$ for every $x \in \phi^{-1}(\bar x)$, so the lifted policy is greedy for $Q^\star$ and optimal [4.5(b)]. ∎

Theorem 14.4 (Approximate abstraction; Abel et al., 2016, Lemma 1)**.** *Suppose $|Q^\star(x_1,a) - Q^\star(x_2,a)| \le \epsilon$ whenever $\phi(x_1) = \phi(x_2)$. Let $\bar Q^\star$ be the optimal*

action value of the abstract MDP, $\tilde{Q}(x,a) := \bar{Q}^\star(\phi(x),a)$ *its lift, and* $\bar{\pi}$ *an abstract greedy policy. Then*

$$\|\tilde{Q} - Q^\star\|_\infty \le \frac{\epsilon}{1-\gamma}, \qquad \|V^\star - V^{\bar{\pi}\circ\phi}\|_\infty \le \frac{2\epsilon}{(1-\gamma)^2}.$$

Proof. Let A average within classes,

$$\mathsf{A}f(x,a) := \sum_{x'' \in \phi^{-1}(\phi(x))} \mathsf{w}(x''\,|\,\phi(x))\, f(x'',a),$$

a nonexpansion with $\|\mathsf{A}Q^\star - Q^\star\|_\infty \le \epsilon$. The computation of [14.3] shows that the lifted abstract Bellman operator acts on ϕ-measurable Q as $\mathsf{A}T^\star Q$, so $\tilde{Q} = \mathsf{A}T^\star\tilde{Q}$. Hence

$$\|\tilde{Q} - Q^\star\|_\infty \le \|\mathsf{A}T^\star\tilde{Q} - \mathsf{A}T^\star Q^\star\|_\infty + \|\mathsf{A}Q^\star - Q^\star\|_\infty \le \gamma\|\tilde{Q} - Q^\star\|_\infty + \epsilon.$$

The lifted policy is greedy for $\tilde{Q}$; apply [4.6(b)]. ∎

14.2 Bisimulation and Its Metric

Definition 14.5 (Bisimulation; Givan et al., 2003)**.** An equivalence relation $\sim$ on $\mathcal{X}$ is a *bisimulation* if $x \sim y$ implies $r(x,a) = r(y,a)$ and $P(C\,|\,x,a) = P(C\,|\,y,a)$ for every action a and every equivalence class C. Equivalently, the quotient map is a model-irrelevant abstraction [14.1(a)].

Theorem 14.6 (Bisimulation metrics; Ferns et al., 2004)**.** *Let* $c_R > 0$ *and* $c_T \in [\gamma, 1)$*. For pseudometrics* d *on* $\mathcal{X}$ *let* W_d *be the transport cost of* d *[6.3] and*

$$\mathsf{F}(d)(x,y) := \max_a \Big[c_R\,|r(x,a) - r(y,a)| + c_T\, W_d\big(P(\cdot\,|\,x,a), P(\cdot\,|\,y,a)\big)\Big].$$

(a) F *maps pseudometrics to pseudometrics and is a* c_T*-contraction in the sup norm; it has a unique fixed point* $d^\star$*, the* bisimulation metric*.*

(b) $c_R\,|V^\star(x) - V^\star(y)| \le d^\star(x,y)$ *for all* x, y*.*

(c) If $\sim$ *is a bisimulation and* $x \sim y$*, then* $d^\star(x,y) = 0$*.*

Proof.

(a) Symmetry and $\mathsf{F}(d)(x,x) = 0$ are clear; the triangle inequality follows from that of W_d [6.4] and of $|\cdot|$, after taking the maximum over a separately. For two pseudometrics, any coupling has

$$\Big|\int d\,d\varsigma - \int d'\,d\varsigma\Big| \le \|d - d'\|_\infty,$$

so $|W_d - W_{d'}| \le \|d - d'\|_\infty$ and

$$\|\mathsf{F}(d) - \mathsf{F}(d')\|_\infty \le c_T \|d - d'\|_\infty.$$

Pseudometrics form a closed subset of $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$; apply [0.8].

(b) Let $V_0 := 0$, $V_{n+1} := T^\star V_n$, $d_0 := 0$, $d_{n+1} := \mathsf{F}(d_n)$. *Claim:* $c_R V_n$ is d_n-Lipschitz. If so, for any coupling ς of $P(\cdot|x,a), P(\cdot|y,a)$, $c_R|PV_n(x,a) - PV_n(y,a)| \le \int d_n \, d\varsigma$, hence $\le W_{d_n}$, and

$$\begin{aligned} c_R|V_{n+1}(x) - V_{n+1}(y)| &\le \max_a \big[c_R|r(x,a) - r(y,a)| \\ &\qquad + \gamma W_{d_n}(P(\cdot|x,a), P(\cdot|y,a))\big] \\ &\le d_{n+1}(x,y), \end{aligned}$$

using $\gamma \le c_T$. Let $n \to \infty$ [4.6(a)], (a).

(c) The set of pseudometrics vanishing on $\sim$ is closed and mapped into itself by F: if d vanishes on $\sim$ and $x \sim y$, then $r(x,a) = r(y,a)$, and coupling $P(\cdot|x,a)$ and $P(\cdot|y,a)$ class by class (equal class masses) gives a coupling supported on $\sim$-related pairs, of cost 0. The fixed point lies in this set.

■

Intuition. Two states are close in the bisimulation metric if they pay similar rewards now and lead to distributions that can be matched cheaply in the same metric later. It is the value-relevant geometry of the state space, built from the same transport cost as the return-law metrics of Chapter 17, and it bounds how much the optimal value can differ between states that an abstraction merges.

14.3 Block MDPs

Definition 14.7 (Block MDPs; Du et al., 2019)**.** A *block MDP* consists of a finite latent MDP $(\mathcal{S}, \mathcal{A}, P_{\text{lat}}, r_{\text{lat}})$, a countable observation space $\mathcal{X}$ partitioned into disjoint *blocks* $\mathcal{X}_s$, $s \in \mathcal{S}$, and emission laws $\mathsf{q}(\cdot|s) \in \mathscr{P}(\mathcal{X}_s)$. The learner observes $x_t \sim \mathsf{q}(\cdot|s_t)$, never s_t. The *decoder* $f^\star : \mathcal{X} \to \mathcal{S}$, $f^\star(x) = s$ for $x \in \mathcal{X}_s$, exists by the block structure but is unknown.

Proposition 14.8 (What block structure buys)**.** *In a block MDP with rewards* $r(x,a) := r_{\text{lat}}(f^\star(x), a)$*:*

(a) the observations form a CMP with kernel

$$P(x'|x,a) = \sum_{s'} P_{\text{lat}}(s'|f^\star(x), a)\, \mathsf{q}(x'|s');$$

(b) the decoder $f^\star$ is a model-irrelevant abstraction (a bisimulation [14.5]) whose abstract MDP is the latent MDP, so optimal latent policies lifted by $f^\star$ are optimal [14.3];

(c) the kernel factorises as $P(x'|x,a) = \langle \varphi(x,a), \mathsf{m}(x') \rangle$ with

$$\varphi(x,a) := P_{\text{lat}}(\cdot|f^\star(x),a) \in \mathbb{R}^{\mathcal{S}}$$

and $\mathsf{m}(x') := (\mathsf{q}(x'|s'))_{s'\in\mathcal{S}}$: a block MDP is a low-rank MDP of dimension $|\mathcal{S}|$ whose features are unknown because $f^\star$ is.

Proof. (a) Given x, the latent state is $f^\star(x)$; draw s' from P_{lat} and x' from $\mathsf{q}(\cdot|s')$. (b) $P(\mathcal{X}_{s'}|x,a) = P_{\text{lat}}(s'|f^\star(x),a)$ depends on x only through $f^\star(x)$, and so does r; any weighting gives $\bar{P} = P_{\text{lat}}$. (c) Rewrite (a). ∎

Fact 14.9 (Learning in block MDPs; Du et al., 2019; Misra et al., 2020)**.** *Under a separability (identifiability) condition and with a regression oracle over a decoder class $\mathcal{G}$, a block MDP with horizon H can be explored and its decoder learned with a number of trajectories polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, H, $1/\mu_{\min}$ (the smallest maximal reaching probability of a latent state) and* $\log|\mathcal{G}|$*, and independent of $|\mathcal{X}|$ (Du et al., 2019, Thm. 4.1); Misra et al. (2020) remove the separability condition by learning* kinematically inseparable *abstractions, which merge observations with identical backward and forward dynamics.*

14.4 The Running Examples

Example 14.10 (Abstraction of the gridworld)**.** In the gridworld CMP [3.9] with the goal $g = (5,5)$, the reflection $(i,j) \mapsto (j,i)$ is a symmetry only after exchanging N with E and S with W; with the action labels kept fixed, two reflected states differ in their next-state laws, so the reflection is *not* a model-irrelevant abstraction in the sense of [14.1], and exploiting it needs abstractions that also map actions. By contrast, if the gridworld is observed through images that differ from cell to cell, each cell's images forming one block, the problem is a block MDP with 25 latent states [14.7], the cell decoder is a bisimulation, and by [14.8(c)] the observed process is a low-rank MDP of dimension 25 however many images there are.

Assumed without proof in this chapter

[14.9] the sample complexity of block-MDP learning.

15

Linear and Low-Rank MDPs

The succession of structures continues from groupings to features. A *linear MDP* has a transition kernel that is linear in a known feature map; every value function of every policy is then linear in the same features, and exploration can be done in the d-dimensional feature space (Jin et al., 2020b; Yang and Wang, 2019). A *low-rank MDP* has the same factorisation with unknown features, so representation learning and exploration must be solved together (Agarwal et al., 2020; Uehara et al., 2022). Both are special cases of MDPs with low *Bellman rank* (Jiang et al., 2017). The assumption on the kernel cannot be weakened to linearity of $Q^\star$ alone: then exponentially many samples are needed (Weisz et al., 2021). The analysis rests on one deterministic inequality, the elliptical-potential lemma, which is the feature-space version of the pigeonhole lemma [13.5]. The setting is episodic as in [13.1], with $\mathcal{X}$ possibly infinite and $\mathcal{A}$ finite.

15.1 Linear MDPs

Definition 15.1 (Linear MDPs; Jin et al., 2020b)**.** An episodic MDP is *linear* with a known feature map $\varphi : \mathcal{Z} \to \mathbb{R}^d$, $\|\varphi(z)\|_2 \le 1$, if for each step h there are signed measures $\mathsf{m}_h = (\mathsf{m}_h^{(1)}, \dots, \mathsf{m}_h^{(d)})$ on $\mathcal{X}$ and $\theta_h \in \mathbb{R}^d$ with

$$P_h(\cdot \mid z) = \langle \varphi(z), \mathsf{m}_h(\cdot) \rangle, \qquad r_h(z) = \langle \varphi(z), \theta_h \rangle,$$
$$\|\mathsf{m}_h(\mathcal{X})\|_2 \le \sqrt{d}, \quad \|\theta_h\|_2 \le \sqrt{d}.$$

A tabular MDP is linear with $d = |\mathcal{Z}|$ and one-hot features; a block MDP is linear with $d = |\mathcal{S}|$ and the features of [14.8(c)], which are unknown.

Proposition 15.2 (Every action value is linear)**.** *In a linear MDP, for every policy π and step h there is $\mathsf{w}_h^\pi \in \mathbb{R}^d$ with $Q_h^\pi(z) = \langle \varphi(z), \mathsf{w}_h^\pi \rangle$ and $\|\mathsf{w}_h^\pi\|_2 \le 2H\sqrt{d}$. In particular $Q_h^\star$ is linear, and every Bellman backup of any bounded function is linear:*

$$r_h(z) + \int P_h(dx' \mid z) f(x') = \langle \varphi(z), \theta_h + \int f \, d\mathsf{m}_h \rangle.$$

Proof.

$$Q_h^\pi(z) = r_h(z) + \int P_h(dx'\,|\,z) V_{h+1}^\pi(x') = \langle \varphi(z), \theta_h + \int V_{h+1}^\pi \, d\mathsf{m}_h \rangle$$

by [4.4]-type backward recursion; $V_{h+1}^\pi \in [0, H]$ gives $\| \int V_{h+1}^\pi d\mathsf{m}_h \|_2 \le H\sqrt{d}$ coordinatewise after splitting m_h into positive and negative parts (as in the paper), and $\|\theta_h\|_2 \le \sqrt{d} \le H\sqrt{d}$. ∎

Intuition. Linearity of the kernel is a closure property: the Bellman backup of *any* function lands in the span of the features. This is what lets least-squares value iteration work without ever projecting out of the class, and it is exactly what fails when only $Q^\star$ is assumed linear, as the lower bound below shows.

Lemma 15.3 (Elliptical potential; Abbasi-Yadkori et al., 2011)**.** *Let $\varphi_1, \dots, \varphi_K \in \mathbb{R}^d$ with $\|\varphi_k\|_2 \le 1$, $\Lambda_k := I + \sum_{j<k} \varphi_j \varphi_j^\top$ and $\|v\|_{\Lambda^{-1}}^2 := v^\top \Lambda^{-1} v$. Then*

$$\sum_{k \le K} \min\left(1, \|\varphi_k\|_{\Lambda_k^{-1}}^2\right) \le 2 \log \det \Lambda_{K+1} \le 2d \log\left(1 + \frac{K}{d}\right),$$

and consequently

$$\sum_{k \le K} \min\left(1, \|\varphi_k\|_{\Lambda_k^{-1}}\right) \le \sqrt{2dK \log(1 + K/d)}.$$

Proof. By the matrix determinant lemma,

$$\det \Lambda_{k+1} = \det \Lambda_k \, (1 + \|\varphi_k\|_{\Lambda_k^{-1}}^2),$$

so

$$\log \det \Lambda_{K+1} = \sum_k \log(1 + \|\varphi_k\|_{\Lambda_k^{-1}}^2),$$

and $\min(1, u) \le 2\log(1+u)$ for $u \ge 0$. By the AM–GM inequality on the eigenvalues,

$$\det \Lambda_{K+1} \le (\operatorname{tr} \Lambda_{K+1}/d)^d \le (1 + K/d)^d.$$

The last claim is Cauchy–Schwarz. ∎

Fact 15.4 (Self-normalised concentration; Abbasi-Yadkori et al., 2011, Thm. 1)**.** *If ζ_k is conditionally R-sub-Gaussian given $\mathcal{F}_{k-1}$ and φ_k is $\mathcal{F}_{k-1}$-measurable, then with probability at least $1 - \delta$, for all K,*

$$\Big\| \sum_{k \le K} \varphi_k \zeta_k \Big\|_{\Lambda_{K+1}^{-1}}^2 \le 2R^2 \log\left(\det(\Lambda_{K+1})^{1/2} / \delta \right).$$

Theorem 15.5 (Regret in linear MDPs; Jin et al., 2020b, Thm. 3.1)**.**

(a) (Mechanism.) Any optimistic algorithm whose per-step optimism gap at the visited pair is bounded by

$$\min\bigl(H,\, 2\beta\|\varphi(z_{k,h})\|_{\Lambda_{k,h}^{-1}}\bigr)$$

has regret at most $\sum_h \bigl(2\beta + H\bigr)\sqrt{2dK\log(1+K/d)}$ *plus a martingale term of order* $H\sqrt{T\log(1/\delta)}$.

(b) (Fact.) Least-squares value iteration with the bonus $\beta\|\varphi(z)\|_{\Lambda^{-1}}$, $\beta = c\,dH\sqrt{\iota}$, $\iota = \log(2dT/\delta)$ *(LSVI-UCB) has optimism gaps of this form, and its regret is* $O(\sqrt{d^3H^3T}\,\iota^2)$ *with probability* $1-\delta$*; neither* $|\mathcal{X}|$ *nor* $|\mathcal{A}|$ *enters.*

Proof. (a) As in the proof of [13.6], the regret is at most the sum of optimism gaps along the visited pairs plus a martingale term; by [15.3] for each h,

$$\sum_k \min\bigl(1,\, \|\varphi(z_{k,h})\|_{\Lambda_{k,h}^{-1}}\bigr) \le \sqrt{2dK\log(1+K/d)},$$

and $\min(H, 2\beta u) \le (2\beta+H)\min(1,u)$. (b) The confidence widths come from [15.4] applied to the regression noise uniformly over the value class, which is where the extra factor d in β arises (Jin et al., 2020b). ∎

Fact 15.6 (Linear $Q^\star$ is not enough; Weisz et al., 2021)**.** *If only* $Q^\star$ *is assumed to lie in the span of known* d*-dimensional features, then even with a generative model every sound planner needs at least* $\min(e^{\Omega(d)}, \Omega(2^H))$ *queries in the fixed-horizon setting and* $e^{\Omega(d)}$ *in the discounted setting.*

15.2 Low-Rank MDPs and Bellman Rank

Definition 15.7 (Low-rank MDPs; Agarwal et al., 2020)**.** An episodic MDP is *low-rank* with dimension d if

$$P_h(x'\,|\,z) = \langle \varphi_h^\star(z),\, \mathsf{m}_h^\star(x')\rangle$$

for unknown $\varphi_h^\star$, $\mathsf{m}_h^\star$ belonging to known function classes Φ, Υ (with the normalisations of [15.1]). A linear MDP is a low-rank MDP with Φ a singleton; a block MDP is a low-rank MDP with d latent states [14.8(c)].

Fact 15.8 (Representation learning with exploration)**.**

(a) FLAMBE learns a model that is accurate for all policies with

$$\tilde{O}\big(H^{22}|\mathcal{A}|^{9}d^{7}\log(|\Phi||\Upsilon|/\delta)/\varepsilon^{10}\big)$$

trajectories, using maximum-likelihood and planning oracles (Agarwal et al., 2020, Thm. 2).

(b) REP-UCB, which adds an elliptical bonus in the learned features to a maximum-likelihood model, finds an ε-optimal policy with

$$\tilde{O}\big(d^{4}|\mathcal{A}|^{2}\log(|\mathcal{M}|/\delta)/((1-\gamma)^{5}\varepsilon^{2})\big)$$

samples in the discounted setting (Uehara et al., 2022, Thm. 4).

Neither bound depends on $|\mathcal{X}|$; both depend on the complexity of the function classes only logarithmically.

Fact 15.9 (Bellman rank; Jiang et al., 2017)**.** *Call a class $\mathcal{Q}$ of candidate action-value functions of* Bellman rank *M if the matrix of average Bellman errors*

$$\mathcal{E}(f,f')_h := \mathbb{E}^{\pi_f}\big[f'_h(Z_h) - r_h(Z_h) - \max_a f'_{h+1}(X_{h+1},a)\big],$$

indexed by the roll-in function f and the tested function f', factorises with inner dimension M at every h. If $Q^\star \in \mathcal{Q}$, the algorithm Olive returns an ε-optimal policy after

$$\tilde{O}\big(M^{2}H^{3}|\mathcal{A}|\varepsilon^{-2}\log(|\mathcal{Q}|/\delta)\big)$$

episodes (Jiang et al., 2017, Thm. 1). In a low-rank MDP the law of X_h under any roll-in is

$$\int \mathbb{P}^{\pi_f}\{Z_{h-1}\in dz\}\,\langle \varphi^\star_{h-1}(z), \mathsf{m}^\star_{h-1}(\cdot)\rangle,$$

so $\mathcal{E}(f,f')_h$ is an inner product of a vector depending on f (the mean feature $\mathbb{E}^{\pi_f}\varphi^\star_{h-1}(Z_{h-1})$) and one depending on f', when the action at step h does not depend on f (as in Olive, which draws it uniformly): the Bellman rank is at most d. Tabular and block MDPs are the cases $d=|\mathcal{X}|$ and $d=|\mathcal{S}|$.

Remark 15.10 (The succession and the successor measure)**.** Tabular $\subset$ block $\subset$ low-rank, and linear $\subset$ low-rank: each step either enlarges the class of kernels or removes knowledge of the representation. In each case the factorisation $P=\langle\varphi,\mathsf{m}\rangle$ makes the shifted successor measure of every policy of rank at most d (Chapter on low-rank structure, Part V), so the reward-free cost of Chapter 16 drops from $|\mathcal{X}|^2|\mathcal{A}|$ towards $d|\mathcal{X}||\mathcal{A}|$ (Stojanović, 2026) and, with known features, to quantities free of $|\mathcal{X}|$. Part V returns to these factorisations as representations of successor measures.

15.3 The Running Examples

Example 15.11 (Features for the goal CMP)**.** In the goal CMP [3.8] the transitions are deterministic, so with one-hot features $\varphi(z) = e_z \in \mathbb{R}^6$ the MDP is linear with $\mathsf{m}(x') = (P(x'|z))_z$. A smaller exact representation exists: $P(\cdot|z) = \delta_{\text{next}(z)}$ takes only the three values $\delta_1, \delta_2, \delta_g$, so $\varphi(z) := e_{\text{next}(z)} \in \mathbb{R}^3$ and $\mathsf{m}(x') := e_{x'}$ give a linear MDP with $d = 3$, and the goal reward is linear in the same features only if it is expressed on the next state. With $d = 3$, the elliptical-potential bound [15.3] replaces the pigeonhole count over six pairs by one over three directions.

Assumed without proof in this chapter

[15.4] self-normalised concentration; [15.5(b)] the regret of LSVI-UCB; [15.6] the lower bound for linear $Q^\star$; [15.8] representation learning bounds; [15.9] the Bellman-rank bound.

16

Reward-Free Exploration

A foundation policy is trained before the task is known. The statistical prototype of this situation is *reward-free exploration* (Jin et al., 2020a): the learner explores without a reward signal, and afterwards must return an ε-optimal policy for every reward function revealed to it. This chapter shows that a reward-free, occupancy-weighted model error controls all rewards at once, proves a reward-free sample bound with a generative model, and states the online upper and lower bounds (Jin et al., 2020a; Kaufmann et al., 2021). The lower bound exceeds the single-reward rate of Chapter 12 by a factor $|\mathcal{X}|$: this is the price of a model that serves all rewards, and the starting point of the representation theory of Part V. The setting is episodic, as in [13.1], with $\mathcal{X}, \mathcal{A}$ finite and rewards $r : \mathcal{Z} \to [0,1]$.

16.1 The Protocol

Definition 16.1 (Reward-free exploration)**.** In the *exploration phase* the learner runs episodes of length H from $x_0 \sim \nu$ with policies of its choice and observes no rewards. It stops after N episodes and outputs an estimate $\hat{P}$ of the kernel. In the *planning phase* any reward r is revealed, and the learner returns $\hat{\pi}_r$, an optimal policy of the MDP $(\hat{P}, r)$. The algorithm is (ε, δ)-*PAC* if with probability at least $1-\delta$, for *all* rewards r simultaneously, $V_0^\star(\nu; r) - V_0^{\hat{\pi}_r}(\nu; r) \le \varepsilon$, where values carry the reward as an argument and are computed under the true P (Kaufmann et al., 2021).

16.2 A Reward-Free Simulation Lemma

Theorem 16.2 (One model error for all rewards)**.** *For a Markov policy π let*

$$\mathsf{e}(\pi) := \sum_{h<H} \mathbb{E}_\nu^\pi \big\| \hat{P}(\cdot|Z_h) - P(\cdot|Z_h) \big\|_{\ell^1},$$

the model error weighted by the occupancy of π under the true kernel. Then for every reward r and every Markov π,

$$\left|V_0^{\pi}(\nu;r) - \hat{V}_0^{\pi}(\nu;r)\right| \le \tfrac{H}{2}\,\mathsf{e}(\pi),$$
$$V_0^{\star}(\nu;r) - V_0^{\hat{\pi}_r}(\nu;r) \le \tfrac{H}{2}\big(\mathsf{e}(\pi_r^{\star}) + \mathsf{e}(\hat{\pi}_r)\big) \le H \sup_{\pi} \mathsf{e}(\pi)$$

where $\hat{V}$ denotes values under $\hat{P}$ and $\pi_r^{\star}$ an optimal Markov policy for (P, r) [4.4]. The quantity $\sup_{\pi} \mathsf{e}(\pi)$ does not depend on the reward.

Proof. Write $\hat{V}_h := \hat{V}_h^{\pi}(\cdot; r) \in [0, H]$. Under the true trajectory measure,

$$\hat{V}_h(X_h) - \hat{V}_{h+1}(X_{h+1}) = r(Z_h) + (\hat{P}\hat{V}_{h+1})(Z_h) - \hat{V}_{h+1}(X_{h+1})$$

[4.4]; sum over $h < H$, take $\mathbb{E}_{\nu}^{\pi}$ and use

$$\mathbb{E}_{\nu}^{\pi}[\hat{V}_{h+1}(X_{h+1}) \mid \mathcal{F}_h] = P\hat{V}_{h+1}(Z_h)$$

[3.3(c)]:

$$\hat{V}_0(\nu) - V_0^{\pi}(\nu;r) = \sum_{h<H} \mathbb{E}_{\nu}^{\pi}\big[(\hat{P} - P)\hat{V}_{h+1}(Z_h)\big].$$

Each term is at most

$$\operatorname{sp}(\hat{V}_{h+1})\,\|\hat{P} - P\|_{\mathrm{TV}}(Z_h) \le \frac{H}{2}\|\hat{P} - P\|_{\ell^1}(Z_h)$$

[0.11(c)]. For the second claim,

$$V_0^{\star} - V_0^{\hat{\pi}_r} = (V_0^{\pi_r^{\star}} - \hat{V}_0^{\pi_r^{\star}}) + (\hat{V}_0^{\pi_r^{\star}} - \hat{V}_0^{\hat{\pi}_r}) + (\hat{V}_0^{\hat{\pi}_r} - V_0^{\hat{\pi}_r}),$$

and the middle term is ≤ 0 because $\hat{\pi}_r$ is optimal for $(\hat{P}, r)$. ∎

Intuition. The reward enters the error only through a value function with range H; what must be small is the model error along the trajectories of the policies that will be compared, and these are unknown when the data are collected. So the data must make $\hat{P}$ accurate in total variation wherever *any* policy goes with appreciable probability. This is coverage, and it is reward-free.

Corollary 16.3 (Reward-free planning with a generative model)**.** *Suppose the learner draws n samples at every pair from a generative model [12.1] and sets $\hat{P}$*

to the empirical kernel. With $L := \log(|\mathcal{Z}|/\delta)$*, with probability at least* $1-\delta$*, simultaneously for all rewards,*

$$V_0^\star(\nu;r) - V_0^{\hat{\pi}_r}(\nu;r) \le H^2\sqrt{\frac{2(|\mathcal{X}|+L)}{n}}.$$

Hence $n = \frac{2H^4(|\mathcal{X}|+L)}{\varepsilon^2}$ *per pair, i.e.*

$$N_{\mathrm{tot}} = O\Big(\frac{|\mathcal{X}|^2|\mathcal{A}|H^4}{\varepsilon^2} + \frac{|\mathcal{X}||\mathcal{A}|H^4\log(|\mathcal{Z}|/\delta)}{\varepsilon^2}\Big)$$

transitions in total, suffice.

Proof. As in [13.3(b)], with probability $\ge 1-\delta$ every row satisfies

$$\|\hat{P}(\cdot|z) - P(\cdot|z)\|_{\ell^1} \le \sqrt{2(|\mathcal{X}|+L)/n}.$$

Then $\mathsf{e}(\pi) \le H\sqrt{2(|\mathcal{X}|+L)/n}$ for every π; apply [16.2]. ∎

16.3 Online Reward-Free Exploration

Fact 16.4 (Upper bounds; Jin et al., 2020a; Kaufmann et al., 2021)**.** *Without a generative model, from a fixed initial law:*

(a) Jin et al. (2020a) first explore with one regret-minimising algorithm per pair (state, step), using the reward that indicates a visit to that state at that step, store the resulting policies, and then collect data with policies drawn from this buffer. This is (ε,δ)*-PAC with*

$$\tilde{O}\Big(\frac{|\mathcal{X}|^2|\mathcal{A}|H^5}{\varepsilon^2} + \frac{|\mathcal{X}|^4|\mathcal{A}|H^7}{\varepsilon}\Big)$$

exploration episodes.

(b) RF-UCRL (Kaufmann et al., 2021, Thm. 5) directly drives down an upper confidence bound on the estimation error of every Q-function for every reward, and is (ε,δ)*-PAC with*

$$\tilde{O}\Big(\frac{|\mathcal{X}||\mathcal{A}|H^4}{\varepsilon^2}\big(\log\frac{1}{\delta} + |\mathcal{X}|\big)\Big)$$

episodes; H^4 *improves to* H^3 *for stationary transitions.*

Fact 16.5 (Lower bound; Jin et al., 2020a)**.** *For every (ε, δ)-PAC reward-free algorithm there is an episodic MDP with stationary transitions on which the expected number of exploration episodes is $\Omega\big(\frac{|\mathcal{X}|^2|\mathcal{A}|H^2}{\varepsilon^2}\big)$ (Kaufmann et al., 2021, Sec. 5). In the discounted normalisation $H \asymp \frac{1}{1-\gamma}$ (and counting transitions) this reads*

$$\Omega\Big(\frac{|\mathcal{X}|^2|\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\Big),$$

a factor $|\mathcal{X}|$ above the single-reward rate [12.4], [12.5] (Stojanović, 2026, p. 19).

Remark 16.6 (Where the factor $|\mathcal{X}|$ comes from)**.** For a single reward, the comparison [12.2] runs along the fixed function $V^\star$: one scalar statistic $(\hat{P}-P)V^\star(z)$ per pair has to be accurate, which costs $O(\varepsilon^{-2})$ samples per pair up to horizon factors. For all rewards, $(\hat{P}-P)f(z)$ must be accurate for every f with range H, which is the total-variation accuracy of the row $\hat{P}(\cdot|z)$, a $|\mathcal{X}|$-dimensional statistic, and costs $\Theta(|\mathcal{X}|\varepsilon^{-2})$. The two bounds bracket the possibilities: a learner that knows the family of rewards in advance, or that knows a low-dimensional structure of P, sits in between. Part V asks which reward-free objects of dimension below $|\mathcal{X}|^2|\mathcal{A}|$ (successor features, low-rank factors, forward–backward embeddings, temporal distances) recover all rewards of interest.

16.4 The Running Examples

Example 16.7 (Coverage in the goal CMP)**.** In the goal CMP [3.8], every policy's trajectory from 1 stays in $\{1, 2, g\}$, and both actions at 1 and 2 have deterministic outcomes. A reward-free dataset containing each of the six pairs once determines P exactly, so $\mathsf{e}(\pi) = 0$ for all π and every reward is solved exactly in the planning phase [16.2]. If the dataset never contains the pair $(2, \mathsf{f})$, then for the reward $r = \mathbf{1}\{x = g\}$ the model cannot distinguish the true CMP from one in which $(2, \mathsf{f})$ leads back to 1, and $\mathsf{e}(\pi_r^\star) \geq 1$: the uncovered pair lies on the optimal path, and no planning with the data can repair this. Coverage, not the amount of data, is the binding constraint.

Assumed without proof in this chapter

[16.4] the online reward-free upper bounds; [16.5] the reward-free lower bound.

Part IV

Distributional Reinforcement Learning as Push-Forward

17

Push-Forward Reinforcement Learning

Following Akgül et al. (2026), a return law is the push-forward of the trajectory measure under the return functional, and its Bellman equation and contraction are the images of $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$ under that functional [6.12]. This chapter adds what the spine chapter does not contain: random rewards, the intermediate layer of the reward process, moments and the limits of statistical closure, the failure of distributional control, all Wasserstein orders, and Lipschitz kernels with the encoder–decoder bound of Akgül et al. (2026, Thm. 3.5). $\mathscr{X}$ and $\mathscr{A}$ are countable and π is stationary unless stated otherwise.

17.1 Random Rewards as Coordinates

Definition 17.1 (Reward kernels and the extended path space)**.** A *reward kernel* $\mathscr{R}(d\rho|z)$ is a Markov kernel from $\mathscr{Z}$ to a countable set $\mathbb{Y} \subseteq [0,1]$ with mean $r(z) := \int \rho\, \mathscr{R}(d\rho|z)$. The *extended pair space* is $\tilde{\mathscr{Z}} := \mathscr{Z} \times \mathbb{Y}$ with points $\tilde{z} = (x, a, \rho)$, and the *extended chain* of a stationary π has kernel

$$(\tilde{P} \otimes \pi)\big((x,a,\rho),(x',a',\rho')\big) := P(x'|x,a)\,\pi(a'|x')\,\mathscr{R}(\rho'|x',a').$$

Its path space $\tilde{\Omega} := \tilde{\mathscr{Z}}^{\mathbb{N}_0}$ carries the trajectory measures $\tilde{\mathbb{P}}^\pi_{\tilde{z}}$ [3.3], the discounted path metric d_γ [6.1] and the trajectory Bellman operator $\tilde{\Phi}^\pi$ [6.5]. All results of Chapter 6 hold verbatim on $\tilde{\Omega}$, since they used only that the pair space is countable.

Intuition. A random reward is one more coordinate of the trajectory. Adjoining it costs nothing in the theory: the trajectory measure and its fixed-point equation are unchanged in form, and the return becomes a deterministic functional of the extended path.

17.2 Three Layers of Push-Forward

Lemma 17.2 (Composition of intertwiners)**.** *Let $F : \Omega \to E$ intertwine with (β_z) [6.7] and $\Psi : E \to E'$ be measurable with $\Psi \circ \beta_z = \beta'_z \circ \Psi$ for maps*

$\beta'_z : E' \to E'$. *Then* $\Psi \circ F$ *intertwines with* (β'_z), *and*

$$\Psi_\# \circ \mathfrak{B}_F = \mathfrak{B}_{\Psi \circ F} \circ \Psi_\#$$

on families of laws. In particular $\mathfrak{Q}^{\Psi \circ F} = \Psi_\# \mathfrak{Q}^F$ *solves the Bellman equation of* $\Psi \circ F$ *because* $\mathfrak{Q}^F$ *solves that of* F.

Proof.

$$\Psi \circ F \circ \mathrm{cons}_z = \Psi \circ \beta_z \circ F = \beta'_z \circ \Psi \circ F.$$

For laws,

$$\Psi_\# (\beta_z)_\# \mathfrak{Q}_{z'} = (\beta'_z)_\# \Psi_\# \mathfrak{Q}_{z'},$$

and push-forward commutes with countable mixtures. The last claim is [6.8(b)] for $\Psi \circ F$, or equivalently the image of [6.8(b)] for F. ∎

Theorem 17.3 (The push-forward hierarchy; Akgül et al., 2026)**.** *On the extended path space* $\tilde{\Omega}$, *consider*

$$\tilde{\Omega} \xrightarrow{\mathsf{F}_R} \mathbb{Y}^{\mathbb{N}_0} \xrightarrow{\ell_\gamma} \left[0, \tfrac{1}{1-\gamma}\right] \xrightarrow{\mathit{mean}} \mathbb{R},$$

with the reward process $\mathsf{F}_R(\omega) := (\rho_t)_{t \geq 0}$ *and the* discounted readout $\ell_\gamma(y) := \sum_t \gamma^t y_t$, *so that the return is* $G = \ell_\gamma \circ \mathsf{F}_R$.

(a) F_R *intertwines with* $\beta^R_{\tilde{z}} := \mathrm{cons}_\rho$ *on* $\mathbb{Y}^{\mathbb{N}_0}$, *and*

$$\ell_\gamma \circ \mathrm{cons}_\rho = \beta_\rho \circ \ell_\gamma$$

with $\beta_\rho(g) := \rho + \gamma g$. *Hence* G *intertwines with* (β_ρ).

(b) The laws of the three layers, $\tilde{\mathbb{P}}^\pi_{\tilde{z}}$,

$$\mathsf{L}^\pi_{\tilde{z}} := (\mathsf{F}_R)_\# \tilde{\mathbb{P}}^\pi_{\tilde{z}}$$

and

$$\eta^\pi_{\tilde{z}} := G_\# \tilde{\mathbb{P}}^\pi_{\tilde{z}} = (\ell_\gamma)_\# \mathsf{L}^\pi_{\tilde{z}},$$

solve their Bellman equations

$$\tilde{\mathbb{P}}^\pi = \tilde{\Phi}^\pi \tilde{\mathbb{P}}^\pi, \qquad \mathsf{L}^\pi = \mathfrak{B}_{\mathsf{F}_R} \mathsf{L}^\pi, \qquad \eta^\pi = \mathscr{T}^\pi \eta^\pi,$$

where

$$(\mathscr{T}^\pi \eta)_{\tilde{z}} = \sum_{\tilde{z}'} (\tilde{P} \otimes \pi)(\tilde{z}, \tilde{z}') (\beta_\rho)_\# \eta_{\tilde{z}'}$$

is the distributional Bellman operator, and each equation is the push-forward of the one to its left.

(c) All three operators are γ-contractions in the transport costs of their discounted metrics, and taking means maps $\mathcal{T}^\pi$ to T^π; for G this is [6.9], for the reward process it follows in the same way from [6.8(c)] because prepending a symbol multiplies $\sum_t \gamma^t |y_t - y_t'|$ by γ.

Proof.

(a)

$$\mathsf{F}_R(\mathrm{cons}_z \omega) = (\rho, \rho_0, \rho_1, \dots) = \mathrm{cons}_\rho(\mathsf{F}_R \omega),$$

and

$$\ell_\gamma(\rho, y_0, y_1, \dots) = \rho + \gamma \ell_\gamma(y).$$

Apply [17.2].

6.6(a) and [17.2] twice.

■

Intuition. Read the diagram from left to right. The trajectory measure is the full object. Forgetting states and actions keeps the reward process; forgetting the order of rewards, up to discounting, keeps the return; forgetting everything but the average keeps the value. Each forgetful map commutes with prepending a step, so each layer inherits the Bellman equation and the contraction of the layer before. Distributional reinforcement learning is the second-to-last layer of this diagram, and it is no more and no less special than the others.

17.3 Moments and Closure

Theorem 17.4 (Moment Bellman equations; Sobel, 1982)**.** *For $m \in \mathbb{N}$ let*

$$\mathsf{M}_m(\omega) := (G(\omega), G(\omega)^2, \dots, G(\omega)^m) \in \mathbb{R}^m.$$

(a) M_m intertwines with the affine maps

$$\beta_\rho^{(m)}(u)_j := \sum_{i=0}^{j} \binom{j}{i} \rho^{j-i} \gamma^i u_i$$

($u_0 := 1$).

(b) The moment functions $\mu_j^\pi(\tilde{z}) := \mathbb{E}_{\tilde{z}}^\pi G^j$ satisfy

$$\mu_j^\pi(\tilde{z}) = \sum_{i=0}^{j} \binom{j}{i} \rho^{j-i} \gamma^i \sum_{\tilde{z}'} (\tilde{P} \otimes \pi)(\tilde{z}, \tilde{z}') \mu_i^\pi(\tilde{z}'),$$

a triangular system in which the j-th moment depends only on moments of order $\le j$.

(c) In particular the variance Σ^π of the return satisfies the variance Bellman equation [12.3(a)].

Proof. (a) $G \circ \mathrm{cons}_{\tilde{z}} = \rho + \gamma G$; expand $(\rho + \gamma G)^j$ binomially. (b) The maps $\beta_\rho^{(m)}$ are affine, so means commute with them: take expectations in [6.8(b)] for M_m (all moments are bounded by $(1-\gamma)^{-m}$). (c) $\Sigma^\pi = \mu_2^\pi - (\mu_1^\pi)^2$; substitute (b) for $j = 1, 2$ and simplify with the law of total variance as in the proof of [12.3(a)]. ∎

Fact 17.5 (Closed statistics; Rowland et al., 2019)**.** *Call a vector of statistics $\psi(\eta) = (\int f_1 \, d\eta, \dots, \int f_m \, d\eta)$ of return laws* Bellman closed *if $\psi(\mathcal{T}^\pi \eta)_{\tilde{z}}$ is a function of $(\psi(\eta_{\tilde{z}'}))_{\tilde{z}'}$ for every MDP. Then ψ is equivalent to the first n moments for some $n \le m$: each ψ_i is an affine function of $(\mu_1, \dots, \mu_n)$ and conversely (Bellemare et al., 2023, Thm. 8.12). In the language of this chapter: a finite-dimensional statistic of the return law admits its own dynamic programming exactly when the statistic, as a functional on path space, intertwines with an affine family, as M_m does.*

17.4 Distributional Control

Proposition 17.6 (Optimal return laws are not unique)**.** *For an MDP with finite $\mathcal{A}$ let $\Pi^\star$ be the set of stationary policies π with $V^\pi = V^\star$. The* optimal return laws *are the push-forwards $\{G_\# \tilde{\mathbb{P}}_{\tilde{z}}^\pi : \pi \in \Pi^\star\}$. They all have the same mean, but in general more than one law: there is no "optimal return law".*

Proof. Equal means is [6.9(d)] with $Q^\pi = Q^\star$ on the support of π. Example: one state, two actions with reward kernels $\delta_{1/2}$ and $\frac{1}{2}(\delta_0 + \delta_1)$, the state absorbing. Both actions are optimal, with return laws $\delta_{1/(2(1-\gamma))}$ and the law of $\sum_t \gamma^t \rho_t$ with i.i.d. fair-coin rewards, which is not a point mass. ∎

Fact 17.7 (The distributional optimality operator; Bellemare et al., 2017)**.** *Replace π in $\mathcal{T}^\pi$ by a policy that is greedy for the means of the current iterate. If the optimal policy is unique, the iterates converge to its return law, geometrically after a finite number of steps determined by the action gap (Bellemare et al., 2023, Thm. 7.9). If it is not unique, the iterates need not converge in any probability metric, although their means converge to $Q^\star$ (Bellemare et al., 2023, Thm. 7.13).*

Intuition. In the push-forward picture the reason is plain. Control selects a policy, hence a trajectory measure, and then pushes it forward. There is one optimal value but a set of optimal trajectory measures, and the return functional

separates them. A greedy rule that looks only at means cannot choose consistently within this set. A criterion that sees more of the law (a risk measure, Chapter 18) chooses within it but leaves the setting of dynamic programming.

17.5 Metrics on Path Space and Their Images

The contraction of the trajectory Bellman operator lives on path space; here it is extended to all Wasserstein orders, and model errors of Lipschitz kernels are propagated along trajectories.

Theorem 17.8 (Wasserstein contractions are images; Akgül et al., 2026, Lemma 3.1)**.** *Let $p \in [1, \infty)$.*

(a) The trajectory Bellman operator is a γ-contraction for $\bar{\mathsf{W}}_p$:

$$\bar{\mathsf{W}}_p(\Phi^\pi \mathbb{Q}, \Phi^\pi \mathbb{Q}') \le \gamma \bar{\mathsf{W}}_p(\mathbb{Q}, \mathbb{Q}').$$

(b) If $F : \Omega \to (E', d')$ is L-Lipschitz from d_γ, then

$$W_p(F_\# \mathbb{Q}_z, F_\# \mathbb{Q}'_z) \le L\, \mathsf{W}_p(\mathbb{Q}_z, \mathbb{Q}'_z);$$

in particular, for the return,

$$W_p(G_\# \mathbb{Q}_z, G_\# \mathbb{Q}'_z) \le \mathrm{sp}(r)\, \mathsf{W}_p(\mathbb{Q}_z, \mathbb{Q}'_z).$$

(c) The distributional Bellman operator is a γ-contraction for $\bar{W}_p$ on all families of laws on $[0, \frac{1}{1-\gamma}]$, and

$$\bar{W}_p((\mathcal{T}^\pi)^k \eta, \eta^\pi) \le \gamma^k \bar{W}_p(\eta, \eta^\pi).$$

Proof.

(a) As in [6.6(b)], with p-th powers of costs: prepending multiplies d_γ by γ, hence d_γ^p by γ^p, and mixtures of near-optimal couplings over the countably many next pairs bound the p-th power of the cost of the mixture by the mixture of p-th powers [6.2(b)].

(b) Push a near-optimal coupling forward by $F \times F$ [6.2(a)]; G is $\mathrm{sp}(r)$-Lipschitz [6.9(c)].

(c) $(\mathcal{T}^\pi \eta)_z$ is the mixture over z' of $(\beta_z)_\# \eta_{z'}$, and β_z multiplies distances by γ; the argument of (a) applies verbatim on $\mathbb{R}$. The rate follows from $\mathcal{T}^\pi \eta^\pi = \eta^\pi$ [6.9(a)].

■

Remark 17.9 (What the image metric adds and what it loses)**.**

(a) Prepending is an isometry for total variation on path space,

$$\|(\mathrm{cons}_z)_\# \mathbb{Q} - (\mathrm{cons}_z)_\# \mathbb{Q}'\|_{\mathrm{TV}} = \|\mathbb{Q} - \mathbb{Q}'\|_{\mathrm{TV}},$$

so Φ^π is not a contraction in total variation. The contraction of [6.6] is a property of the discounted metric, not of the measures.

(b) On the image $\mathbb{R}$, the Cramér-type distances $\ell_p(\eta, \eta') = (\int |F_\eta - F_{\eta'}|^p)^{1/p}$ between distribution functions make $\mathcal{T}^\pi$ a contraction with modulus $\gamma^{1/p}$ (Bellemare et al., 2023, Prop. 4.20). This uses the scaling of Lebesgue measure on $\mathbb{R}$ under $g \mapsto \gamma g$, a structure of the image space that has no counterpart on path space; it is the one contraction of distributional reinforcement learning that is not inherited from paths.

17.6 Lipschitz Kernels

Definition 17.10 (Lipschitz kernels; Akgül et al., 2026, Def. 3.2)**.** Let (E, d), (E', d') be countable metric spaces. A Markov kernel K from E to E' is *L-Lipschitz* (for W_p) if $W_p(K(x,\cdot), K(y,\cdot)) \le L\, d(x,y)$ for all x, y. A map $f : E \to E'$ is M-Lipschitz if $d'(f(x), f(y)) \le M d(x,y)$; it is the kernel $x \mapsto \delta_{f(x)}$.

Lemma 17.11 (Calculus of Lipschitz kernels; Akgül et al., 2026, Lemma 3.4 and Lemma A.1)**.** *Let K be L-Lipschitz from E to E' and K' be L'-Lipschitz from E' to E''.*

(a) For $\mu, \mu' \in \mathcal{P}(E)$, $W_p(\mu K, \mu' K) \le L\, W_p(\mu, \mu')$.

(b) KK' is LL'-Lipschitz; in particular, for an M-Lipschitz map f, the kernel $x \mapsto K'(f(x), \cdot)$ is ML'-Lipschitz.

(c) For $\mu \in \mathcal{P}(E)$ and kernels K_1, K_2,

$$W_p(\mu K_1, \mu K_2) \le \Big(\sum_x \mu(x) W_p(K_1(x,\cdot), K_2(x,\cdot))^p \Big)^{1/p} \le \sup_x W_p(K_1(x,\cdot), K_2(x,\cdot)).$$

Proof.

(a) Let ς couple μ, μ' with $\int d^p \, d\varsigma \le W_p(\mu, \mu')^p + \epsilon$, and for each of the countably many pairs (x, y) let $\kappa_{x,y}$ couple $K(x,\cdot), K(y,\cdot)$ with $\int d'^p \, d\kappa_{x,y} \le L^p d(x,y)^p + \epsilon$. Then $\sum_{x,y} \varsigma(x,y)\kappa_{x,y}$ couples $\mu K, \mu' K$ and its cost is

$$\le \sum \varsigma(x,y)\big(L^p d(x,y)^p + \epsilon\big) \le L^p (W_p(\mu, \mu')^p + \epsilon) + \epsilon.$$

The factor is L^p inside the p-th power, so L after the root. Let $\epsilon \to 0$.

(b)
$$\begin{aligned} W_p(KK'(x,\cdot), KK'(y,\cdot)) &= W_p(K(x,\cdot)K', K(y,\cdot)K') \\ &\le L' W_p(K(x,\cdot), K(y,\cdot)) \\ &\le L' L d(x,y) \end{aligned}$$

by (a) for K'.

(c) Mix near-optimal couplings of $K_1(x,\cdot), K_2(x,\cdot)$ with weights $\mu(x)$ [6.2(b)].

■

17.7 Compounding of Model Error

Theorem 17.12 (Model error along trajectories)**.** *Let (E,d) be a countable metric space, Q, Q' Markov kernels on E with Q' L-Lipschitz for W_1, $\epsilon_Q := \sup_x W_1(Q(x,\cdot), Q'(x,\cdot))$, and $r : E \to [0,1]$ L_r-Lipschitz. Let $\mathbb{P}_x^Q$, $\mathbb{P}_x^{Q'}$ be the laws of the chains from x [1.1] and $G(\omega) := \sum_t \gamma^t r(\omega_t)$.*

(a) There is a coupling of $\mathbb{P}_x^Q$ and $\mathbb{P}_x^{Q'}$, itself a Markov chain on $E \times E$, under which $\mathbb{E}\, d(X_t, X_t') \le \epsilon_Q \sum_{j<t} L^j$ for all t.

(b) If $\gamma L < 1$, then

$$\sup_x W_1\big(G_\# \mathbb{P}_x^Q, G_\# \mathbb{P}_x^{Q'}\big) \le \frac{\gamma\, L_r\, \epsilon_Q}{(1-\gamma)(1-\gamma L)}.$$

Proof.

(a) Fix $\epsilon > 0$. For each pair $(y,y') \in E^2$ choose a coupling $\kappa((y,y'),\cdot)$ of $Q(y,\cdot)$ and $Q'(y',\cdot)$ with cost $\le W_1(Q(y,\cdot), Q'(y',\cdot)) + \epsilon$. Then κ is a Markov kernel on the countable space E^2; let (X_t, X_t') be its chain from (x,x) [1.1]. The first coordinate is a Markov chain with kernel Q, because $\kappa((y,y'),\cdot)$ has first marginal $Q(y,\cdot)$ whatever y' is; likewise the second has kernel Q'. So the chain couples $\mathbb{P}_x^Q$ and $\mathbb{P}_x^{Q'}$ [1.2(a)]. By the Markov property [1.2(b)], the triangle inequality [6.4] and the Lipschitz property of Q',

$$\begin{aligned} \mathbb{E}[d(X_{t+1}, X_{t+1}') \mid X_t, X_t'] &\le W_1(Q(X_t,\cdot), Q'(X_t,\cdot)) \\ &\quad + W_1(Q'(X_t,\cdot), Q'(X_t',\cdot)) + \epsilon \\ &\le \epsilon_Q + \epsilon + L\, d(X_t, X_t'). \end{aligned}$$

Iterate from $d(X_0, X_0') = 0$: $\mathbb{E}d(X_t, X_t') \le (\epsilon_Q + \epsilon) \sum_{j<t} L^j$. Taking $\epsilon_t := \epsilon 2^{-t}$ at step t instead of ϵ and letting $\epsilon \to 0$ gives the claim in the limit along a sequence of couplings; for (b) this suffices.

(b)
$$|G(\omega) - G(\omega')| \le L_r \sum_t \gamma^t d(\omega_t, \omega_t'),$$

so by (a) the coupled returns differ in mean by at most

$$L_r(\epsilon_Q + \epsilon) \sum_t \gamma^t \sum_{j<t} L^j = L_r(\epsilon_Q + \epsilon) \sum_j L^j \frac{\gamma^{j+1}}{1-\gamma} = \frac{\gamma L_r(\epsilon_Q + \epsilon)}{(1-\gamma)(1-\gamma L)}.$$

This bounds W_1 of the return laws [6.2(a)].

■

Intuition. A model error of size ϵ_Q per step is injected at every step and then transported by the dynamics; a Lipschitz constant L amplifies distances by at most L per step, and discounting damps by γ. Errors compound geometrically at the rate γL. When $\gamma L < 1$ the compounded error stays bounded, otherwise this argument gives nothing.

Corollary 17.13 (Encoders and decoders; Akgül et al., 2026, Thm. 3.5)**.** *Let Q be an L_Q-Lipschitz kernel on (E, d), $\mathsf{S} : E \to E_{\text{lat}}$ an L_E-Lipschitz* encoder *into a countable latent metric space, and P_D an L_D-Lipschitz* decoder *kernel from E_{lat} to E. The* auto-encoded dynamics *is*

$$Q_{\text{ae}}(x, \cdot) := \sum_y Q(x, y)\, P_D(\mathsf{S}(y), \cdot).$$

Let

$$\epsilon_{\text{rec}} := \sup_y W_1(\delta_y, P_D(\mathsf{S}(y), \cdot))$$

be the reconstruction error. Then Q_{ae} is $L_Q L_E L_D$-Lipschitz,

$$\sup_x W_1(Q(x, \cdot), Q_{\text{ae}}(x, \cdot)) \le \epsilon_{\text{rec}},$$

and if $\gamma L_Q L_E L_D < 1$,

$$\sup_x W_1\big(G_\# \mathbb{P}_x^Q,\ G_\# \mathbb{P}_x^{Q_{\text{ae}}}\big) \le \frac{\gamma\, L_r\, \epsilon_{\text{rec}}}{(1-\gamma)(1-\gamma L_Q L_E L_D)}.$$

Proof. $Q_{\mathrm{ae}} = Q \cdot (P_D \circ \mathsf{S})$ is a composition of kernels with constants L_Q, L_E (a map) and L_D [17.11(b)]. With $K := P_D \circ \mathsf{S}$,

$$W_1(Q(x,\cdot), Q(x,\cdot)K) \le \sum_y Q(x,y) W_1(\delta_y, K(y,\cdot)) \le \epsilon_{\mathrm{rec}}$$

by [17.11(c)] with K_1 the identity kernel. Apply [17.12(b)] with $Q' := Q_{\mathrm{ae}}$ and $L := L_Q L_E L_D$. ∎

Remark 17.14 (The bottleneck trade-off, made quantitative)**.** Akgül et al. (2026) argue that a narrow information bottleneck (small L_E) speeds up the Bellman backups, while a decoder that must undo strong compression has a large L_D. In [17.13] the product $L_E L_D$ enters through the compounding rate $\gamma L_Q L_E L_D$, and the compression enters a second time through ϵ_{rec}, which grows when S merges states that the dynamics separate. The two effects pull in opposite directions; the bound is small only when the latent space merges states that are close in the metric *and* close in their dynamics.

17.8 The Running Examples

Example 17.15 (The hierarchy for the goal chain)**.** For the goal CMP [3.8] under π_{u} from state 1 with $\gamma = 0.9$ (deterministic rewards, $\mathbb{Y} = \{0, 1\}$):

(a) the trajectory measure is the law of the uniform walk on $\{1, 2, g\}$ absorbed at g;

(b) the reward process is $(0, \dots, 0, 1, 1, \dots)$ with the switch at τ_g, so its law is the law of τ_g, namely $\mathbb{P}\{\tau_g = t\} = F_{t-1}/2^t$ [6.14];

(c) the return is $\gamma^{\tau_g}/(1-\gamma)$, with mean 5.827, second moment

$$\frac{\gamma^4}{(4 - 2\gamma^2 - \gamma^4)(1-\gamma)^2} \approx 38.06,$$

and variance ≈ 4.10, which also solve the moment equations [17.4(b)];

(d) the value is the mean, $V^{\pi_{\mathrm{u}}}(1) \approx 5.83$.

Here the second layer loses nothing about rewards: a single switch time carries the whole reward process. In the gridworld [3.9] the same holds for the goal reward, while for a reward that depends on the path to the goal (for example a cost per visited boundary cell) the reward process carries strictly more than the return.

Example 17.16 (An encoder for the goal chain)**.** Put on $\{1, 2, g\}$ the path metric $d(1,2) = d(2,g) = 1$, $d(1,g) = 2$, and let $Q := P_{\pi_{\mathrm{u}}}$ [1.24]. The rows

differ by $W_1(Q(1), Q(2)) = \frac{1}{2}$, $W_1(Q(2), Q(g)) = 1$, $W_1(Q(1), Q(g)) = \frac{3}{2}$, so $L_Q = 1$; the goal reward has $L_r = 1$. The encoder that merges 1 and 2 into one latent state s (latent distance 1 to s_g) has $L_E = 1$, and the decoder $P_D(s) = \frac{1}{2}(\delta_1 + \delta_2)$, $P_D(s_g) = \delta_g$ has $L_D = \frac{3}{2}$ and $\epsilon_{\text{rec}} = \frac{1}{2}$. The compounding rate $\gamma L_Q L_E L_D = \frac{3}{2}\gamma$ is below 1 only for $\gamma < \frac{2}{3}$, and even then the bound of [17.13] exceeds the range of the return. The encoder is bad for a reason visible in [2.13]: it merges the two states whose asymmetric temporal relation to the goal ($\kappa(1, g) = 6, \kappa(2, g) = 4$) is what the return law records.

Assumed without proof in this chapter

[17.5] the characterisation of Bellman-closed statistics; [17.7] the behaviour of the distributional optimality operator; Minkowski's inequality in L^p; [17.9(b)] the Cramér-distance contraction.

18

Push-Forward Policy Iteration

An algorithm cannot store a law on $\mathbb{R}$, let alone a law on path space. It stores a finite-dimensional surrogate of the push-forward and updates it by a projected image of the trajectory Bellman operator. This chapter treats the surrogate families used in practice, categorical (Bellemare et al., 2017) and quantile (Dabney et al., 2018b,a), proves that projection preserves the contraction inherited from paths whenever it is nonexpansive, and casts the resulting algorithms as *push-forward policy iteration* (Akgül et al., 2026, Alg. 2): evaluate the push-forward by fitting a surrogate to its Bellman image, improve the policy by a functional of the fitted law. With the mean as functional this is policy iteration; with a risk functional it is risk-sensitive control, which leaves dynamic programming. The objects are families $\eta = (\eta_z)_{z\in\mathbb{Z}}$ of laws on $[0, \frac{1}{1-\gamma}]$.

18.1 Surrogates and Projections

Definition 18.1 (Surrogate families and projections)**.** Let F_η denote the distribution function and $F_\eta^{-1}(u) := \inf\{g : F_\eta(g) \geq u\}$ the quantile function of $\eta \in \mathcal{P}(\mathbb{R})$.

(a) The *categorical family* with support $\mathsf{g}_1 < \cdots < \mathsf{g}_m$ is

$$\mathfrak{F}_{\mathrm{C}} := \{\sum_i p_i \delta_{\mathsf{g}_i}\};$$

its *projection* Π_{C} moves the mass at $g \in [\mathsf{g}_i, \mathsf{g}_{i+1}]$ to the two neighbouring atoms in proportions that preserve the mean (mass at g outside $[\mathsf{g}_1, \mathsf{g}_m]$ goes to the nearest end) (Bellemare et al., 2017).

(b) The *quantile family* is $\mathfrak{F}_{\mathrm{Q}} := \{\frac{1}{m}\sum_i \delta_{\theta_i}\}$; its *projection* is

$$\Pi_{\mathrm{Q}}\eta := \frac{1}{m}\sum_i \delta_{F_\eta^{-1}(u_i)}$$

at the midpoints $u_i := \frac{2i-1}{2m}$ (Dabney et al., 2018b).

(c) Π acts on families coordinatewise, and the *projected push-forward operator* is $\Pi\mathcal{T}^\pi$.

(d) (Fact; Bellemare et al., 2023, Prop. 4.18.) For laws on $\mathbb{R}$,

$$W_p(\eta,\eta')^p = \int_0^1 |F_\eta^{-1}(u) - F_{\eta'}^{-1}(u)|^p \, du$$

for $p \in [1,\infty)$, the infimum over couplings being attained by $(F_\eta^{-1}(U), F_{\eta'}^{-1}(U))$ with U uniform on $(0,1)$.

$$W_\infty(\eta,\eta') := \sup_{u\in(0,1)} |F_\eta^{-1}(u) - F_{\eta'}^{-1}(u)|.$$

Lemma 18.2 (Projected contractions)**.** *Let $\bar d$ be the supremum over z of a metric d on laws, $\mathcal{T}^\pi$ a c-contraction for $\bar d$, $c<1$, and Π nonexpansive for d with values in a family closed under $\Pi\mathcal{T}^\pi$. Then $\Pi\mathcal{T}^\pi$ is a c-contraction, its fixed point $\hat\eta$ (if it exists) is unique, iterates converge to it at rate c^k, and*

$$\bar d(\hat\eta,\eta^\pi) \le \frac{\bar d(\Pi\eta^\pi,\eta^\pi)}{1-c}.$$

Proof.

$$\bar d(\Pi\mathcal{T}^\pi\eta,\Pi\mathcal{T}^\pi\eta') \le \bar d(\mathcal{T}^\pi\eta,\mathcal{T}^\pi\eta') \le c\,\bar d(\eta,\eta').$$

Uniqueness and the rate as in [0.8]. Finally

$$\bar d(\hat\eta,\eta^\pi) \le \bar d(\Pi\mathcal{T}^\pi\hat\eta,\Pi\mathcal{T}^\pi\eta^\pi) + \bar d(\Pi\eta^\pi,\eta^\pi) \le c\,\bar d(\hat\eta,\eta^\pi) + \bar d(\Pi\eta^\pi,\eta^\pi),$$

using $\mathcal{T}^\pi\eta^\pi = \eta^\pi$. ∎

Proposition 18.3 (The quantile projection; Dabney et al., 2018b)**.** *Π_Q is nonexpansive in W_∞. Hence $\Pi_Q\mathcal{T}^\pi$ is a γ-contraction in $\bar W_\infty$ with a unique fixed point $\hat\eta_Q$, and*

$$\bar W_\infty(\hat\eta_Q,\eta^\pi) \le \frac{1}{1-\gamma}\bar W_\infty(\Pi_Q\eta^\pi,\eta^\pi).$$

Proof. The quantile function of $\Pi_Q\eta$ is the step function equal to $F_\eta^{-1}(u_i)$ on $(\frac{i-1}{m},\frac{i}{m}]$. So

$$W_\infty(\Pi_Q\eta,\Pi_Q\eta') = \max_i |F_\eta^{-1}(u_i) - F_{\eta'}^{-1}(u_i)| \le W_\infty(\eta,\eta').$$

$\mathcal{T}^\pi$ is a γ-contraction in $\bar W_\infty$ by the argument of [17.8(c)] with the supremum norm of the coupled difference (the comonotone coupling of [18.1(d)] gives W_∞ on $\mathbb{R}$). The family $\mathfrak{F}_Q^{\mathcal{F}}$ is closed under $\Pi_Q\mathcal{T}^\pi$ by construction, and it is complete for $\bar W_\infty$ (its points are vectors of m atoms per pair, with the sup distance between sorted atoms). Existence follows from [0.8], and the rest from [18.2]. ∎

Fact 18.4 (The categorical projection)**.** *Π_{C} is the orthogonal projection onto $\mathfrak{F}_{\mathrm{C}}$ for the Cramér distance ℓ_2 (restricted to laws on $[\mathsf{g}_1, \mathsf{g}_m]$), hence nonexpansive, and $\Pi_{\mathrm{C}}\mathcal{T}^{\pi}$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$ (Bellemare et al., 2023, Lemma 5.23–5.24). This is the one place where the image-space contraction of [17.9(b)], not the path contraction, carries an algorithm.*

18.2 Learning the Push-Forward from Samples

Proposition 18.5 (Quantile regression)**.** *For $u \in (0,1)$ let $\mathsf{p}_u(v) := v\,(u - \mathbf{1}\{v < 0\})$ be the* check (pinball) loss*. For a law η on a bounded interval, the minimisers of $\vartheta \mapsto \int \mathsf{p}_u(g - \vartheta)\,\eta(dg)$ are the ϑ with $F_\eta(\vartheta^-) \le u \le F_\eta(\vartheta)$; in particular $F_\eta^{-1}(u)$ is a minimiser. Consequently, minimising*

$$\sum_i \int \varrho_{u_i}(g - \vartheta_i)\,(\mathcal{T}^{\pi}\eta)_z(dg)$$

over $\vartheta \in \mathbb{R}^m$ returns the atoms of $(\Pi_{\mathrm{Q}}\mathcal{T}^{\pi}\eta)_z$, and the loss can be estimated without bias from sampled transitions (z, ϱ, z') and draws $g' \sim \eta_{z'}$ via $g = \varrho + \gamma g'$.

Proof. The function

$$\mathscr{L}(\vartheta) := \int \mathsf{p}_u(g - \vartheta)\,\eta(dg)$$

is convex and finite, with right derivative $-u + F_\eta(\vartheta)$ and left derivative $-u + F_\eta(\vartheta^-)$ (differentiate $\mathsf{p}_u(g-\vartheta)$ in ϑ, and exchange derivative and integral by dominated convergence [0.2(c)]). A convex function is minimal exactly where the left derivative is $\le 0 \le$ the right derivative. The sample form is [6.9]: a draw from $(\mathcal{T}^{\pi}\eta)_z$ is $\varrho + \gamma g'$ with (ϱ, z') from the extended chain [17.1] and $g' \sim \eta_{z'}$. ∎

Definition 18.6 (Push-forward policy iteration; Akgül et al., 2026)**.** Fix a surrogate family $\mathfrak{F}$ and a *decision functional* $\phi : \mathcal{P}([0, \frac{1}{1-\gamma}]) \to \mathbb{R}$. *Push-forward policy iteration* (PPI) alternates

(a) *evaluation*: compute (or approximate) the fixed point $\eta_k \in \mathfrak{F}^{\mathcal{Z}}$ of $\Pi\mathcal{T}^{\pi_k}$, for instance by iterating the quantile regression of [18.5], or by minimising

$$\int_0^1 (F^{-1}_{\eta_z}(u) - F^{-1}_{(\mathcal{T}^{\pi_k}\eta)_z}(u))^2\,du = W_2(\eta_z, (\mathcal{T}^{\pi_k}\eta)_z)^2$$

[18.1(d)], as in Akgül et al. (2026, Alg. 2);

(b) *improvement*: $\pi_{k+1}(x) \in \arg\max_a \phi(\eta_{k,(x,a)})$.

Theorem 18.7 (PPI with the mean is policy iteration)**.** *Let* $\phi(\eta) := \int g\,\eta(dg)$.

(a) If evaluation is exact, $\eta_k = \eta^{\pi_k}$*, then PPI is policy iteration [4.7]:* $\phi(\eta_{k,z}) = Q^{\pi_k}(z)$*, so* $Q^\star - Q^{\pi_k} \to 0$ *at rate* γ^k*, and for finite* $\mathcal{X}$ *an optimal policy is reached after finitely many steps.*

(b) If evaluation is approximate with $\bar{W}_1(\eta_k, \eta^{\pi_k}) \le \epsilon$*, then*

$$\limsup_k \|V^\star - V^{\pi_k}\|_\infty \le \frac{2\gamma\epsilon}{(1-\gamma)^2}.$$

(c) With the quantile family, [18.3] controls the evaluation error only in $\bar{W}_\infty$*, through*

$$\frac{1}{1-\gamma}\bar{W}_\infty(\Pi_Q \eta^{\pi_k}, \eta^{\pi_k}),$$

and $W_1 \le W_\infty$ *[18.1(d)] makes (b) applicable with this* ϵ*.*

Proof. (a) The mean of $\eta_z^{\pi_k}$ is $Q^{\pi_k}(z)$ [6.9(d)], so the improvement step is greedy for Q^{π_k}; apply [4.7]. (b)

$$|\phi(\eta_{k,z}) - Q^{\pi_k}(z)| \le W_1(\eta_{k,z}, \eta_z^{\pi_k}) \le \epsilon$$

[6.2(d)], so PPI is approximate policy iteration with evaluation error ϵ; apply [9.7]. (c) $W_1 \le W_\infty$ by [18.1(d)], and [18.3]. For atomic laws $W_\infty(\Pi_Q \eta, \eta)$ is bounded below by half the largest gap between consecutive atoms carrying a quantile cell boundary, so it need not vanish as $m \to \infty$; the guarantee of (c) is therefore a statement about the surrogate family, not an approximation rate. ∎

18.3 Risk

Definition 18.8 (Risk functionals)**.** A *risk functional* is a map ϕ from laws on $[0, \frac{1}{1-\gamma}]$ to $\mathbb{R}$ (Bellemare et al., 2023, Def. 7.14). Examples: the mean; the mean–variance criterion $\int g\,d\eta - c\,\mathrm{Var}(\eta)$; the *conditional value-at-risk* at level $\alpha \in (0, 1]$,

$$\mathrm{CVaR}_\alpha(\eta) := \frac{1}{\alpha}\int_0^\alpha F_\eta^{-1}(u)\,du,$$

the mean of the worst α-fraction of returns.

Proposition 18.9 (Risk is Lipschitz in the push-forward)**.**

$$|\mathrm{CVaR}_\alpha(\eta) - \mathrm{CVaR}_\alpha(\eta')| \le \frac{1}{\alpha} W_1(\eta, \eta').$$

Consequently every error bound for return laws in W_1 (path-level contractions [17.8], model-error compounding [17.12], encoder–decoder bounds [17.13], projections [18.3]) transfers to CVaR_α *with the factor* $\frac{1}{\alpha}$.

Proof.

$$\frac{1}{\alpha}\int_0^\alpha |F_\eta^{-1} - F_{\eta'}^{-1}| \le \frac{1}{\alpha}\int_0^1 |F_\eta^{-1} - F_{\eta'}^{-1}| = \frac{1}{\alpha} W_1(\eta, \eta')$$

[18.1(d)]. ∎

Fact 18.10 (Risk-sensitive control leaves dynamic programming)**.** *For risk functionals other than the mean (and its affine transformations), maximising $\phi(\eta_\nu^\pi)$ over policies need not have a stationary Markov optimal policy, and the optimal policy may need to randomise (Bellemare et al., 2023, Sec. 7.7, Ex. 7.19); PPI with such a ϕ is then a heuristic, not policy iteration. For* CVaR_α*, optimality is restored by enlarging the state with the return accumulated so far (Bellemare et al., 2023, Sec. 7.8).*

Intuition. The push-forward picture explains the asymmetry. Evaluation is always a fixed-point problem, because the trajectory measure of a fixed policy is the fixed point of Φ^π and every statistic inherits it. Optimisation is a fixed-point problem only when the criterion is affine in the trajectory measure, because only then does choosing an action at the first step commute with taking the criterion of the rest. The mean is affine; risk is not. History-dependence is the price, and it is exactly the information that the path space carries and the state space forgets.

18.4 The Running Examples

Example 18.11 (Quantile PPI on the goal chain)**.** For the goal CMP [3.8] with $\gamma = 0.9$, the return law from state 1 under π_{u} is atomic with atoms $\gamma^t/(1-\gamma)$ and weights $F_{t-1}/2^t$ [17.15]: $\mathbb{P}\{\tau_g = 2\} = \frac{1}{4}$, $\mathbb{P}\{\tau_g \le 3\} = \frac{3}{8}$, $\mathbb{P}\{\tau_g \le 4\} = \frac{1}{2}$. With $m = 2$ quantile levels $u \in \{\frac{1}{4}, \frac{3}{4}\}$, $\Pi_{\mathrm{Q}}\eta_1^{\pi_{\mathrm{u}}}$ has the atoms $F^{-1}(\frac{1}{4}) = \gamma^8/(1-\gamma) \approx 4.30$ and $F^{-1}(\frac{3}{4}) = \gamma^3/(1-\gamma) = 7.29$ (the return decreases in τ_g;

$$\mathbb{P}\{\tau_g \ge 8\} \approx 0.27 \ge \frac{1}{4} > \mathbb{P}\{\tau_g \ge 9\}$$

and $\mathbb{P}\{\tau_g \geq 3\} = \frac{3}{4}$). The conditional value-at-risk $\mathrm{CVaR}_{1/4} \approx 2.94$ is about half the mean 5.83: the slowest quarter of trajectories wanders between 1 and 2 for at least eight steps. A CVaR-greedy improvement at state 1 favours f even more strongly than the mean-greedy step of [4.11], because b creates the long excursions that populate the lower tail.

Assumed without proof in this chapter

[18.1(d)] the quantile representation of W_p on $\mathbb{R}$; [18.4] the categorical projection; [18.10] the failure of dynamic programming for risk.

Part V

Successor Measures and Temporal Geometry

19

Successor Measures and Successor Features

Reward-free exploration [16.6] showed that a model accurate for all rewards costs a factor $|\mathcal{X}|$ more than a single value. The successor measure sits between the two: it is the second functional of the trajectory measure [6.10], it is reward-free, and it computes the value of every reward by one integral. This chapter develops successor measures as learning targets (Dayan, 1993; Blier et al., 2021), their compression to successor features for linear rewards, and generalised policy improvement (Barreto et al., 2017), which turns a library of successor features into a policy for a new task without further learning. $\mathcal{X}$, $\mathcal{A}$ are countable; π is stationary.

19.1 The Successor Measure as a Learning Target

Proposition 19.1 (Densities and temporal-difference targets; Blier et al., 2021)**.** *Let $\beta \in \mathcal{P}(\mathcal{X})$ with $\beta > 0$ be a reference law (for example the data distribution), and $m^\pi(z, z') := M^\pi(z, \{z'\})/\beta(z')$ the* successor density.

(a)

$$m^\pi(z, z') = \frac{\mathbf{1}\{z = z'\}}{\beta(z')} + \gamma \sum_{z''} (P \otimes \pi)(z, z'')\, m^\pi(z'', z').$$

(b) For every reward r,

$$Q^\pi(z) = \sum_{z'} \beta(z')\, m^\pi(z, z')\, r(z') = \mathbb{E}_{Z' \sim \beta}[m^\pi(z, Z') r(Z')] :$$

a value is a β-expectation of the successor density.

(c) The singular term $\mathbf{1}\{z = z'\}/\beta(z')$ is the image of the first step of the trajectory under the occupancy functional O; in continuous spaces it becomes a Dirac mass, and $M^\pi(z, \cdot)$ has no density with respect to β. Defining $\tilde{m}^\pi$ through $M^\pi - I = \gamma(P \otimes \pi)M^\pi$, which starts the count at $t = 1$, removes it.

Proof. (a) Divide the successor-measure Bellman equation [6.10(a)] at $\{z'\}$ by $\beta(z')$. (b) [6.10(b)]. (c) $M^\pi = I + \gamma(P \otimes \pi)M^\pi$, and $(P \otimes \pi)M^\pi$ is a mixture of laws of Z_t, $t \geq 1$, which in a continuous state space are absolutely continuous whenever P is. ∎

Intuition. The successor measure answers the question "starting here, how much discounted time will I spend there?" for all pairs of places at once. It is a model of the environment without trajectories: it forgets the order of visits, keeps their discounted frequency, and is still exactly what a value needs.

19.2 Successor Features

Proposition 19.2 (Successor features; Barreto et al., 2017)**.** *Let $\varphi : \mathcal{Z} \to \mathbb{R}^d$ be bounded. The* feature occupancy $\mathsf{O}_\varphi(\omega) := \sum_t \gamma^t \varphi(z_t)$ *intertwines with $\beta_z(u) := \varphi(z) + \gamma u$, and its mean*

$$\psi^\pi(z) := \mathbb{E}_z^\pi \mathsf{O}_\varphi = \int M^\pi(z, dz')\varphi(z'),$$

the successor features*, satisfies*

(a) $\psi^\pi = \varphi + \gamma(P \otimes \pi)\psi^\pi$, a γ-contraction in $\sup_z \|\cdot\|$;

(b) for every linear reward *$r_w(z) := \langle \varphi(z), w\rangle$, $w \in \mathbb{R}^d$, $Q_w^\pi = \langle \psi^\pi, w\rangle$.*

With indicator features $\varphi(z) = (\mathbf{1}\{z = z'\})_{z'}$, ψ^π is the successor measure.

Proof.

$$\mathsf{O}_\varphi \circ \mathrm{cons}_z = \varphi(z) + \gamma \mathsf{O}_\varphi.$$

Take means in [6.8(b)] (the map is affine); (b) by linearity of $w \mapsto r_w$ and [6.10(b)]. ∎

Theorem 19.3 (Generalised policy improvement; Barreto et al., 2017, Thm. 1)**.** *Let $\mathcal{A}$ be finite, $\pi_1, \dots, \pi_n$ stationary policies and $\tilde{Q}_i$ functions with $\|\tilde{Q}_i - Q^{\pi_i}\|_\infty \le \epsilon$, all for the same reward. Let $\pi(x) \in \arg\max_a \max_i \tilde{Q}_i(x, a)$. Then*

$$Q^\pi \ge \max_i Q^{\pi_i} - \frac{2\gamma\epsilon}{1-\gamma}.$$

Proof. Let $\bar{Q} := \max_i Q^{\pi_i}$ (pointwise). For every x',

$$\begin{aligned}
\bar{Q}(x', \pi(x')) &\ge \max_i \tilde{Q}_i(x', \pi(x')) - \epsilon \\
&= \max_a \max_i \tilde{Q}_i(x', a) - \epsilon \\
&\ge \max_a \bar{Q}(x', a) - 2\epsilon \\
&\ge \max_i V^{\pi_i}(x') - 2\epsilon,
\end{aligned}$$

since $V^{\pi_i}(x')$ is an average of $Q^{\pi_i}(x', \cdot)$ [4.3]. Hence

$$\begin{aligned} T^{\pi}\bar{Q}(z) &= r(z) + \gamma \int P(dx'|z)\bar{Q}(x', \pi(x')) \\ &\geq \max_i \left(r(z) + \gamma \int P(dx'|z) V^{\pi_i}(x') \right) - 2\gamma\epsilon \\ &= \bar{Q}(z) - 2\gamma\epsilon. \end{aligned}$$

By monotonicity and constant shift [4.2],

$$(T^{\pi})^k \bar{Q} \geq \bar{Q} - 2\gamma\epsilon \sum_{j<k} \gamma^j,$$

and $(T^{\pi})^k \bar{Q} \to Q^{\pi}$ [4.3]. ∎

Fact 19.4 (Transfer with successor features; Barreto et al., 2017, Thm. 2)**.** *Let π_j be optimal for r_{w_j}, $j \leq n$, and let π be the GPI policy [19.3] for the new reward r_w computed from $\tilde{Q}_j := \langle \tilde{\psi}^{\pi_j}, w \rangle$ with $|\langle \tilde{\psi}^{\pi_j} - \psi^{\pi_j}, w \rangle| \leq \epsilon$. Then*

$$Q^{\star}_w - Q^{\pi}_w \leq \frac{2}{1-\gamma}\left(\varphi_{\max} \min_j \|w - w_j\| + \epsilon\right),$$

$\varphi_{\max} := \sup_z \|\varphi(z)\|$.

Remark 19.5 (What successor features buy)**.** Successor features of n policies are nd numbers per pair, learned without rewards by the temporal-difference recursion [19.2(a)] (Chapters 6–7 apply verbatim). They solve every linear reward *for the policies in the library* exactly, and GPI improves on all of them at once; optimality for a new reward is guaranteed only up to its distance to the rewards the library was built for [19.4]. Removing that dependence requires the policies themselves to be indexed by rewards, which is the forward–backward construction of Chapter 22.

19.3 The Running Examples

Example 19.6 (GPI on the goal CMP)**.** In the goal CMP [3.8] with $\gamma = 0.9$ and the goal reward, take the library $\{\pi_{\mathrm{b}}, \pi_{\mathrm{u}}\}$ with exact values. At state 2, $Q^{\pi_{\mathrm{u}}}(2, \mathrm{f}) = \gamma \cdot 10 = 9$ exceeds

$$Q^{\pi_{\mathrm{u}}}(2, \mathrm{b}) = \gamma V^{\pi_{\mathrm{u}}}(1) \approx 5.24$$

and all values of π_{b} (which are 0 off g); at state 1,

$$Q^{\pi_{\mathrm{u}}}(1, \mathrm{f}) = \gamma V^{\pi_{\mathrm{u}}}(2) \approx 6.41 > Q^{\pi_{\mathrm{u}}}(1, \mathrm{b}) \approx 5.24.$$

So the GPI policy is π_{f}, optimal after one step, with successor measures from [6.14] as the only learned objects. For the reward $\mathbf{1}\{x = 2\}$, the library values are

$$Q^{\pi_{\mathsf{u}}}(1, \mathsf{f}) \approx 1.43 > Q^{\pi_{\mathsf{u}}}(1, \mathsf{b}) \approx 1.17$$

and $Q^{\pi_{\mathsf{u}}}(2, \mathsf{b}) \approx 2.17$, above every other library value at 2; GPI therefore alternates $1 \xrightarrow{\mathsf{f}} 2 \xrightarrow{\mathsf{b}} 1$, which is optimal for this reward (reaching g ends all reward) and coincides with neither library policy.

Assumed without proof in this chapter

[19.4] the transfer bound for successor features.

20

Spectral Theory of the Successor Representation

For a reversible chain the successor measure is a function of the transition operator in the sense of spectral calculus: it has the eigenvectors of P_π and the eigenvalues $1/(1-\gamma\chi)$. This single fact connects four lines of work: proto-value functions, which use the slow eigenvectors of a graph Laplacian as a basis for values (Mahadevan and Maggioni, 2007); Laplacian representations learned by minimising a Dirichlet energy (Wu et al., 2019); eigenoptions, which turn eigenvectors of the successor representation into temporally extended behaviour (Machado et al., 2017, 2023); and the spectral formulas for hitting times, which express the temporal geometry of Chapter 23 through the deviation matrix. $\mathcal{X}$ is finite, π stationary, $P := P_\pi$ irreducible with invariant law μ [1.14], and $\langle f, f' \rangle_\mu := \sum_x \mu(x) f(x) f'(x)$.

20.1 Hitting Times from the Deviation Matrix

Theorem 20.1 (Hitting times are displacements of the deviation matrix)**.** *Let D be the deviation matrix of P [2.9]. For all $y, g \in \mathcal{X}$,*

$$\kappa(y,g) = \frac{D(g,g) - D(y,g)}{\mu(g)}.$$

Hence $\kappa(y,g) = \langle e_g - e_y, \omega_g \rangle$ with $\omega_g := D(\cdot,g)/\mu(g)$: every irreducible chain has an exact displacement representation of its mean hitting times with one-hot features (compare the displacement theorem of the hitting-time chapter), and $\sum_y \mu(y)\kappa(y,g) = D(g,g)/\mu(g)$.

Proof. Let $h := \kappa(\cdot, g)$, finite by [2.4(c)]. For $y \neq g$, $(I-P)h(y) = 1$ [2.3(b)]. At g, $h(g) = 0$ and $Ph(g) = \mathbb{E}_g \tau_g^+ - 1 = \frac{1}{\mu(g)} - 1$ by the Markov property and [1.14], so $(I-P)h(g) = 1 - \frac{1}{\mu(g)}$. Thus $(I-P)h = \mathbf{1} - \mathbf{1}_{\{g\}}/\mu(g)$. The right side has μ-mean 0, so it equals $(I-P^*)$ of itself, and by [2.10(a)] the function $-D\mathbf{1}_{\{g\}}/\mu(g)$ solves the same equation (note $D\mathbf{1} = Z\mathbf{1} - P^*\mathbf{1} = 0$). Solutions of $(I-P)v = f$ differ by P-harmonic functions, which are constant for irreducible P (as in the proof of [2.11]). Hence $h = c - D(\cdot,g)/\mu(g)$, and $h(g) = 0$ fixes $c = D(g,g)/\mu(g)$. Finally $\mu D = 0$, because every row of P^* is μ [2.8(d)] and $P^* D = 0$ [2.10(a)]. ∎

20.2 Reversible Chains

Theorem 20.2 (Spectral calculus of the successor representation)**.** *Let P be reversible with respect to μ [1.17]. Then:*

(a) P is self-adjoint in $\langle\cdot,\cdot\rangle_\mu$; there is a μ-orthonormal eigenbasis $f_1 = \mathbf{1}, f_2, \dots, f_n$ with real eigenvalues $1 = \chi_1 \geq \chi_2 \geq \dots \geq \chi_n \geq -1$, and $P^t(x,y) = \mu(y)\sum_i \chi_i^t f_i(x) f_i(y)$.

(b) The state successor measure is

$$M_\gamma(x,y) = \mu(y)\sum_{i=1}^{n} \frac{f_i(x)f_i(y)}{1-\gamma\chi_i},$$

so its eigenvectors (as an operator on $L^2(\mu)$) are those of P, its eigenvalues are $\frac{1}{1-\gamma\chi_i}$, and the order of the eigenvalues is preserved.

(c) The deviation matrix is

$$D(x,y) = \mu(y)\sum_{i\geq 2} \frac{f_i(x)f_i(y)}{1-\chi_i},$$

and $\mathbb{E}_\mu \tau_g = \sum_{i\geq 2} \frac{f_i(g)^2}{1-\chi_i}$; Kemeny's constant [2.11] is $\bar{\kappa} = \sum_{i\geq 2} \frac{1}{1-\chi_i}$.

Proof.

(a) Detailed balance gives

$$\langle f, Pf'\rangle_\mu = \sum_{x,y} \mu(x)P(x,y)f(x)f'(y) = \langle Pf, f'\rangle_\mu.$$

The finite-dimensional spectral theorem for self-adjoint operators gives the basis; $|\chi_i| \leq 1$ since P is a contraction in $\|\cdot\|_\infty$; $P\mathbf{1} = \mathbf{1}$. Expanding $\delta_y/\mu(y)$ in the basis gives the formula for P^t.

(b) Sum (a) against γ^t [2.6].

(c) $P^*(x,y) = \mu(y)$ [2.8(d)], so $Z = (I - P + P^*)^{-1}$ has eigenvalue 1 on f_1 and $\frac{1}{1-\chi_i}$ on f_i, $i \geq 2$ (irreducibility gives $\chi_i < 1$ for $i \geq 2$); subtract P^*. The hitting formula is [20.1] with $\mu D = 0$, and

$$\bar{\kappa} = \sum_g \mu(g)\mathbb{E}_\mu \tau_g = \sum_{i\geq 2} \frac{\|f_i\|_\mu^2}{1-\chi_i}.$$

■

Intuition. The successor representation amplifies the slow modes of the chain: an eigenvalue close to 1 becomes an eigenvalue close to $\frac{1}{1-\gamma}$, and a fast mode becomes almost 1. The top eigenvectors of the successor representation are therefore the directions in which the chain mixes slowly, the bottlenecks and the far-apart regions, and the same slow modes dominate the hitting times.

Corollary 20.3 (The successor representation and the normalised Laplacian; Machado et al., 2023, Thm. 2)**.** *Let $P = \mathsf{D}^{-1}\mathsf{W}$ be the random walk on a connected graph with symmetric weights W and degrees $\mathsf{D} = \operatorname{diag}(\mathsf{W}\mathbf{1})$, and $\mathscr{L} := I - \mathsf{D}^{-1/2}\mathsf{W}\mathsf{D}^{-1/2}$ its normalised Laplacian. Then $\mathscr{L}e = \nu e$ iff $f := \mathsf{D}^{-1/2}e$ satisfies $Pf = (1-\nu)f$, and f is an eigenvector of the successor representation with eigenvalue $\chi_{\mathrm{SR}} = \frac{1}{1-\gamma(1-\nu)}$, i.e. $\nu = 1 - \frac{1-\chi_{\mathrm{SR}}^{-1}}{\gamma}$. Small Laplacian eigenvalues correspond to large SR eigenvalues.*

Proof. $\mathsf{D}^{-1/2}\mathscr{L}\mathsf{D}^{1/2} = I - P$, so $\mathscr{L}e = \nu e$ iff $(I-P)\mathsf{D}^{-1/2}e = \nu \mathsf{D}^{-1/2}e$. The random walk is reversible with $\mu \propto \mathsf{D}\mathbf{1}$; apply [20.2(b)]. ∎

20.3 Variational Characterisations and Representations

Proposition 20.4 (Dirichlet form and the Laplacian objective)**.** *For reversible P and $f : \mathcal{X} \to \mathbb{R}$,*

$$\begin{aligned}\langle f, (I-P)f\rangle_\mu &= \tfrac{1}{2}\sum_{x,y}\mu(x)P(x,y)\big(f(x)-f(y)\big)^2 \\ &= \tfrac{1}{2}\mathbb{E}\big[(f(X_0)-f(X_1))^2\big], \quad X_0 \sim \mu.\end{aligned}$$

Consequently

$$\min\{\langle f,(I-P)f\rangle_\mu : \langle f,\mathbf{1}\rangle_\mu = 0,\ \|f\|_\mu = 1\} = 1-\chi_2,$$

attained at f_2. The objective of Wu et al. (2019), which minimises $\mathbb{E}\|\phi(X_0) - \phi(X_1)\|^2$ over $\phi : \mathcal{X} \to \mathbb{R}^k$ with μ-orthonormal coordinates orthogonal to $\mathbf{1}$, is the trace of the Dirichlet form, and its minimum $\sum_{i=2}^{k+1}(1-\chi_i)$ is attained by $(f_2,\dots,f_{k+1})$ (Ky Fan's principle, a Fact).

Proof. Expand the square and use $\mu P = \mu$ and detailed balance:

$$\sum \mu(x)P(x,y)f(x)^2 = \sum \mu(x)P(x,y)f(y)^2 = \|f\|_\mu^2,$$

and the cross term is $-2\langle f, Pf\rangle_\mu$. For the minimum, write $f = \sum_{i\ge 2} c_i f_i$ with $\sum c_i^2 = 1$:

$$\langle f,(I-P)f\rangle_\mu = \sum_i c_i^2(1-\chi_i) \ge 1-\chi_2.$$

■

Theorem 20.5 (Value approximation with slow eigenvectors; Mahadevan and Maggioni, 2007)**.** *For reversible P, let Π_k be the μ-orthogonal projection onto* $\operatorname{span}(f_1, \dots, f_k)$ *and $V = M_{\mathfrak{X}} r$ the value of a reward r on $\mathfrak{X}$. Then*

$$\|V - \Pi_k V\|_\mu \le \frac{\|(I - \Pi_k) r\|_\mu}{1 - \gamma \chi_{k+1}},$$

$$\left\| M_{\mathfrak{X}} - I - \sum_{i \le k} \frac{\gamma \chi_i}{1 - \gamma \chi_i} f_i \langle f_i, \cdot \rangle_\mu \right\|_{\mu \to \mu} = \max_{i > k} \frac{\gamma |\chi_i|}{1 - \gamma \chi_i}.$$

Rewards that vary slowly along the dynamics (small high-frequency components) have values that are well represented by a few slow eigenvectors, and the shifted successor measure is low rank up to an error governed by the $(k+1)$-st eigenvalue (compare (Chapter on low-rank structure, Part V)).

Proof. By [20.2(b)], $V = \sum_i \frac{\langle r, f_i \rangle_\mu}{1 - \gamma \chi_i} f_i$, so

$$\|V - \Pi_k V\|_\mu^2 = \sum_{i > k} \frac{\langle r, f_i \rangle_\mu^2}{(1 - \gamma \chi_i)^2};$$

for $i > k, \chi_i \le \chi_{k+1}$ gives $1 - \gamma \chi_i \ge 1 - \gamma \chi_{k+1} > 0$. For the second claim, $M_{\mathfrak{X}} - I$ has eigenvalues

$$\frac{1}{1 - \gamma \chi_i} - 1 = \frac{\gamma \chi_i}{1 - \gamma \chi_i}$$

on f_i, and the remainder after removing the first k is a self-adjoint operator with the stated largest eigenvalue modulus. ■

20.4 Eigenoptions

Definition 20.6 (Eigenoptions; Machado et al., 2017)**.** For an eigenvector e of the successor representation (or of the Laplacian), the *eigenpurpose* is the intrinsic reward $r_e(x, a, x') := e(x') - e(x)$. The *eigenoption* for e follows an optimal policy for r_e and terminates at the states x with $\max_a Q_e^\star(x, a) \le 0$, where $Q_e^\star$ is the optimal action value for r_e.

Fact 20.7 (Eigenoptions terminate; Machado et al., 2023, Thm. 1)**.** *In a finite ergodic MDP with $\gamma < 1$, every eigenoption has at least one state where it terminates.*

Intuition. The eigenpurpose rewards moving along the eigenvector, so an eigenoption travels to the extreme of a slow mode: through a bottleneck, or to a far corner. Options built from the slowest modes are exactly the behaviours that reduce the large hitting times of [20.2(c)], which is why they help exploration. Chapter 26 places options in the general theory of temporal abstraction.

20.5 The Running Examples

Example 20.8 (Spectrum of the gridworld)**.** The uniform walk on the 5×5 grid [1.25] is reversible with uniform μ. Its Kemeny constant $\bar{\kappa} = \sum_{i \geq 2} \frac{1}{1-\chi_i} \approx 54.08$ agrees with the value found from hitting times in [2.13], as [20.2(c)] requires. The slowest non-constant eigenvectors vary linearly along the rows or columns, so the corresponding eigenoptions travel from one side of the grid to the opposite side.

Assumed without proof in this chapter

The finite-dimensional spectral theorem; Ky Fan's principle in [20.4]; [20.7] termination of eigenoptions.

21

Low-Rank Structure and Entry-Wise Estimation

A successor measure on $\mathcal{Z}$ is an $|\mathcal{Z}| \times |\mathcal{Z}|$ object, and learning it in total variation for all rewards costs what reward-free exploration costs [16.5]. Low-rank structure of the dynamics is the simplest assumption under which this cost drops: one state dimension disappears from the sample complexity (Stojanović, 2026). This chapter shows that a low-rank transition kernel makes the successor measure the identity plus a low-rank term, explains why control needs entry-wise rather than spectral or Frobenius accuracy, and states the entry-wise estimation results on which low-rank reinforcement learning rests. $\mathcal{Z}$ is finite with $|\mathcal{Z}| = n_Z$; matrices are indexed by $\mathcal{Z}$.

21.1 Low-Rank Dynamics

Theorem 21.1 (Low-rank kernels give identity-plus-low-rank successor measures)**.** *Suppose the state-action kernel factorises as $P \otimes \pi = \Phi\Psi^\top$ with $\Phi, \Psi \in \mathbb{R}^{n_Z \times d}$. Then*

$$M^\pi = I + \gamma\, \Phi\, (I_d - \gamma \Psi^\top \Phi)^{-1} \Psi^\top,$$

so the shifted successor measure *$M^\pi - I = \gamma(P \otimes \pi) M^\pi$ has rank at most d, and every value is*

$$Q_r^\pi = r + \gamma \Phi\, (I_d - \gamma \Psi^\top \Phi)^{-1} \Psi^\top r :$$

a d-dimensional summary $\Psi^\top r$ of the reward suffices.

Proof. $(\Phi\Psi^\top)^t = \Phi(\Psi^\top\Phi)^{t-1}\Psi^\top$ for $t \geq 1$. The nonzero eigenvalues of $\Psi^\top\Phi$ are those of $\Phi\Psi^\top = P \otimes \pi$, whose spectral radius is at most 1 [0.9(c)], so $\varrho(\gamma\Psi^\top\Phi) \leq \gamma < 1$ and

$$\sum_{s \geq 0} (\gamma \Psi^\top \Phi)^s = (I_d - \gamma \Psi^\top \Phi)^{-1}$$

[0.9(a)]. Then

$$M^\pi = I + \sum_{t \geq 1} \gamma^t \Phi (\Psi^\top \Phi)^{t-1} \Psi^\top = I + \gamma \Phi \Big(\sum_s (\gamma \Psi^\top \Phi)^s \Big) \Psi^\top$$

[6.10(a)]. The value formula is [6.10(b)]. ∎

Intuition. The identity term is the first step of the trajectory, which is never low-rank; the rest of the trajectory passes through the d-dimensional bottleneck of the kernel at every step and inherits its rank. This is why one should learn the shifted successor measure, not the successor measure itself.

21.2 Which Error Matters for Control

Definition 21.2 (Matrix norms)**.** For $A \in \mathbb{R}^{n\times m}$: $\|A\|_{\max} := \max_{i,j} |A_{ij}|$; $\|A\|_{2\to\infty} := \max_i \|A_{i\cdot}\|_2$ (largest row norm); $\|A\|_2$ the operator norm; $\|A\|_F$ the Frobenius norm. For a rank-d $A = U\,\mathrm{diag}(\sigma)V^\top$ (compact singular value decomposition), A is *μ-incoherent* if $\|U\|_{2\to\infty} \le \sqrt{\mu d/n}$ and $\|V\|_{2\to\infty} \le \sqrt{\mu d/m}$.

Proposition 21.3 (Entry-wise error and control)**.**

(a)

$$\|A\|_{\max} \le \|A\|_{2\to\infty} \le \|A\|_2 \le \|A\|_F \le \sqrt{nm}\,\|A\|_{\max},$$

and each inequality can be an equality.

(b) (Action-gap certificate.) If

$$\|\hat{Q} - Q^\star\|_{\max} < \frac{1}{2}\big(Q^\star(x,\pi^\star(x)) - \max_{a\neq\pi^\star(x)} Q^\star(x,a)\big)$$

at a state x, then $\pi_{\hat{Q}}(x) = \pi^\star(x)$; and in general

$$\|V^\star - V^{\pi_{\hat{Q}}}\|_\infty \le \frac{2\|\hat{Q} - Q^\star\|_{\max}}{1-\gamma}$$

[4.6(b)].

(c) (Delocalisation.) If A is rank-d and μ-incoherent, then $\|A\|_{\max} \le \frac{\mu d}{\sqrt{nm}}\|A\|_2$.

Proof.

(a) An entry is bounded by its row norm, a row norm by $\|A^\top e_i\|_2 \le \|A\|_2$, and $\|A\|_2^2 \le \sum \sigma_i^2 = \|A\|_F^2$; the last is the count of entries. Equality: a single nonzero entry gives equality in the first three, a constant matrix in the last.

(b) If every entry moves by less than half the gap at x, the order between the optimal and every other action at x is preserved. The second claim is [4.6(b)] with the sup norm equal to $\|\cdot\|_{\max}$.

(c)

$$|A_{ij}| = |U_{i\cdot}\,\mathrm{diag}(\sigma)V_{j\cdot}^\top| \le \|U_{i\cdot}\|_2\,\sigma_1\,\|V_{j\cdot}\|_2 \le \sqrt{\mu d/n}\,\|A\|_2\sqrt{\mu d/m}.$$

■

Intuition. Control is decided entry by entry, at the pairs where actions compete. A Frobenius error budget can be spent on one entry, where it flips an action, or spread over all entries, where it does no harm. Incoherence is the assumption that the matrix itself spreads its information, so that estimation errors spread too.

Fact 21.4 (Entry-wise estimation of low-rank objects; Stojanović, 2026)**.**

(a) The classical Davis–Kahan bound controls singular subspaces in operator norm and does not, by itself, give entry-wise or two-to-infinity control; leave-one-out arguments for incoherent low-rank matrices give two-to-infinity subspace bounds and hence entry-wise estimation guarantees from noisy or sampled entries (Papers A, B).

(b) Under low-rank transition kernels, the reward-free sample complexity improves from order $|\mathcal{X}|^2|\mathcal{A}|$ to $d|\mathcal{X}||\mathcal{A}|$, and under low-rank action-value structure the model-free complexity improves from $|\mathcal{X}||\mathcal{A}|$ to $d(|\mathcal{X}|+|\mathcal{A}|)$, up to horizon and logarithmic factors; low-rank structure removes one state dimension, not all of them (thesis, concluding chapter).

21.3 The Running Examples

Example 21.5 (A truncation of $Q^\star$)**.** For the goal CMP with $\gamma = 0.9$, $Q^\star$ [4.11] as a 3×2 matrix has singular values ≈ 21.26 and 0.94. Its best rank-1 approximation $\hat{Q}$ has relative Frobenius error $\approx 4.4\%$ and entry-wise error $\|\hat{Q} - Q^\star\|_{\max} \approx 0.51$. The action gaps of $Q^\star$ are 0.81 at state 1 and 1.71 at state 2. The certificate [21.3(b)] guarantees the greedy action only at state 2 ($0.51 < 0.855$); at state 1 the truncated matrix happens to keep f, but nothing in the Frobenius error rules out a flip. The kernel P_{π_u} has full rank 3 and $M_{\mathcal{X}}^{\pi_u} - I$ has singular values $\approx (13.0, 1.95, 0.21)$: low rank holds only approximately, and its quality must be judged in the norm that decides actions.

Assumed without proof in this chapter

[21.4] the entry-wise estimation results and the low-rank sample complexities.

22

Forward–Backward Representations and Zero-Shot RL

Successor features solve linear rewards for a fixed library of policies [19.5]. A *zero-shot* agent must instead return an optimal policy for any reward revealed after training, with no further learning or planning. The forward–backward (FB) representation of Touati and Ollivier (2021) achieves this by indexing policies by reward embeddings and factorising their successor measures: the successor measure of the policy indexed by w is $F(\cdot, w)^\top B(\cdot)$ times a reference law. This chapter proves that an exact FB representation yields optimal policies for all rewards, that one exists in finite spaces, and derives its temporal-difference equation from the successor-measure Bellman equation, which is itself the second functional of the trajectory measure [6.10]. $\mathcal{Z}$ is finite, $\mathcal{A}$ finite, and $\beta \in \mathcal{P}(\mathcal{Z})$ is a reference law with $\beta > 0$.

22.1 Exact FB Representations

Definition 22.1 (Forward–backward representation; Touati and Ollivier, 2021)**.** A triple $(F, B, (\pi_w)_{w \in \mathbb{R}^d})$ with $F : \mathcal{Z} \times \mathbb{R}^d \to \mathbb{R}^d$, $B : \mathcal{Z} \to \mathbb{R}^d$ and stationary deterministic policies π_w is an *FB representation* with respect to β if, for all $w \in \mathbb{R}^d$ and $z, z' \in \mathcal{Z}$,

$$\pi_w(x) \in \arg\max_a F(x, a, w)^\top w, \qquad M^{\pi_w}(z, \{z'\}) = F(z, w)^\top B(z')\, \beta(z').$$

For a reward r, its *embedding* is $w_r := \sum_z \beta(z) r(z) B(z) = \mathbb{E}_{Z \sim \beta}[r(Z) B(Z)]$.

Theorem 22.2 (FB encodes all optimal policies; Touati and Ollivier, 2021, Thm. 2)**.** *Let $(F, B, (\pi_w))$ be an FB representation. For every reward r, the policy π_{w_r} is optimal for r, and $Q_r^\star(z) = F(z, w_r)^\top w_r$.*

Proof. For any w, by [6.10(b)] and the factorisation,

$$Q_r^{\pi_w}(z) = \sum_{z'} M^{\pi_w}(z, \{z'\}) r(z') = F(z, w)^\top \sum_{z'} B(z') \beta(z') r(z') = F(z, w)^\top w_r.$$

At $w = w_r$, π_{w_r} is greedy for $F(\cdot, w_r)^\top w_r = Q_r^{\pi_{w_r}}$. A policy that is greedy for its own action values satisfies $T^\star Q^\pi = T^\pi Q^\pi = Q^\pi$, so $Q^\pi = Q^\star$ by uniqueness of the fixed point of $T^\star$ [4.5]. ∎

Proposition 22.3 (Existence in finite spaces; Touati and Ollivier, 2021, Prop. 6)**.** *For $d = |\mathcal{Z}|$ an FB representation exists: index $\mathbb{R}^d$ by $\mathcal{Z}$, put $B(z') := e_{z'}$, let π_w be an optimal stationary deterministic policy for the reward $r_w(z') := w_{z'}/\beta(z')$ [4.5], and $F(z,w)_{z'} := M^{\pi_w}(z,\{z'\})/\beta(z')$.*

Proof. $F(z,w)^\top B(z')\beta(z') = M^{\pi_w}(z,\{z'\})$ by construction. And

$$F(x,a,w)^\top w = \sum_{z'} M^{\pi_w}(x,a,\{z'\}) r_w(z') = Q^{\pi_w}_{r_w}(x,a) = Q^\star_{r_w}(x,a)$$

[6.10(b)], whose argmax contains $\pi_w(x)$. ∎

Intuition. B embeds a place by how it can be reached, F embeds a starting pair (and a task) by where it leads; their inner product is the discounted time spent going from one to the other. A reward is turned into a task embedding by averaging B over where the reward is. Optimality is a consistency condition: the policy indexed by a task must be greedy for the values that its own successor measure assigns to that task.

22.2 Learning and Approximation

Proposition 22.4 (The FB temporal-difference equation)**.** *$(F,B,(\pi_w))$ with greedy π_w is an FB representation iff, for all w, z, z',*

$$F(z,w)^\top B(z') = \frac{\mathbf{1}\{z = z'\}}{\beta(z')} + \gamma \sum_{x''} P(x''|z)\, F\big((x'', \pi_w(x'')), w\big)^\top B(z').$$

Proof. This is the successor-density equation [19.1(a)] for π_w, whose unique solution is m^{π_w} (it is a γ-contraction in sup-norm after multiplying by $\beta(z')$ [6.10(d)]). ∎

Fact 22.5 (Approximate FB; Touati and Ollivier, 2021)**.** *If for a reward r the error $F(z_0,w_r)^\top B(z) - M^{\pi_{w_r}}(z_0,\{z\})/\beta(z)$ is at most ε on average over $z \sim \beta$, for every z_0, then π_{w_r} is $\frac{3\varepsilon\|r\|_\infty}{1-\gamma}$-optimal for r. In infinite spaces exact FB requires $d = \infty$, and finite d gives a rank-d approximation of the successor measures.*

Remark 22.6 (Rank, identity and coverage)**.**

(a) $M^{\pi_w} = I + \gamma(P \otimes \pi_w)M^{\pi_w}$ always contains the identity, which has full rank; a finite-d FB representation can be exact only through its shifted part [21.1] plus a separate treatment of the first step (rewards at the current pair are known at test time).

(b) The FB object is reward-free, so the reward-free lower bound [16.5] constrains how much data its training needs on worst-case MDPs; low-rank dynamics [21.4] are the structural assumption under which FB with small d can beat it.

(c) In the hierarchy of [6.12], FB keeps the second functional (discounted occupancy means) of a whole family of trajectory measures $\{\mathbb{P}^{\pi_w}\}_w$, and nothing of their return laws or hitting times.

22.3 The Running Examples

Example 22.7 (Zero-shot on the goal CMP)**.** For the goal CMP [3.8] with β uniform on the six pairs, the construction [22.3] gives $d = 6$. For the goal reward $r = \mathbf{1}\{x = g\}$, $w_r = \frac{1}{6}(e_{(g,\mathsf{f})} + e_{(g,\mathsf{b})})$, so $r_{w_r} = r$ and $\pi_{w_r} = \pi_{\mathsf{f}}$; for $r = \mathbf{1}\{x = 2\}$ the same F, B return the alternating policy of [19.6]. One trained object serves both tasks, whereas the successor features of [19.6] needed a library that happened to contain π_{u}. The price is that F depends on the task embedding, so the representation is a function on $\mathcal{Z} \times \mathbb{R}^d$ rather than a fixed table.

Assumed without proof in this chapter

[22.5] the approximate-FB optimality bound.

23

Hitting Times and Displacement Geometry

The third functional of the trajectory measure is the hitting time of a goal [6.11]. Its mean is the minimal solution of a Poisson equation [2.3(b)] and a quasimetric on the state space [2.5]: it is directional and it composes. This chapter asks when a learned feature map can represent mean hitting times linearly, following Boock et al. (2026). The answer is a displacement representation

$$\kappa^{\pi}(x, g) = \langle \varphi(g) - \varphi(x), \omega_g^{\pi} \rangle,$$

exact under linear closure of the latent dynamics and a range condition, stable under approximate closure with an amplification factor equal to the largest mean hitting time, and impossible for any symmetric representation when the dynamics are directed. $\mathcal{X}$ is finite, $\mathcal{A}$ finite, π stationary, and $g \in \mathcal{X}$ is a goal accessible under P_π from every state, so $\kappa^{\pi}(\cdot, g) < \infty$ [2.4(c)].

23.1 Exact Displacement Representations

Definition 23.1 (Linear closure)**.** Let $\mathbb{H}$ be a real Hilbert space and $\varphi : \mathcal{X} \to \mathbb{H}$. The pair (φ, K) with a bounded linear $\mathsf{K} : \mathbb{H} \to \mathbb{H}$ is a *linear closure of P_π with error $\epsilon \geq 0$* if

$$\| \sum_{x'} P_\pi(x, x') \varphi(x') - \mathsf{K}\varphi(x) \|_{\mathbb{H}} \leq \epsilon$$

for all $x \neq g$; it is *exact* if $\epsilon = 0$ (Boock et al., 2026, Def. 2.1). The *unit representer* for g is any $\mathsf{w}_1 \in \mathbb{H}$ with $\langle \varphi(x), \mathsf{w}_1 \rangle = 1$ for all $x \neq g$.

Theorem 23.2 (Hitting-time isomorphism; Boock et al., 2026, Thm. 2.2)**.** *Let (φ, K) be an exact linear closure of P_π. For $\omega \in \mathbb{H}$ define*

$$\hat{\kappa}_\omega(x) := \langle \varphi(g) - \varphi(x), \omega \rangle.$$

Then $\hat{\kappa}_\omega = \kappa^{\pi}(\cdot, g)$ if and only if

$$\langle \varphi(x), (\mathsf{K} - I)^* \omega \rangle = 1 \qquad \text{for all } x \neq g,$$

i.e. iff $(\mathsf{K} - I)^* \omega$ is a unit representer. Consequently $\kappa^{\pi}(\cdot, g)$ admits a displacement representation iff a unit representer lies in the range of $(\mathsf{K} - I)^*$.

Proof. $\hat{\kappa}_\omega(g) = 0$. For $x \neq g$, by linearity and exact closure,

$$P_\pi \hat{\kappa}_\omega(x) = \langle \varphi(g), \omega \rangle - \Big\langle \sum_{x'} P_\pi(x, x')\varphi(x'), \omega \Big\rangle = \langle \varphi(g), \omega \rangle - \langle \mathsf{K}\varphi(x), \omega \rangle,$$

so

$$\hat{\kappa}_\omega(x) - 1 - P_\pi \hat{\kappa}_\omega(x) = \langle (\mathsf{K} - I)\varphi(x), \omega \rangle - 1 = \langle \varphi(x), (\mathsf{K} - I)^* \omega \rangle - 1.$$

Thus $\hat{\kappa}_\omega$ solves the Poisson equation [2.3(b)] iff the stated condition holds. That equation has the unique solution $\kappa^\pi(\cdot, g)$ among functions vanishing at g [2.4(c)]. ∎

Corollary 23.3 (Tabular case)**.** *With one-hot features $\varphi(x) = e_x \in \mathbb{R}^{\mathcal{X}}$ and K the matrix with $\mathsf{K}e_x = \sum_{x'} P_\pi(x, x') e_{x'}$ (the transpose of P_π), the representer is $\omega_g^\pi = -\kappa^\pi(\cdot, g)$, up to adding a constant vector: the hitting-time function is its own representer.*

Proof.

$$\langle e_x, (\mathsf{K} - I)^* \omega \rangle = \langle (\mathsf{K} - I)e_x, \omega \rangle = P_\pi \omega(x) - \omega(x),$$

which equals 1 for $\omega = -\kappa^\pi(\cdot, g)$ and $x \neq g$ by [2.3(b)]. Adding $c\mathbf{1}$ changes neither $P_\pi \omega - \omega$ nor $\hat{\kappa}_\omega(x) = \omega(g) - \omega(x)$. Apply [23.2]. ∎

Fact 23.4 (Identifiability; Boock et al., 2026, Thm. 2.3)**.** *If two exact representations (φ, K) and $(\varphi^*, \mathsf{K}^*)$ both represent the mean hitting times of a* separating *policy for all goals, and the maps sending a displacement d to the hitting-time profile $h \mapsto \langle d, \omega_h \rangle$ are bounded and bounded below into $L^2(\nu)$ for a full-support ν, then there is a unique bounded linear isomorphism M between the displacement spans with*

$$\mathsf{M}(\varphi(g) - \varphi(x)) = \varphi^*(g) - \varphi^*(x),$$

and M intertwines the latent dynamics:

$$\mathsf{M}(\varphi(g) - \mathsf{K}_a \varphi(x)) = \varphi^*(g) - \mathsf{K}_a^* \varphi^*(x)$$

for every action a.

23.2 Approximate Closure

Theorem 23.5 (Stability of the displacement representation; Boock et al., 2026, Thm. 3.3)**.** *Let (φ, K) be a linear closure of P_π with error ϵ, and ω satisfy $\langle \varphi(x), (\mathsf{K} - I)^* \omega \rangle = 1$ for $x \neq g$. Then*

$$\sup_x \left| \kappa^\pi(x, g) - \hat{\kappa}_\omega(x) \right| \leq \Big(\max_x \kappa^\pi(x, g) \Big) \, \|\omega\|_{\mathbb{H}} \, \epsilon.$$

Since

$$\max_x \kappa^\pi(x,g) = \|(I - P_{\pi,C})^{-1}\|_\infty \geq \frac{1}{1-\varrho(P_{\pi,C})}$$

with $C := \mathcal{X} \setminus \{g\}$, the amplification is an effective horizon of the killed chain.

Proof. As in [23.2], for $x \neq g$ the Poisson residual is

$$\mathsf{e}(x) := \hat{\kappa}_\omega(x) - 1 - P_\pi \hat{\kappa}_\omega(x) = \Big\langle \sum_{x'} P_\pi(x,x')\varphi(x') - \mathsf{K}\varphi(x), \omega \Big\rangle,$$

so $|\mathsf{e}(x)| \leq \|\omega\|\epsilon$ by Cauchy–Schwarz. Since $\hat{\kappa}_\omega(g) = 0$, on C we have $(I - P_{\pi,C})\hat{\kappa}_\omega = \mathbf{1} + \mathsf{e}$ and $(I - P_{\pi,C})\kappa^\pi = \mathbf{1}$, so $\hat{\kappa}_\omega - \kappa^\pi = U_C \mathsf{e}$ with $U_C = (I - P_{\pi,C})^{-1} \geq 0$ [2.4(b)]. Hence

$$|\hat{\kappa}_\omega - \kappa^\pi| \leq U_C \mathbf{1} \cdot \|\omega\|\epsilon = \kappa^\pi(\cdot, g)\|\omega\|\epsilon$$

[2.4(c)]. The operator norm of the nonnegative matrix U_C is its largest row sum, $\max_x \kappa^\pi(x,g)$, and it is at least $\varrho(U_C) = \sum_t \varrho(P_{\pi,C})^t$, because $\varrho(P_{\pi,C})$ is an eigenvalue of the nonnegative matrix $P_{\pi,C}$ with a nonnegative eigenvector (Perron–Frobenius), which is then an eigenvector of U_C. ∎

Intuition. A one-step error in the latent dynamics is paid once per step of the walk to the goal, and the walk lasts κ^π steps on average. The far-away goals are the hard ones, in the precise sense that the error bound at x is proportional to the mean hitting time from x itself.

23.3 Symmetric Representations Cannot Be Directional

Proposition 23.6 (Directionality)**.** *Let $\mathsf{s} : \mathcal{X} \times \mathcal{X} \to [0,\infty)$ be symmetric, for example $\mathsf{s}(x,g) = \|\varphi(x) - \varphi(g)\|$ as in the Hilbert representations of Park et al. (2024). If $\kappa^\pi(x,g) \neq \kappa^\pi(g,x)$ for some x, g (both finite), then $\mathsf{s} \neq \kappa^\pi$ on $\{(x,g),(g,x)\}$, and the representation error*

$$\max\{|\mathsf{s} - \kappa^\pi|(x,g), |\mathsf{s} - \kappa^\pi|(g,x)\}$$

is at least $\frac{1}{2}|\kappa^\pi(x,g) - \kappa^\pi(g,x)|$. A displacement representation $\langle \varphi(g) - \varphi(x), \omega_g \rangle$ is not symmetric, because ω_g depends on the goal.

Proof.

$$|\kappa(x,g) - \kappa(g,x)| \leq |\kappa(x,g) - \mathsf{s}(x,g)| + |\mathsf{s}(g,x) - \kappa(g,x)|.$$

∎

23.4 The Running Examples

Example 23.7 (Displacement geometry of the running examples)**.**

(a) In the goal chain, the tabular representer of [23.3] for the goal g is $\omega = -(6, 4, 0)$, and $\hat{\kappa}_\omega(x) = \langle e_g - e_x, \omega \rangle = \kappa(x, g)$. The amplification constant of [23.5] is $\max_x \kappa(x, g) = 6$, while $\frac{1}{1-\xi(P_C)} \approx 5.24$ [2.12].

(b) In the gridworld under the uniform policy, the corner-to-centre and centre-to-corner hitting times are 36 and ≈ 90.74 [2.13]. Any symmetric representation errs by at least 27 on one of these two pairs [23.6], although the chain is reversible.

Assumed without proof in this chapter

[23.4] identifiability of the displacement geometry; the Perron–Frobenius theorem for nonnegative matrices in [23.5].

24

Quasimetric and Reachability Representations

Mean hitting times are a quasimetric [2.5], but they are undiscounted and need not be finite. The fourth functional of the trajectory measure, the discounted reach value $\mathbb{E}_x\gamma^{\tau_g}$ [6.11(b)], is always finite, and its negative logarithm is again a quasimetric. This chapter shows that this *reachability distance* is the successor distance of Myers et al. (2024), computable from the successor measure alone, that it is controlled above and below by hitting times, that it composes through waypoints, and that in deterministic CMPs the shortest-path quasimetric is the largest quasimetric compatible with one-step transitions, which is the principle behind quasimetric RL (Wang et al., 2023). $\mathcal{X}$ and $\mathcal{A}$ are countable.

24.1 Quasimetrics

Definition 24.1 (Quasimetric)**.** A *quasimetric* on $\mathcal{X}$ is a map $\mathsf{q} : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ with $\mathsf{q}(x, x) = 0$ and $\mathsf{q}(x, z) \le \mathsf{q}(x, y) + \mathsf{q}(y, z)$ for all x, y, z; symmetry is not required.

24.2 The Reachability Distance

Theorem 24.2 (Reachability distance; Myers et al., 2024, Thm. 3.4)**.** *For a Markov chain P on $\mathcal{X}$ and $\gamma \in (0, 1)$ let $q^\gamma(x, g) := \mathbb{E}_x\gamma^{\tau_g}$ [2.3(c)] and $\mathsf{d}_\gamma(x, g) := -\log q^\gamma(x, g) \in [0, \infty]$.*

(a) (Successor-measure form.) With the state successor measure $M(x, g) := \sum_t \gamma^t P^t(x, g)$, $q^\gamma(x, g) = M(x, g)/M(g, g)$, so $\mathsf{d}_\gamma(x, g) = \log \frac{M(g,g)}{M(x,g)}$ is the successor distance of Myers et al. (2024).

(b) d_γ is a quasimetric.

(c)

$$\log \frac{1}{\gamma} \cdot \mathsf{s}(x, g) \le \mathsf{d}_\gamma(x, g) \le \log \frac{1}{\gamma} \cdot \kappa(x, g),$$

where $\mathsf{s}(x, g)$ is the length of a shortest positive-probability path and $\kappa(x, g) = \mathbb{E}_x\tau_g$; and if $\mathbb{E}_x\tau_g^2 < \infty$, $\mathsf{d}_\gamma(x, g)/\log\frac{1}{\gamma} \to \kappa(x, g)$ as $\gamma \uparrow 1$.

Proof.

(a)
$$\sum_t \gamma^t \mathbf{1}\{X_t = g\} = \gamma^{\tau_g} \sum_t \gamma^t \mathbf{1}\{X_t = g\} \circ \theta_{\tau_g}$$
on $\{\tau_g < \infty\}$; take $\mathbb{E}_x$ and use the strong Markov property [1.2(c)]: $M(x,g) = q^\gamma(x,g)M(g,g)$, with $M(g,g) \geq 1$.

(b) $q^\gamma(x,x) = 1$. On $\{\tau_y < \infty\}$, $\tau_z \leq \tau_y + \tau_z \circ \theta_{\tau_y}$, so
$$\gamma^{\tau_z} \geq \gamma^{\tau_y} \gamma^{\tau_z \circ \theta_{\tau_y}};$$
by [1.2(c)], $q^\gamma(x,z) \geq q^\gamma(x,y) q^\gamma(y,z)$. Take $-\log$.

(c) $\tau_g \geq \mathsf{s}(x,g)$ $\mathbb{P}_x$-a.s., giving the lower bound. Jensen for the convex $u \mapsto \gamma^u$ gives $\mathbb{E}_x \gamma^{\tau_g} \geq \gamma^{\mathbb{E}_x \tau_g}$, the upper bound. With $s := \log \frac{1}{\gamma} \downarrow 0$,
$$-\log \mathbb{E} e^{-s\tau} = s\mathbb{E}\tau + O(s^2 \mathbb{E}\tau^2)$$
by $e^{-u} \leq 1 - u + u^2/2$ and $\log(1-v) \geq -v - v^2$ for small v.

■

Corollary 24.3 (Controlled reachability distance)**.** *For a CMP with finite $\mathcal{A}$, let*
$$q^{\gamma,\star}(x,g) := \sup_{\pi \in \Pi^{\mathrm{HR}}} \mathbb{E}_x^\pi \gamma^{\tau_g}$$
and $\mathsf{d}_\gamma^\star := -\log q^{\gamma,\star}$.

(a) $q^{\gamma,\star}(\cdot, g) = (1-\gamma)V^\star$ *for the MDP in which g is made absorbing with reward $\mathbf{1}\{x = g\}$; it is attained by a stationary deterministic policy [4.5].*

(b) $\mathsf{d}_\gamma^\star$ *is a quasimetric, and* $\mathsf{d}_\gamma^\star \leq \mathsf{d}_\gamma^\pi$ *for every stationary π.*

(c) In a deterministic CMP,
$$\mathsf{d}_\gamma^\star = \log \frac{1}{\gamma} \cdot \mathsf{s}^\star,$$
where $\mathsf{s}^\star(x,g)$ *is the length of a shortest controlled path.*

Proof. (a) The return with an absorbing rewarded goal is $\gamma^{\tau_g}/(1-\gamma)$ [3.8], and making g absorbing does not change τ_g. (b) Given $\epsilon > 0$, run a policy ϵ-optimal for reaching y, and upon reaching y switch to one ϵ-optimal for reaching z: this history-dependent policy reaches z with
$$\mathbb{E}\gamma^{\tau_z} \geq (q^{\gamma,\star}(x,y) - \epsilon)(q^{\gamma,\star}(y,z) - \epsilon)$$
by the strong Markov property; take the supremum and let $\epsilon \to 0$. (c) A deterministic CMP under a deterministic policy has deterministic τ_g, and the optimum is the shortest path. ■

Intuition. The reachability distance measures how hard it is to get from x to g in units of discount: a goal reachable in n steps for sure is at distance $n \log \frac{1}{\gamma}$, and a goal reachable only with probability p costs an extra $\log \frac{1}{p}$. Unlike the mean hitting time it is finite whenever reaching is possible at all, and unlike the successor measure it is a distance.

24.3 Quasimetric Learning

Theorem 24.4 (Maximal local quasimetric; Wang et al., 2023)**.** *Let the CMP be deterministic, with transition graph $x \to x'$ iff $x' = P(x, a)$ for some a, and let $\mathcal{Q}_1$ be the set of quasimetrics q with $\mathsf{q}(x, x') \le 1$ for every edge. Then $\mathsf{s}^\star \in \mathcal{Q}_1$ and $\mathsf{q} \le \mathsf{s}^\star$ for every $\mathsf{q} \in \mathcal{Q}_1$. Hence for every $\nu \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ with full support, $\mathsf{s}^\star$ is the unique maximiser of $\mathsf{q} \mapsto \int \mathsf{q}\, d\nu$ over $\mathcal{Q}_1$ (when the maximum is finite).*

Proof. $\mathsf{s}^\star$ is a quasimetric (concatenate paths), and an edge is a path of length 1. For $\mathsf{q} \in \mathcal{Q}_1$ and a shortest path $x = x_0 \to \dots \to x_n = g$, the triangle inequality gives

$$\mathsf{q}(x, g) \le \sum_i \mathsf{q}(x_{i-1}, x_i) \le n = \mathsf{s}^\star(x, g)$$

(if no path exists, $\mathsf{s}^\star = \infty$). Uniqueness: a maximiser $\mathsf{q} \le \mathsf{s}^\star$ with $\int \mathsf{q}\, d\nu = \int \mathsf{s}^\star d\nu < \infty$ must agree with $\mathsf{s}^\star$ wherever $\nu > 0$. ∎

Proposition 24.5 (Composition through waypoints)**.** *For every quasimetric q and waypoints $x = w_0, w_1, \dots, w_k = g$, $\mathsf{q}(x, g) \le \sum_i \mathsf{q}(w_{i-1}, w_i)$. For $\mathsf{d}^\star_\gamma$ in a deterministic CMP, equality holds iff the waypoints lie, in order, on a shortest path; for the uncontrolled d_γ, equality holds iff every path from w_{i-1} to w_{i+1} passes through w_i almost surely and the hitting times factorise, as in the goal chain below.*

Proof. Iterate the triangle inequality. Deterministic case: $\log \frac{1}{\gamma}$ times path lengths [24.3(c)]. Uncontrolled case: equality in

$$\mathbb{E}\gamma^{\tau_z} \ge \mathbb{E}\gamma^{\tau_y + \tau_z \circ \theta_{\tau_y}}$$

forces $\tau_z = \tau_y + \tau_z \circ \theta_{\tau_y}$ a.s. (the inequality between the integrands is pointwise), i.e. z is not reached before y. ∎

Intuition. A quasimetric turns long-horizon planning into shortest-path search over waypoints whose pairwise distances were learned locally. This is the multi-stage planning that hierarchical methods and the hitting-time isomorphism of Chapter 23 exploit. Directionality matters at every leg: a symmetric distance would allow plans that use one-way passages backwards.

24.4 The Running Examples

Example 24.6 (Reachability distances in the goal chain)**.** Under π_{u} with $\gamma = 0.9$: $q^{\gamma}(1, g) \approx 0.583$, $q^{\gamma}(2, g) \approx 0.712$, and

$$q^{\gamma}(1, 2) = \sum_{t \geq 1} 2^{-t} \gamma^{t} = \frac{\gamma/2}{1 - \gamma/2} \approx 0.818.$$

Hence $\mathsf{d}_{\gamma}(1, g) \approx 0.540$, $\mathsf{d}_{\gamma}(2, g) \approx 0.339$, $\mathsf{d}_{\gamma}(1, 2) \approx 0.201$, and $\mathsf{d}_{\gamma}(2, 1) = \infty$. The triangle inequality through the waypoint 2 is an equality, $0.540 = 0.201 + 0.339$, because every path from 1 to g passes through 2 [24.5]. The hitting-time bounds of [24.2(c)] read $0.211 \leq 0.540 \leq 0.632$ ($\mathsf{s} = 2$, $\kappa = 6$, $\log \frac{1}{\gamma} \approx 0.105$). The controlled distance is

$$\mathsf{d}^{\star}_{\gamma}(1, g) = 2 \log \frac{1}{\gamma} \approx 0.211,$$

attained by π_{f}.

Assumed without proof in this chapter

Nothing beyond earlier chapters.

25

Goal-Conditioned Reinforcement Learning

A goal-conditioned agent must reach any state it is asked to reach. Kaelbling (1993) observed that one transition carries information about every goal at once: learning to reach all goals is no harder, per sample, than learning to reach one. This chapter develops that observation into the theory of goal-conditioned reinforcement learning: the goal-augmented MDP and its universal value function (Schaul et al., 2015), all-goals updating and its convergence, the shortest-path relaxations that exploit the triangle inequality of temporal distances, hindsight relabelling (Andrychowicz et al., 2017) with the precise reason it is valid for one-step targets and biased for Monte Carlo targets, and the classification view of goal reaching (Eysenbach et al., 2021, 2022), which identifies goal-conditioned values with successor densities. $\mathcal{X}$, $\mathcal{A}$ are finite; goals are states.

25.1 Goal-Augmented MDPs

Definition 25.1 (Goal-conditioned values)**.** For a CMP $(\mathcal{X}, \mathcal{A}, P)$ and a goal $g \in \mathcal{X}$, let $P^{(g)}$ make g absorbing and let $r_g(x,a) := \mathbf{1}\{x = g\}$. The *goal-conditioned action value* of π is $Q^\pi(z,g)$ for $(P^{(g)}, r_g)$, and $Q^\star(z,g)$ its optimum; with $\gamma < 1$, $V^\star(x,g) = q^{\gamma,\star}(x,g)/(1-\gamma)$ [24.3(a)]. The *goal-augmented MDP* has states $(x,g) \in \mathcal{X}\times\mathcal{X}$, kernel $P((x',g)\,|\,(x,a,g)) := P^{(g)}(x'\,|\,x,a)$ (the goal never changes) and reward $r_g(x,a)$. A function of (x,a,g) approximating $Q^\star(x,a,g)$ is a *universal value function* (Schaul et al., 2015).

Theorem 25.2 (All-goals Q-learning; Kaelbling, 1993)**.** *Let a single trajectory (Z_k) of the CMP be generated by any behaviour policy, and for every goal g and pair $z = Z_k$ update*

$$Q_{k+1}(z,g) = Q_k(z,g) + \alpha_k(z)\Big(\mathbf{1}\{X_k = g\} + \gamma \max_{a'} Q_k(Y^g_{k+1}, a', g) - Q_k(z,g)\Big),$$

with $Y^g_{k+1} := X_{k+1}$ if $X_k \neq g$ and $Y^g_{k+1} := g$ if $X_k = g$ (the goal is absorbing for its own problem). If every pair is visited infinitely often and $\alpha_k(z)$ satisfies the Robbins–Monro conditions, then $Q_k(\cdot,g) \to Q^\star(\cdot,g)$ almost surely for every g simultaneously. The behaviour policy need not depend on, or even know, any goal.

Proof. For a fixed g, $Y^g_{k+1} \sim P^{(g)}(\cdot|Z_k)$ given $\mathcal{F}_k$, so the update is Q-learning [8.4] for $(P^{(g)}, r_g)$ driven by the behaviour's pairs. The proof of [8.5] applies to the index set $\mathcal{Z} \times \mathcal{X}$, because the coordinate (z, g) is updated exactly when z is visited, for every g, and the mean of the update is a γ-contraction of the error coordinatewise in g. ∎

Intuition. Goals do not influence the dynamics, so a transition (x, a, x') is a valid sample for the goal-augmented kernel at every goal. Exploration is needed over pairs, not over pairs and goals: the price of learning all goals is computation, not data.

25.2 Shortest-Path Relaxations

Proposition 25.3 (Relaxation preserves bounds; Kaelbling, 1993)**.**

(a) (Times.) If $\mathsf{D} \geq \kappa^\star$ *pointwise on* $\mathcal{X} \times \mathcal{X}$*, then*

$$\mathsf{D}'(x, g) := \min\big(\mathsf{D}(x, g), \mathsf{D}(x, w) + \mathsf{D}(w, g)\big)$$

satisfies $\kappa^\star \leq \mathsf{D}' \leq \mathsf{D}$ *for every waypoint* w*.*

(b) (Reach values.) If $\mathsf{q} \leq q^{\gamma,\star}$ *pointwise, then*

$$\mathsf{q}'(x, g) := \max\big(\mathsf{q}(x, g), \mathsf{q}(x, w)\mathsf{q}(w, g)\big)$$

satisfies $\mathsf{q} \leq \mathsf{q}' \leq q^{\gamma,\star}$*.*

(c) In a deterministic CMP, starting from $\mathsf{D}_0(x, g) := 1$ *if* g *is a one-step successor of* x*,* $\mathsf{D}_0(x, x) = 0$ *and* ∞ *otherwise, repeated relaxation over all waypoints (the Floyd–Warshall recursion) terminates with* $\mathsf{D} = \mathsf{s}^\star = \kappa^\star$*.*

Proof. (a) $\kappa^\star$ is a quasimetric [5.5], so

$$\kappa^\star(x, g) \leq \kappa^\star(x, w) + \kappa^\star(w, g) \leq \mathsf{D}(x, w) + \mathsf{D}(w, g).$$

(b)

$$q^{\gamma,\star}(x, g) \geq q^{\gamma,\star}(x, w)\, q^{\gamma,\star}(w, g)$$

by the concatenation argument of [24.3(b)]. (c) Every relaxed value is the length of a concatenation of edges, hence $\geq \mathsf{s}^\star$; and after the rounds over all w every shortest path, whose intermediate states are distinct, has been assembled (induction on the largest intermediate state in the order of relaxation). In a deterministic CMP $\kappa^\star = \mathsf{s}^\star$. ∎

25.3 Hindsight Relabelling

Theorem 25.4 (When relabelling is valid; Andrychowicz et al., 2017)**.**

(a) (One-step targets.) Let a replay buffer contain transitions (x, a, x') collected with any goal, and let each be stored with substitute goals chosen by any rule, including goals achieved later in the same trajectory. If every (z, g) receives Robbins–Monro updates, all-goals Q-learning [25.2] on the relabelled data converges to $Q^\star(\cdot, g)$ for every g.

(b) (Monte Carlo targets are biased.) Estimating the probability of reaching g by the fraction of trajectories that reach the goal they were relabelled with is biased: in the CMP with one state x_0, one action that moves to g with probability p and stays at x_0 otherwise, and horizon 1, relabelling each trajectory with its final state gives the estimate 1 for the reach probability of g from x_0 on every relabelled sample with goal g, whereas the truth is p.

Proof. (a) The relabelled tuple $((x, g), a, (x', g))$ is distributed as a transition of the goal-augmented kernel at (x, a, g), because the next state is drawn from $P(\cdot|x, a)$ whatever goal is attached; the choice of goals affects only which coordinates are updated and when, which [7.5] allows. (b) A trajectory is relabelled with goal g only if it reached g; the conditional frequency of reaching g given this selection is 1. ∎

Intuition. Relabelling is an off-policy device for the goal coordinate. A one-step target asks what happens after (x, a), which does not depend on why a was taken; a Monte Carlo target asks how often the goal is reached, and choosing the goal after seeing the outcome selects the successes. Goal-conditioned supervised learning (Ghosh et al., 2021) imitates relabelled trajectories and optimises a lower bound on the goal-reaching objective whose slack grows with the distance between the data policy and the learned one (Ghosh et al., 2021, Thm. 3.1).

25.4 Goal Reaching as Classification

Theorem 25.5 (Classifier odds are successor densities; Eysenbach et al., 2021)**.** *Fix a stationary π and a goal law $\nu > 0$ on $\mathcal{X}$. Let*

$$p_+(g|z) := (1-\gamma)\sum_{t\geq 0}\gamma^t \mathbb{P}_z^\pi\{X_{t+1} = g\},$$

the normalised discounted law of future states from time 1 *on. Consider pairs* (z, g) *with* $g \sim p_+(\cdot|z)$ *labelled* $F = 1$ *and* $g \sim \nu$ *labelled* $F = 0$*, with equal prior weights, and the Bayes-optimal classifier* $C(z, g) := \mathbb{P}\{F = 1 \mid z, g\}$.

(a) $\frac{C(z,g)}{1-C(z,g)} = \frac{p_+(g|z)}{\nu(g)}$.

(b) (C-learning Bellman equation.) The odds satisfy

$$\frac{C(z,g)}{1-C(z,g)} = (1-\gamma)\frac{P(g|z)}{\nu(g)} + \gamma \sum_{z'} (P \otimes \pi)(z, z') \frac{C(z', g)}{1 - C(z', g)}.$$

(c) For the reward $r(x') := \mathbf{1}\{x' = g\}$ *collected on arrival,*

$$Q^\pi(z) := \mathbb{E}^\pi_z \sum_{t \geq 0} \gamma^t \mathbf{1}\{X_{t+1} = g\} = \frac{p_+(g|z)}{1-\gamma},$$

so the log-odds equal $\log Q^\pi(z)$ *up to an additive constant depending on* g *only: acting greedily on the classifier is acting greedily on the goal-conditioned values (Eysenbach et al., 2022).*

Proof. (a) Bayes' rule with equal priors. (b) $p_+(g|\cdot)/(1-\gamma)$ is the shifted successor measure $\gamma^{-1}\big((M^\pi_{\mathcal{X}} - I)\big)$ read at g along state-action pairs, so it satisfies

$$\frac{p_+(g|z)}{1-\gamma} = P(g|z) + \gamma \sum_{z'} (P \otimes \pi)(z, z') \frac{p_+(g|z')}{1-\gamma}$$

by the successor-measure equation [6.10(a)] applied from time 1 (equivalently [19.1(c)]); divide by $\nu(g)$ and use (a). (c) Rearrange (a). ∎

Fact 25.6 (Contrastive critics; Eysenbach et al., 2022)**.** *The optimal critic of the InfoNCE objective that contrasts one future state with* K *samples from* ν *is* $f(z, g) = \log \frac{p_+(g|z)}{\nu(g)} + c(z)$*; with the factorised critic* $f(z, g) = \langle \phi(z), \psi(g) \rangle$ *this is a low-rank representation of the successor density, of the same type as the forward–backward factorisation [22.1].*

Proposition 25.7 (Goal-conditioned values are quasimetrics; Wang et al., 2023, Thm. 1)**.** *With unit costs, the optimal goal-conditioned cost* $\kappa^\star(x, g)$ *is a quasimetric [5.5], and so is* $-\log q^{\gamma,\star}(x, g)$ *[24.3(b)]. Conversely, every quasimetric on a finite set is the optimal cost-to-go of some goal-conditioned deterministic MDP on that set (Fact, Wang et al., 2023): quasimetrics are exactly the right function class for goal-conditioned values.*

Proof. The forward direction is cited. For the converse (Fact), the cited theorem builds, for a quasimetric q, the MDP whose action from x to any y costs $\mathsf{q}(x, y)$; its shortest-path distances are q by the triangle inequality. ∎

25.5 The Running Examples

Example 25.8 (All goals in the goal CMP)**.** In the goal CMP [3.8], the uniform behaviour policy with restarts visits all six pairs; by [25.2] the same data yield $Q^\star(\cdot, g)$ for each of the goals 1, 2, g. For the goal 2, $\kappa^\star(1,2) = 1$ by f, while $\kappa^\star(2,1) = 1$ by b and $\kappa^\star(g,1) = \infty$ since g is absorbing; the optimal temporal distances are asymmetric, and $\kappa^\star(1,g) = 2 = \kappa^\star(1,2) + \kappa^\star(2,g)$, so Kaelbling's relaxation [25.3] through the waypoint 2 is tight. Relabelling the trajectory $1 \xrightarrow{\mathsf{f}} 2 \xrightarrow{\mathsf{f}} g$ with the achieved goal 2 gives a valid one-step sample $(1, \mathsf{f}, 2)$ for goal 2 [25.4(a)].

Assumed without proof in this chapter

The lower bound of goal-conditioned supervised learning in the intuition after [25.4]; [25.6] the contrastive critic; the converse in [25.7].

26

Options and Hierarchical Reinforcement Learning

Temporal abstraction lets an agent decide at the level of extended behaviours: go to the door, reach the subgoal, run the eigenoption. The *options* framework of Sutton et al. (1999b) makes this precise: an option is a policy with a stopping rule, and an MDP with options is a semi-Markov decision process. In the language of this book an option is a stopping time on the trajectory measure, and the *option model* is the push-forward of the trajectory measure under the stopped return and the discounted terminal state, the same construction as the reach value of a goal [6.11(b)] with a random target. This chapter proves the SMDP Bellman equations, the intra-option equations and the convergence of intra-option Q-learning, derives the option-critic gradient theorems (Bacon et al., 2017) from the policy-gradient theorem [10.1] on an augmented chain, and places MAXQ (Dietterich, 2000), feudal and goal-conditioned hierarchies (Dayan and Hinton, 1993; Nachum et al., 2018) and the regret theory of options (Fruit and Lazaric, 2017) in this frame. $\mathcal{X}$, $\mathcal{A}$ are finite.

26.1 Options as Stopping Times

Definition 26.1 (Options; Sutton et al., 1999b)**.** A (Markov) *option* is a triple $o = (\mathsf{I}_o, \pi_o, \beta_o)$ of an initiation set $\mathsf{I}_o \subseteq \mathcal{X}$, a stationary policy π_o and a termination function $\beta_o : \mathcal{X} \to [0,1]$. Started at $x \in \mathsf{I}_o$, it runs π_o and, after each transition into X_t, stops with probability $\beta_o(X_t)$; its *duration* $\sigma_o \geq 1$ is the first such time. Formally, adjoin i.i.d. uniform variables U_t to the path (as rewards were adjoined in [17.1]) and set $\sigma_o := \inf\{t \geq 1 : U_t \leq \beta_o(X_t)\}$, a stopping time. A primitive action a is the option with $\pi = \delta_a$, $\beta \equiv 1$, $\mathsf{I} = \mathcal{X}$. A *policy over options* $\mu(o|x)$ chooses an option at each termination.

Definition 26.2 (Option models)**.** The *model* of o consists of the reward model and the discounted termination kernel

$$r^o(x) := \mathbb{E}_x^{\pi_o}\sum_{t<\sigma_o}\gamma^t r(Z_t), \qquad p^o(x,y) := \mathbb{E}_x^{\pi_o}\big[\gamma^{\sigma_o}\mathbf{1}\{X_{\sigma_o}=y\}\big].$$

Both are expectations of the push-forward of the trajectory measure under

$$\omega \mapsto \Big(\sum_{t<\sigma_o}\gamma^t r(z_t),\ \gamma^{\sigma_o}\delta_{x_{\sigma_o}}\Big).$$

For $\beta_o = \mathbf{1}_B$ and $r = 0$, $p^o(x,\cdot)$ is the law of $X_{\tau_B^+}$ weighted by $\gamma^{\tau_B^+}$, whose total mass is the reach value of B from the next step [2.3(c)].

Theorem 26.3 (SMDP Bellman equations; Sutton et al., 1999b, Thm. 1)**.** *Let $\mathcal{O}$ be a finite set of options and μ a Markov policy over options.*

(a)
$$V^\mu(x) = \sum_o \mu(o|x)\Big[r^o(x) + \sum_y p^o(x,y)V^\mu(y)\Big].$$

(b) $\sum_y p^o(x,y) = \mathbb{E}_x^{\pi_o}\gamma^{\sigma_o} \le \gamma$*, so the operator*
$$\mathcal{T}_{\mathcal{O}}V(x) := \max_{o:x\in \mathsf{I}_o}\Big[r^o(x) + \sum_y p^o(x,y)V(y)\Big]$$
is a γ-contraction; its fixed point is $V^\star_{\mathcal{O}} := \sup_\mu V^\mu$, attained by a greedy policy over options.

(c) $V^\star_{\mathcal{O}} \le V^\star$, with equality if $\mathcal{O}$ contains the primitive actions.

Proof.

(a) The return splits at the first termination:
$$G = \sum_{t<\sigma}\gamma^t r(Z_t) + \gamma^\sigma G\circ\theta_\sigma.$$
Take expectations and apply the strong Markov property [1.2(c)] at σ on the path space with adjoined uniforms; after termination the process restarts from X_σ under μ.

(b) $\sigma_o \ge 1$ gives the bound; the contraction follows as in [4.2]. The optimality statement is [4.5] for the semi-Markov decision process whose "actions" are options and whose discounted kernel is p^o: the supermartingale argument of [4.5(a)] applies at the termination times.

(c) Every policy over options is a (history-dependent) policy of the MDP, so $V^\mu \le V^\star$ [4.5(a)]. With primitive options, $\mathcal{T}_{\mathcal{O}} \ge T^\star$ on values, so $V^\star_{\mathcal{O}} \ge V^\star$.

■

Proposition 26.4 (Intra-option equations)**.** *For a Markov option o,*
$$r^o(x) = \sum_a \pi_o(a|x)\Big[r(x,a) + \gamma\sum_{x'}P(x'|x,a)\big(1-\beta_o(x')\big)r^o(x')\Big],$$
$$p^o(x,y) = \gamma\sum_a \pi_o(a|x)\sum_{x'}P(x'|x,a)\cdot\Big[\big(1-\beta_o(x')\big)p^o(x',y) + \beta_o(x')\mathbf{1}\{x'=y\}\Big].$$

Equivalently, with the substochastic kernel $K_o(x,x') := P_{\pi_o}(x,x')(1-\beta_o(x'))$, $r^o = (I-\gamma K_o)^{-1} r_{\pi_o}$: an option model is the resolvent of a chain killed by termination, as the killed potentials of Chapter 2.

Proof. Condition on the first step [3.4(c)]: after the transition to x' the option stops with probability $\beta_o(x')$ (contributing $\gamma \mathbf{1}\{x'=y\}$ to p^o and nothing more to r^o) and otherwise restarts from x'. The resolvent form is [0.9(a)] since $\|\gamma K_o\| \le \gamma$. ∎

Fact 26.5 (Interruption; Sutton et al., 1999b, Thm. 2)**.** *If each option is terminated early whenever continuing is worse than restarting, i.e. at histories ending in x with $Q^{\mu}(x,o) < V^{\mu}(x)$, then the interrupted policy μ' satisfies $V^{\mu'} \ge V^{\mu}$, strictly at every state from which an interruption occurs with positive probability.*

Theorem 26.6 (Intra-option Q-learning; Sutton et al., 1999b, Thm. 3)**.** *Let every option in $\mathcal{O}$ be Markov with a deterministic policy, and define*

$$U^\star(x',o) := (1-\beta_o(x'))Q^\star_{\mathcal{O}}(x',o) + \beta_o(x') \max_{o'} Q^\star_{\mathcal{O}}(x',o').$$

Then

$$Q^\star_{\mathcal{O}}(x,o) = r(x,\pi_o(x)) + \gamma \sum_{x'} P(x'\mid x,\pi_o(x))\, U^\star(x',o).$$

If, after each transition (x,a,x'), every option o with $\pi_o(x)=a$ is updated by

$$Q(x,o) \leftarrow Q(x,o) + \alpha\big(r(x,a) + \gamma U_Q(x',o) - Q(x,o)\big),$$

with Robbins–Monro step sizes for each (x,o) and every $(x,\pi_o(x))$ visited infinitely often, then $Q \to Q^\star_{\mathcal{O}}$ almost surely, whatever the behaviour.

Proof. The displayed equation is the first-step decomposition of [26.3] for Markov options: after one step the option continues with probability $1-\beta_o(x')$ (value $Q^\star_{\mathcal{O}}(x',o)$) or terminates and the greedy policy over options restarts. The map $Q \mapsto r + \gamma P U_Q$ is a γ-contraction in the sup norm, because U_Q is a convex combination of $Q(x',o)$ and $\max_{o'} Q(x',o')$. The learning claim is [7.5] with index set $\mathcal{X}\times\mathcal{O}$, exactly as in [8.5]: every option consistent with the taken action receives an unbiased sample of the target. ∎

26.2 Learning Options by Gradient

Theorem 26.7 (Option-critic gradients; Bacon et al., 2017)**.** *Let options have policies $\pi_{o,\vartheta}$ and terminations $\beta_{o,\vartheta'}$ differentiable in parameters, and fix a policy*

over options μ. Define on the augmented state (x, o)*, with o the option in force,*

$$\begin{aligned} Q_\Omega(x,o) &:= \sum_a \pi_o(a|x) Q_U(x,o,a), \\ Q_U(x,o,a) &:= r(x,a) + \gamma \sum_{x'} P(x'|x,a) U(x',o), \\ U(x',o) &:= (1-\beta_o(x')) Q_\Omega(x',o) + \beta_o(x') V_\Omega(x'), \\ V_\Omega(x') &:= \sum_{o'} \mu(o'|x') Q_\Omega(x',o'), \end{aligned}$$

and let $\mu_\Omega(\cdot|x_0,o_0)$ be the discounted occupancy of the augmented chain [6.10] started at (x_0,o_0).

(a) (Intra-option policy gradient.)

$$\nabla_\vartheta Q_\Omega(x_0,o_0) = \sum_{x,o} \mu_\Omega(x,o|x_0,o_0) \sum_a \nabla_\vartheta \pi_{o,\vartheta}(a|x)\, Q_U(x,o,a).$$

(b) (Termination gradient.)

$$\nabla_{\vartheta'} U(x_1,o_0) = -\sum_{x',o} \mu'_\Omega(x',o|x_1,o_0)\, \nabla_{\vartheta'} \beta_{o,\vartheta'}(x') \big(Q_\Omega(x',o) - V_\Omega(x')\big),$$

with μ'_Ω the discounted occupancy of the augmented chain observed at arrivals, started at (x_1,o_0).

Proof. The augmented chain on $\mathcal{X} \times \mathcal{O}$ moves from (x,o) by drawing $a \sim \pi_o(\cdot|x)$, $x' \sim P(\cdot|x,a)$, and keeping o with probability $1-\beta_o(x')$ or drawing $o' \sim \mu(\cdot|x')$ otherwise; Q_Ω is its value [4.3]. (a) Differentiate the defining equations in ϑ:

$$\nabla Q_\Omega(x,o) = u(x,o) + \gamma \sum_a \pi_o(a|x) \sum_{x'} P(x'|x,a) \nabla U(x',o)$$

with $u(x,o) := \sum_a \nabla \pi_o(a|x) Q_U(x,o,a)$, and $\nabla U(x',o)$ is the average of $\nabla Q_\Omega(x',\cdot)$ under the option-switching law. So

$$\nabla Q_\Omega = u + \gamma P_{\text{aug}} \nabla Q_\Omega,$$

whose solution is the resolvent of the augmented chain applied to u, as in the proof of [10.1]. (b) Differentiate in ϑ': π and Q_U's immediate reward do not depend on ϑ', and

$$\begin{aligned} \nabla U(x',o) = &-\nabla \beta_o(x') \big(Q_\Omega(x',o) - V_\Omega(x')\big) \\ &+ \big[(1-\beta_o(x')) \nabla Q_\Omega(x',o) + \beta_o(x') \nabla V_\Omega(x')\big], \end{aligned}$$

while

$$\nabla Q_\Omega(x',o) = \gamma \sum_a \pi_o(a|x') \sum_{x''} P(x''|x',a) \nabla U(x'',o).$$

Substituting the second identity into the first gives $\nabla U = b + \gamma \tilde{P}\, \nabla U$ for the arrival-indexed augmented kernel $\tilde{P}$ and

$$b(x',o) := -\nabla \beta_o(x') (Q_\Omega(x',o) - V_\Omega(x'));$$

solve by the resolvent. ∎

Intuition. Both gradients are policy gradients of one Markov chain, the chain of (state, option in force). The termination gradient pushes an option to stop where its advantage over the policy over options is negative, which is the gradient form of the interruption theorem [26.5].

26.3 Hierarchies

Proposition 26.8 (MAXQ value decomposition; Dietterich, 2000)**.** *Let a subtask i be solved by a policy over child subtasks (options) a, and let $V(a,x)$ be the expected discounted reward accumulated while child a runs from x (its reward model $r^a(x)$) and $C(i,x,a) := \sum_y p^a(x,y) V(i,y)$ the* completion function*, the value of finishing i after a terminates. Then $Q(i,x,a) = V(a,x) + C(i,x,a)$, and applying this recursively expresses the root value as a sum of child values and completions along the hierarchy.*

26.3(a). for the policy of subtask i, with the option a chosen at x. ∎

Remark 26.9 (Goal-conditioned hierarchies)**.** In feudal hierarchies (Dayan and Hinton, 1993) and their deep descendants (Nachum et al., 2018), a high-level policy chooses a subgoal g every c steps, and a low-level goal-conditioned policy [25.1] tries to reach it: each subgoal is an option whose policy is $\pi(\cdot|\cdot, g)$ and whose termination is "reach g or c steps elapse". For the low level, hindsight relabelling of goals is valid for one-step targets [25.4(a)]. For the high level, the subgoal is an action that changes the low-level behaviour, so a stored high-level transition is stale once the low level has changed; Nachum et al. (2018) correct for this by relabelling the stored subgoal with one that makes the observed low-level actions likely, an off-policy correction that has no analogue at the low level.

Fact 26.10 (Regret with options; Fruit and Lazaric, 2017)**.** *Optimism at the level of options (UCRL-SMDP) has regret*

$$\tilde{O}\big((D\sqrt{|\mathcal{X}|}+C)R_{\max}\sqrt{|\mathcal{X}||\mathcal{O}|n}\big)$$

after n decisions, where D is a diameter of the SMDP and C depends on the tails of option durations and rewards; a matching lower bound in n, $|\mathcal{X}|$, $|\mathcal{O}|$ holds up to these constants. Options help exactly when they shorten the effective diameter more than they increase the uncertainty of their own models.

26.4 The Running Examples

Example 26.11 (An option in the goal CMP)**.** In the goal CMP [3.8], let the option "to the goal" run π_{f} and terminate at g. From 1 its duration is 2, so $p^o(1,g)=\gamma^2$ and, with the goal reward counted on arrival, $r^o(1)=0$; from 2, $p^o(2,g)=\gamma$. With $\mathcal{O}=\{o,\mathsf{b}\}$ and the value $V^\star(g)=\frac{1}{1-\gamma}$ already correct, a single SMDP backup [26.3(b)] gives $V(1)=p^o(1,g)V^\star(g)=V^\star(1)$, where primitive backups [4.11] need two (first at 2, then at 1): the option has compressed the path to the goal. The option model is the reach value of [24.6] under the deterministic policy, $p^o(x,g)=q_g^{\gamma,\pi_{\mathrm{f}}}(x)$.

Assumed without proof in this chapter

[26.5] the interruption theorem; [26.10] regret with options.

Part VI

Trajectory Measures and Foundation Policies

27

The Bridge: Occupancy Laws

Parts IV and V took two different statistics of the same trajectory measure: the law of the return, for one reward at a time, and the mean of the discounted occupancy, for all rewards at once. This chapter takes the law of the discounted occupancy itself, the *distributional successor measure* of Wiltzer et al. (2024), and shows that it is the common refinement of both. It is a push-forward of the trajectory measure in the sense of Akgül et al. (2026); its Bellman equation and contraction are inherited from [6.6]; it determines the return law of every reward and the successor measure; the successor measure alone does not determine return laws; and for $\gamma < \frac{1}{2}$ the occupancy law determines the trajectory measure itself. $\mathscr{X}, \mathscr{A}$ are countable and π is stationary.

27.1 The Occupancy Law

Definition 27.1 (Occupancy law)**.** Let $\mathsf{O}(\omega) = \sum_t \gamma^t \delta_{z_t}$ be the discounted occupancy [6.10], a measure on $\mathscr{Z}$ of mass $\frac{1}{1-\gamma}$, and give the set $\mathbb{O}$ of such measures the total-variation norm $\|m - m'\| := \sup_{\|f\|_\infty \le 1} |mf - m'f|$. The *occupancy law* of π from z is $\Xi_z^\pi := \mathsf{O}_\# \mathbb{P}_z^\pi \in \mathscr{P}(\mathbb{O})$ (Wiltzer et al., 2024).

Theorem 27.2 (Bellman equation of the occupancy law)**.**

(a) O *intertwines with* $\beta_z(m) := \delta_z + \gamma m$*, which is* γ*-Lipschitz on* $(\mathbb{O}, \|\cdot\|)$*, and* O *is 2-Lipschitz from* (Ω, d_γ)*.*

(b) $\Xi^\pi = \mathfrak{B}_{\mathsf{O}} \Xi^\pi$*, where*

$$(\mathfrak{B}_{\mathsf{O}} \Xi)_z = \sum_{z'} (P \otimes \pi)(z, z') (\beta_z)_\# \Xi_{z'} :$$

an occupancy drawn from z *is* δ_z *plus* γ *times an occupancy drawn from the next pair.*

(c) $\mathfrak{B}_{\mathsf{O}}$ *is a* γ*-contraction for* $\bar{W}_1$ *on* $\mathscr{P}(\mathbb{O})$*, and for every path kernel* $\mathbb{Q}$*,*

$$\bar{W}_1\big(\mathsf{O}_\# (\Phi^\pi)^k \mathbb{Q}, \Xi^\pi\big) \le \frac{2\gamma^k}{1-\gamma}.$$

(d) The mean of Ξ_z^π *is the successor measure* $M^\pi(z, \cdot)$*.*

Proof. (a) $\mathsf{O} \circ \mathrm{cons}_z = \delta_z + \gamma \mathsf{O}$; $\|\beta_z(m) - \beta_z(m')\| = \gamma \|m - m'\|$; and [6.10(d)]. (b) [6.8(b)]. (c) The mixture-and-push-forward argument of [6.8(c)] uses only countable mixtures over z' and the Lipschitz property of β_z, so it applies on $\mathbb{O}$; the rate is [6.8(d)] with $L = 2$ and [6.6(c)]. (d) $\mathbb{E}_z^\pi \mathsf{O}(C) = M^\pi(z, C)$ [6.10]. ∎

27.2 One Law for All Rewards

Theorem 27.3 (The occupancy law determines all return laws)**.** *For every bounded reward r on $\mathcal{Z}$, the return is the linear readout $G_r = \langle \mathsf{O}, r \rangle$, so its law is $\eta^\pi_{r,z} = (m \mapsto \langle m, r \rangle)_\# \Xi^\pi_z$. More generally, for rewards $r_1, \dots, r_k$ the joint law of $(G_{r_1}, \dots, G_{r_k})$ under $\mathbb{P}^\pi_z$ is the push-forward of Ξ^π_z under $m \mapsto (\langle m, r_i \rangle)_i$. In particular every risk functional [18.8] of every reward, and the value $Q^\pi_r(z) = \langle M^\pi(z, \cdot), r \rangle$, are functions of Ξ^π_z. The readout $m \mapsto \langle m, r \rangle$ is $\|r\|_\infty$-Lipschitz, so W_1-errors in occupancy laws transfer to return laws with factor $\|r\|_\infty$.*

Proof. $\langle \mathsf{O}(\omega), r \rangle = \sum_t \gamma^t r(z_t)$, and $\langle \beta_z(m), r \rangle = r(z) + \gamma \langle m, r \rangle$: the readout intertwines β_z with the affine maps of the return, so [17.2] identifies its push-forward with the return law. The joint statement is the same with a vector-valued readout. The Lipschitz bound: $|\langle m, r \rangle - \langle m', r \rangle| \le \|r\|_\infty \|m - m'\|$. ∎

Proposition 27.4 (The successor measure does not determine return laws)**.** *There are two Markov chains, on the same three states and from the same start, with the same successor measure and different return laws for the same reward.*

Proof. States $\{s, u, v\}$, start s. Chain A: $s \to u$ or $s \to v$ with probability $\frac{1}{2}$ each, u and v absorbing. Chain B: from each state, move to u or v with probability $\frac{1}{2}$ each. In both, $\mathbb{P}\{X_t = u\} = \mathbb{P}\{X_t = v\} = \frac{1}{2}$ for $t \ge 1$, so

$$M(s, \cdot) = \delta_s + \frac{\gamma}{1 - \gamma} \cdot \frac{1}{2}(\delta_u + \delta_v)$$

in both. For $r = \mathbf{1}_u$, in chain A the return is $\frac{\gamma}{1-\gamma}$ or 0 with probability $\frac{1}{2}$ each; in chain B it is $\sum_{t \ge 1} \gamma^t B_t$ with i.i.d. fair bits B_t, which has variance

$$\frac{1}{4} \sum_{t \ge 1} \gamma^{2t} = \frac{\gamma^2}{4(1 - \gamma^2)},$$

strictly below the variance $\frac{\gamma^2}{4(1-\gamma)^2}$ of chain A. ∎

27.3 When the Occupancy Is the Trajectory

Theorem 27.5 (Injectivity of the occupancy for small discount)**.**

(a) If $\gamma < \frac{1}{2}$, then $\mathsf{O} : \Omega \to \mathbb{O}$ is injective; hence Ξ_z^π determines $\mathbb{P}_z^\pi$, and every functional of the trajectory measure (hitting-time laws, reward-process laws, return laws) is a function of the occupancy law.

(b) If $\gamma = \frac{1}{2}$, O is not injective: the paths $(z, z', z', \dots)$ and $(z', z, z, \dots)$ with $z \neq z'$ have the same occupancy $\delta_z + \delta_{z'}$.

Proof.

(a) Fix $z \in \mathscr{X}$; then $\mathsf{O}(\omega)(\{z\}) = \sum_t \gamma^t b_t$ with $b_t := \mathbf{1}\{z_t = z\} \in \{0, 1\}$. If two bit sequences first differ at t_0, the two sums differ by at least

$$\gamma^{t_0} - \sum_{t > t_0} \gamma^t = \gamma^{t_0} \frac{1 - 2\gamma}{1 - \gamma} > 0.$$

So $\mathsf{O}(\omega)$ determines, for every z, the set of times at which ω visits z, hence ω. O is measurable and injective between standard spaces, so images of measurable sets are measurable and $\mathbb{P}_z^\pi = (\mathsf{O}^{-1})_\# \Xi_z^\pi$ on the range (Çınlar, 2011, §I.2).

(b) $\frac{1}{2} \sum_{t \geq 1} 2^{-(t-1)} = 1$: both paths give z and z' total discounted time 1 each. ∎

Intuition. With a small discount, the occupancy is a base-$\frac{1}{\gamma}$ expansion of the trajectory, and nothing is lost. With a large discount, many trajectories share one occupancy, and the occupancy law keeps exactly what discounted sums of rewards can see. In both regimes the successor measure, the mean of the occupancy law, keeps strictly less.

Theorem 27.6 (The bridge)**.** *For a stationary policy, the following diagram of push-forwards and means commutes, and every Bellman equation on it is the image of $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$:*

$$\begin{array}{ccccc} \mathbb{P}^\pi & \xrightarrow{\mathsf{O}_\#} & \Xi^\pi & \xrightarrow{\text{mean}} & M^\pi \\ & & \downarrow{\scriptstyle \langle \cdot, r \rangle_\#} & & \downarrow{\scriptstyle \langle \cdot, r \rangle} \\ & & \eta_r^\pi & \xrightarrow{\text{mean}} & Q_r^\pi \end{array}$$

Moreover the reach value and the reachability distance of every goal are functions of the top-right corner, $q_g^{\gamma,\pi}(x) = M_{\mathscr{X}}^\pi(x, g) \,/\, M_{\mathscr{X}}^\pi(g, g)$ [24.2(a)]; return laws and risk require the middle column; hitting-time laws require the left column unless $\gamma < \frac{1}{2}$ [27.5].

Proof. Collect [27.2], [27.3], [6.9(d)], [6.10(b)], [24.2(a)] (for the state chain [3.4(a)]) and [27.5]; commutativity is linearity of means and of the readout. ∎

27.4 The Running Examples

Example 27.7 (The occupancy law of the goal chain)**.** For the goal CMP under π_{u} from 1, the occupancy is

$$\mathsf{O} = \sum_{t<\tau_g} \gamma^t \delta_{Z_t} + \frac{\gamma^{\tau_g}}{1-\gamma}\delta_{(g,\cdot)},$$

where the goal mass is spread over the actions chosen at g. Its law records the excursions between 1 and 2, not just their number. With $\gamma = 0.4 < \frac{1}{2}$ the occupancy determines the whole path [27.5(a)]. With $\gamma = 0.9$ it still determines τ_g (the goal mass is $\gamma^{\tau_g}/(1-\gamma)$, injective in τ_g), hence the return law of the goal reward, but two different excursion patterns of the same length may share an occupancy whenever their discounted visit sums at 1 and 2 coincide, as in [27.5(b)].

Assumed without proof in this chapter

The measurability of images under injective measurable maps between standard spaces in [27.5(a)].

28

What a Foundation Policy Must Compute

A foundation policy is trained once, without rewards, and then serves an open class of tasks without further learning. The preceding chapters determined, for each object that current methods learn, which functional of the trajectory measure it is and what can be read off from it. This chapter turns these findings into definitions and theorems: four properties that a foundation representation may have, a table of which objects have which properties (each entry a theorem of the book), separation results showing that no mean-level object answers risk queries and no symmetric object answers directional ones, and a statistical floor that every representation solving all rewards must pay. The properties follow the classification of Boock et al. (2026, Table 1). $\mathcal{X}, \mathcal{A}$ are finite.

28.1 Four Properties

Definition 28.1 (Foundation representations)**.** A *foundation representation* assigns to each CMP $(\mathcal{X}, \mathcal{A}, P)$ an object $\mathsf{R}(P)$, together with read-out maps that produce from $\mathsf{R}(P)$ and a task a policy or a prediction. It has

(F1) *reward-freeness* if $\mathsf{R}(P)$ is a functional of the family of trajectory measures $\{\mathbb{P}^\pi\}$ and is estimable from reward-free transitions;

(F2) *directionality* if it determines a function $\mathsf{D}_\mathsf{R} : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ that equals a temporal distance of the CMP (a mean hitting time [2.5], a reachability distance [24.2], [24.3], or a shortest-path length [24.4]), which is asymmetric in general;

(F3) *compositionality* if D_R is a quasimetric [24.1];

(F4) *zero-shot optimality* for a reward class $\mathscr{R}$ if for every $r \in \mathscr{R}$ the read-out policy is optimal when R is exact, and ϵ-optimal with ϵ bounded by a function of the representation error that vanishes with it.

A representation *supports risk* if the return law η_r^π of every $r \in \mathscr{R}$ (for the policies it represents) is a function of $\mathsf{R}(P)$.

28.2 What Each Object Computes

Theorem 28.2 (Properties of the learned objects)**.** *With $\mathscr{R}$ the bounded rewards (or linear rewards where stated):*

(a) Successor features of a library *with GPI [19.2], [19.3]: F1; F4 only up to the library and for linear rewards [19.4]; no risk.*

(b) Forward–backward representations *[22.1]: F1 (temporal-difference objective [22.4]); F4 for all bounded rewards [22.2], with approximation bound [22.5]; F2 and F3 for goals made absorbing, through the reachability distance* $-\log\big((1-\gamma)V_g^\star\big)$ *computed from F, B [22.2], [24.3(a),(b)], although the stored bilinear quantity $F^\top B$ is not itself a quasimetric; no risk.*

(c) Occupancy laws $\{\Xi^\pi\}$ *[27.1]: F1; F4 for every reward whenever the family is indexed as in FB, since the means are successor measures [27.2(d)]; F2 and F3 as in (b); supports risk for every reward [27.3]; for $\gamma < \frac{1}{2}$, every functional of the trajectory measure [27.5].*

(d) Displacement representations *[23.2]: F1 when (φ, K) is learned from transitions; F2, since the goal-dependent representer allows asymmetry [23.6]; F3 when exact, because the represented quantity is a mean hitting time [2.5]; stability with amplification* $\max \kappa$ *[23.5]; no risk.*

(e) Symmetric Hilbert representations $\|\varphi(x) - \varphi(g)\|$ *(Park et al., 2024): F1 and F3 (a norm satisfies the triangle inequality); not F2 [23.6].*

(f) Maximal local quasimetrics *[24.4]: F1, F2, F3 in deterministic CMPs; in stochastic CMPs the local constraint does not characterise a temporal distance.*

Proof. Each item collects the cited results; the reachability statement in (b) and (c) uses that, for an absorbing goal with the goal reward, $q^{\gamma,\star} = (1-\gamma)V_g^\star$ [24.3(a)] and $V_g^\star$ is read out exactly by an exact FB representation [22.2]. ∎

Remark 28.3 (The published classification)**.** Boock et al. (2026, Table 1) classify offline goal-conditioned methods by directional distance, triangle inequality, temporal-difference training and unsupervised policy training: successor-measure methods (FB, PSM, C-learning, CRL, CMD) are directional and mostly lack the triangle inequality; reachability methods (HIQL, QRL, TMD) are directional with mixed triangle-inequality status; hitting-time methods with Euclidean features (GAS, HILP) satisfy the triangle inequality but are not directional; and their displacement method is listed with all four. Theorem [28.2] refines the triangle-inequality column: for successor-measure methods it concerns the stored quantity, not the information content, since the quasimetric

$\log \frac{M(g,g)}{M(x,g)}$ is computable from any successor measure [24.2(a)].

28.3 Separations

Theorem 28.4 (Separations)**.**

(a) No symmetric representation has F2 on a CMP whose temporal distance is asymmetric; the error on some pair is at least half the asymmetry [23.6].

(b) No representation that is a function of successor measures alone supports risk: there are CMPs with equal successor measures and different return laws [27.4].

(c) A representation that supports risk jointly for the indicator rewards $\mathbf{1}_{\{z\}}$, $z \in \mathcal{Z}$, determines the occupancy laws: the joint return law of these rewards is Ξ^{π}_{z} [27.3]. Risk for all rewards jointly is therefore exactly as demanding as the occupancy law.

Proof. (a), (b) as cited. (c) The vector of returns $(\langle \mathsf{O}, \mathbf{1}_{\{z\}} \rangle)_{z \in \mathcal{Z}}$ is O itself, so its law is Ξ^{π}_{z}. ∎

Theorem 28.5 (The statistical floor)**.** *Suppose an algorithm learns, from reward-free episodes, a representation from which an ε-optimal policy for every reward $r : \mathcal{Z} \to [0,1]$ is read out with probability at least $1-\delta$. Then it is an (ε,δ)-PAC reward-free exploration algorithm [16.1], and on worst-case episodic CMPs it needs $\Omega\big(|\mathcal{X}|^2|\mathcal{A}|H^2/\varepsilon^2\big)$ episodes [16.5]. Only structure, such as low-rank dynamics [21.4] or a restricted reward class, lowers this floor.*

Proof. The planning phase of [16.1] may use any procedure applied to the collected data; reading out a policy from the learned representation is one such procedure. (The definition in [16.1] fixes the plug-in planner; the lower bound of Jin et al. (2020a) applies to arbitrary planners.) Apply [16.5]. ∎

Intuition. The theorems say what to ask of a foundation policy. Directionality and compositionality are properties of a temporal distance, which exists in every CMP and is a functional of its trajectory measures; a representation either computes it or does not. Zero-shot optimality for all rewards needs at least the successor measures of an indexed family of policies; risk needs their occupancy laws; and all of it has a reward-free sample cost that no representation can avoid without structure.

28.4 The Running Examples

Example 28.6 (The properties on the goal CMP)**.** On the goal CMP: the mean hitting times $\kappa^{\pi_u}(1,2) = 2$ and $\kappa^{\pi_u}(2,1) = \infty$ [2.13] rule out every symmetric representation (F2 fails). The FB representation of [22.7] returns π_f for the goal reward and the reachability distance $\mathrm{d}^{\star}_{\gamma}(1,g) = 2\log\frac{1}{\gamma}$ [24.6] (F2, F3, F4). The variance ≈ 4.10 of the return from 1 under π_u [17.15] is not a function of M^{π_u} alone, as the two chains of [27.4] show; it is a function of Ξ^{π_u} [27.3].

Assumed without proof in this chapter

Nothing beyond earlier chapters; [28.5] relies on the Fact [16.5].

29

The Trajectory-Measure Standard and Open Problems

The book began with one measure on infinite trajectories and one equation, $\mathbb{P}^\pi = \Phi^\pi \mathbb{P}^\pi$ [6.6]. Everything since has been a functional of that measure: values and their dynamic programming (Part I), the stochastic approximation of its images from samples (Part II), the statistical cost of learning enough of it (Part III), the laws of its returns (Part IV), the means of its occupancies and its temporal distances (Part V), and the laws of its occupancies (Chapter 27). This chapter states the resulting standard for foundation policies as a hierarchy of certificates, each backed by a theorem of the book, and lists the problems that the hierarchy leaves open.

29.1 The Hierarchy of Certificates

Definition 29.1 (Levels of a foundation representation)**.** A foundation representation [28.1] for a family of stationary policies $\{\pi_w\}$ is *certified at level*

(L0) *value* if it determines $Q_r^{\pi_w}$ for every reward r in a class and every w;

(L1) *occupancy mean* if it determines the successor measures M^{π_w};

(L2) *occupancy law* if it determines the occupancy laws Ξ^{π_w};

(L3) *trajectory* if it determines the trajectory measures $\mathbb{P}^{\pi_w}$;

each with an error bound for its read-outs in terms of a stated representation error.

Theorem 29.2 (What each level certifies)**.**

(a) L3 $\Rightarrow$ L2 $\Rightarrow$ L1 $\Rightarrow$ L0 (for all bounded rewards), by push-forward, means and linear read-out [27.6]; the converse implications fail: L0 for one reward $\not\Rightarrow$ L1, L1 $\not\Rightarrow$ L2 [27.4], and L2 $\not\Rightarrow$ L3 for $\gamma \geq \frac{1}{2}$ [27.5(b)], while L2 $\Leftrightarrow$ L3 for $\gamma < \frac{1}{2}$ [27.5(a)].

(b) L1 for an FB-indexed family certifies zero-shot optimality for all rewards [22.2], [22.5] and the reachability quasimetric of absorbing goals [28.2(b)] (properties F2–F4).

(c) L2 additionally certifies every risk functional of every reward, with Lipschitz transfer of errors [18.9], [27.3].

(d) The errors of every level are images of a path-level error: if the trajectory kernel is within $\bar{\mathsf{W}}$ of the truth, then occupancy laws are within $2\bar{\mathsf{W}}$, re-

turn laws within $\mathrm{sp}(r)\bar{\mathsf{W}}$ *and values within* $\mathrm{sp}(r)\bar{\mathsf{W}}$ *[6.8(d)], [6.9(c)], [27.2(a)], [6.2(d)]; and model errors in a Lipschitz kernel compound at rate* γL *[17.12], [17.13].*

(e) Any level that certifies L0 for all bounded rewards pays the reward-free floor [28.5].

Proof. Collect the cited results. For L0 $\not\Rightarrow$ L1: the values of one reward are one function on $\mathscr{Z}$, which does not determine the $|\mathscr{Z}|\times|\mathscr{Z}|$ successor measure. ■

Intuition. The standard asks a foundation policy to state its level and to prove its read-out bounds at that level. A method that learns values for a fixed reward family is at L0; FB, successor features with GPI and their relatives are at L1; distributional successor measures are at L2; a learned world model that samples trajectories aims at L3. Higher levels answer more questions; every level pays the same reward-free floor unless the environment has structure that the representation exploits.

29.2 Open Problems

Problem 29.3 (Entry-wise certificates for successor measures)**.** Under low-rank dynamics [21.1], give finite-sample, entry-wise error bounds for the shifted successor measure $M^\pi - I$ learned by temporal differences, of the form needed by the action-gap certificate [21.3(b)], and determine whether the $d|\mathscr{X}||\mathscr{A}|$ reward-free rate [21.4(b)] is attainable at level L1 for an FB-indexed family.

Problem 29.4 (Reward-free floors under function approximation)**.** The floor [28.5] is tabular. Determine the reward-free sample complexity of level L1 when the successor measures of the family admit an exact rank-d factorisation $F^\top B$ [22.1], as a function of d and the effective horizon, and decide whether it matches the linear-MDP rates [13.8].

Problem 29.5 (Learning occupancy laws)**.** Level L2 is defined by the fixed point of a γ-contraction on laws of measures [27.2(c)]. Give sample-complexity bounds for estimating Ξ^π in $\bar{W}_1$ from transitions, by analogy with [12.4] for values, and identify the analogue of the total-variance mechanism [12.3] at this level.

Problem 29.6 (Risk-sensitive zero-shot control)**.** Risk-sensitive optimisation leaves stationary dynamic programming [18.10]. At level L2 the occupancy law of a history-augmented process is again a push-forward of a trajectory measure.

Characterise the smallest augmentation (for instance by the occupancy accumulated so far) under which zero-shot CVaR-optimal control is a fixed-point problem, and bound its cost.

Problem 29.7 (Identifiability above $\gamma = \frac{1}{2}$)**.** For $\gamma \geq \frac{1}{2}$ the occupancy is not injective [27.5(b)]. Characterise the σ-algebra on path space generated by O, and hence exactly which functionals of the trajectory measure (hitting-time laws, orders of visits) are identifiable from occupancy laws.

Problem 29.8 (Compounding beyond $\gamma L < 1$)**.** The encoder–decoder bound [17.13] is vacuous when $\gamma L_Q L_E L_D \geq 1$, which happens already in the goal chain [17.16]. Find conditions (contractive latent dynamics in a different metric, or a coupling adapted to the target statistic) under which compressed trajectory measures retain controlled return-law errors.

Problem 29.9 (Finite-sample directional geometry)**.** The displacement representation is exact under linear closure and stable with amplification $\max_x \kappa^\pi(x, g)$ [23.5]. Give finite-sample bounds for learning $(\varphi, \mathsf{K}, \omega_g)$ from offline trajectories that are uniform over goals, and relate the required coverage to the concentrability coefficients of [9.5].

29.3 Closing

Remark 29.10 (The thesis of the book)**.** Reinforcement learning theory has many objects (values, Bellman operators, successor measures, return distributions, temporal distances, representations) and many Bellman equations. They are one object seen through different functionals: the trajectory measure of a policy, the unique fixed point of the trajectory Bellman operator. Each functional that commutes with prepending a step inherits a Bellman equation; each that is Lipschitz for the discounted path metric inherits a contraction; each that is affine inherits dynamic programming for control. The question to ask of any new method is therefore not which equation it solves, but which functional of the trajectory measure it learns, at which level of the hierarchy, and with what certificate.

Assumed without proof in this chapter

Nothing beyond earlier chapters.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the International Conference on Machine Learning*, 2016.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, 2020.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Alekh Agarwal, Kianté Brantley, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms, 2026. Working draft, June 27, 2026.

Abdullah Akgül, Gulcin Baykal, Manuel Haußmann, Mustafa Mert Çelikok, and Melih Kandemir. Distributional active inference. In *Proceedings of the International Conference on Machine Learning*, 2026.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3):357–367, 1967.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado P. van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. The MIT Press, Cambridge, MA, 2023.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(4):679–684, 1957.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33:719–726, 1962.

David Blackwell. Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1):226–235, 1965.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *arXiv preprint arXiv:2101.07123*, 2021.

Magnus Victor Boock, Abdullah Akgül, Mustafa Mert Çelikok, and Melih Kandemir. Hitting time isomorphism for multi-stage planning with foundation policies. *arXiv preprint arXiv:2605.06470*, 2026.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Texts and Readings in Mathematics*. Hindustan Book Agency and Springer, second edition, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47 (2):119–137, 1979.

Erhan Çınlar. *Probability and Stochastics*, volume 261 of *Graduate Texts in Mathematics*. Springer, New York, 2011.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018b.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2021.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, 1993.

Cyrus Derman. *Finite State Markovian Decision Processes*. Academic Press, New York, 1970.

Thomas G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13: 227–303, 2000.

Wolfgang Doeblin. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Revue Mathématique de l'Union Interbalkanique*, 2:77–105, 1938.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the International Conference on Machine Learning*, 2019.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *Proceedings of the International Conference on Learning Representations*, 2021.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Ruslan Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Mattie Fellows, Matthew J. A. Smith, and Shimon Whiteson. Why target networks stabilise temporal difference methods. In *Proceedings of the International Conference on Machine Learning*, 2023.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Ronan Fruit and Alessandro Lazaric. Exploration-exploitation in MDPs with options. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the International Conference on Machine Learning*, 2018.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2021.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Ronald A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

Cassius T. Ionescu Tulcea. Mesures dans les espaces produits. *Atti della Accademia Nazionale dei Lincei. Rendiconti*, 7:208–211, 1949.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51): 1563–1600, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the International Conference on Machine Learning*, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the Conference on Learning Theory*, 2020b.

Leslie Pack Kaelbling. Learning to achieve goals. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1993.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, 2001.

Sham M. Kakade and John C. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2002.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2021.

Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Van Nostrand, New York, 1960.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017. With contributions by Elizabeth L. Wilmer.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2016.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Marlos C. Machado, André Barreto, Doina Precup, and Michael Bowling. Temporal abstraction in reinforcement learning with the successor representation. *Journal of Machine Learning Research*, 24:1–69, 2023.

Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007.

Shie Mannor, Yishay Mansour, and Aviv Tamar. Reinforcement learning: Foundations, 2025. Book draft, October 2025.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the International Conference on Machine Learning*, 2003.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. In *Proceedings of the International Conference on Machine Learning*, 2024.

Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with Hilbert representations. In *Proceedings of the International Conference on Machine Learning*, 2024.

Martin L. Puterman and Timothy C. Y. Chan. *Markov Decision Processes and Reinforcement Learning*. Cambridge University Press, 2025. Prepublication version, August 7, 2025.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*. Academic Press, 1971.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2019.

Tom Schaul, Dan Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the International Conference on Machine Learning*, 2015.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, 2018.

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.

Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Matthew J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Stefan Stojanović. *Low-Rank Structures for Reinforcement Learning: Sample Efficiency, Generalization, and Representation*. PhD thesis, KTH Royal Institute of Technology, 2026.

Ralph E. Strauch. Negative dynamic programming. *The Annals of Mathematical Statistics*, 37:871–890, 1966.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999a.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999b.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, 1993.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Advances in Neural Information Processing Systems*, 2021.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Hado van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2010.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2021.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, Yunhao Tang, André Barreto, Will Dabney, Marc G. Bellemare, and Mark Rowland. A distributional analogue to the successor representation. In *Proceedings of the International Conference on Machine Learning*, 2024.

Yifan Wu, George Tucker, and Ofir Nachum. The Laplacian in RL: Learning representations with efficient approximations. In *Proceedings of the International Conference on Learning Representations*, 2019.

Lin F. Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of the International Conference on Machine Learning*, 2019.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*, 2021.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023.