CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Risk-Aware Safety in Multi-Agent Systems

Online CVaR Q-Learning, Q-Function–Multiplier Decoupling, and the Centralised Scaling Bottleneck

Master's thesis in Computer Science and Engineering

Xi Cheng

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Master's thesis 2026

Risk-Aware Safety in Multi-Agent Systems

Online CVaR Q-Learning,
Q-Function–Multiplier Decoupling,
and the Centralised Scaling Bottleneck

Xi Cheng

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Data Science and AI
OPTIMA Group
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Risk-Aware Safety in Multi-Agent Systems
Online CVaR Q-Learning,
Q-Function–Multiplier Decoupling,

and the Centralised Scaling Bottleneck
Xi Cheng

Supervisor: Anna Gautier, Department of Computer Science and Engineering
Examiner: Peter Damaschke, Department of Computer Science and Engineering

Master's Thesis 2026
Department of Computer Science and Engineering
Data Science and AI
OPTIMA Group
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Risk-Aware Safety in Multi-Agent Systems
Online CVaR Q-Learning, Q-Function–Multiplier Decoupling, and the Centralised Scaling Bottleneck
Xi Cheng
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Some plans look good on average and still go badly wrong. A taxi dispatcher routing a fleet under shifting demand cannot judge its strategy by the typical shift alone: the difference between strategies shows up on the worst shifts, when calls outpace the fleet and passengers are left waiting. The standard framework for such problems is the Markov decision process (MDP), and a standard way to control upper-tail severity is Conditional Value-at-Risk (CVaR), which measures the average cost in the worst few per cent of outcomes.

In practice the dispatcher has no advance model of how the shift will unfold. Each shift differs from the last: an evening event reshapes demand, weather closes routes, vehicles break down; the dispatcher must learn its policy from experience. The problem is one of online reinforcement learning. Borkar and Jain (2014) gave the canonical algorithm for this setting: a tabular Q-learner whose policy, tail threshold, and constraint multiplier evolve together on three timescales. They proved that it converges in the limit. Its finite-time behaviour, and its scaling to multi-agent systems, remain open.

Here we show that the cost Q-function in this algorithm plays two roles — it scores actions, and it drives the constraint multiplier — which coincide in the limit but separate in finite time. We propose three changes that retain the asymptotic CVaR target: we correct the literal cost-side update with a Bellman-consistent one; we decouple the reward and cost Q-functions, running them as separate recursions that recombine only at action selection; and we source the multiplier signal from a rolling buffer of return samples rather than from the cost Q-function itself.

We then stress-test the resulting learner under centralisation, where one learner acts over the joint state-action space of all agents. It clears the five-component Karush–Kuhn–Tucker (KKT)-based operational gate at the calibrated single-agent point and in the centralised lift with one and two agents; the wider single-agent grid characterises the operating envelope. Under the same frozen centralised protocol and per-seed training budget, it stalls on every seed at three agents. We trace the stall to two structural quantities of the joint representation — the volume of its Q-function table and the variance of its cumulative cost — which are not targeted by C1–C3 at fixed centralised representation, and motivate per-agent risk-contribution allocation as the forward construction examined in the final chapter.

Keywords: Conditional Value-at-Risk; online Q-learning; three-timescale stochastic approximation; Lagrangian relaxation; Q-function–multiplier decoupling; finite-time analysis; centralised multi-agent reinforcement learning; risk-contribution allocation.

Acknowledgements

I thank my supervisor, Anna Gautier, for her guidance, critical feedback, and patience throughout this thesis. When I first entered the risk-aware reinforcement learning literature, I read papers linearly, expecting the structure to become clear only after repeated reading. Anna redirected that instinct with a checklist she handed me early on: how risk is defined (chance constraint or CVaR), what the precise optimisation problem is, what algorithmic tools are used, and which references are worth tracing. The reading shifted from passive to structured, and a clearer view of the field's assumptions and open problems emerged.

Her later guidance helped me connect CVaR-constrained decision making, risk contributions, and risk-aware multi-agent coordination into a coherent research direction. She also insisted on precise academic terminology, treating the specialised vocabulary of the field as part of the research itself. She did not push for quick answers when progress stalled; she asked me to sit with the difficulty and think, which over time I came to recognise as part of the methodology.

I thank my examiner, Peter Damaschke, for his time and for supporting the examination process. I am grateful to the Department of Computer Science and Engineering at Chalmers University of Technology and the University of Gothenburg for providing the environment in which this work was carried out, and to the friends and family who supported me through the period.

AI-assisted tools were used for language editing; all scientific claims, experiments, interpretations, and final text were reviewed by and remain the responsibility of the author. Any remaining errors are my own.

Xi Cheng
Gothenburg, 2026-06-06

Contents

List of Figures

List of Tables

List of Acronyms

Below is the list of acronyms used throughout this thesis.

CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
CMDP	Constrained Markov Decision Process
CVaR	Conditional Value-at-Risk
i.i.d.	Independent and Identically Distributed
IQR	Interquartile Range
KKT	Karush–Kuhn–Tucker
LLN	Law of Large Numbers
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MILP	Mixed-Integer Linear Programming
MMDP	Multi-Agent Markov Decision Process
oRMDP	online Risk-constrained MDP
PID	Proportional–Integral–Derivative
RCA	Risk-Contribution Allocation
RL	Reinforcement Learning
R–U	Rockafellar–Uryasev
SA	Stochastic Approximation
SNR	Signal-to-Noise Ratio
TD	Temporal Difference
TV	Total Variation
VaR	Value-at-Risk

Nomenclature

Below is the nomenclature of indices, sets, parameters, variables, and functions used throughout this thesis. Where symbols are first introduced or defined formally, the section or equation reference is given in the body of the thesis rather than here.

Indices

t	Time step within an episode, $t \in \{0, 1, \ldots, T\}$
k	Episode index; also the stochastic-approximation iteration index
τ	Dummy summation index over time steps
i	Agent index in a multi-agent system, $i \in \{1, \ldots, N\}$
s, a	Generic state and action symbols; a also denotes an action argument in Q-function and score expressions such as $J(z, a) = Q^R(z, a) - \lambda Q^C(z, a)$

Sets and spaces

$\mathcal{S}$	State space (per-agent in the multi-agent setting)
$\mathcal{A}$	Action space (per-agent in the multi-agent setting)
$\mathcal{B}$	Rolling buffer of π_b-rollout returns, capacity W
$\mathbb{R}$	Set of real numbers
$\mathbb{N}$	Set of natural numbers

Parameters and step sizes

The thesis is finite-horizon throughout; there is no discount factor.

α	CVaR risk level, fixed across the thesis, $\alpha \in (0, 1)$
β	Recursive VaR proxy on the intermediate timescale; the R–U threshold variable
$\bar{\beta}$	Polyak-window average of β used in diagnostics and stopping signals
$\widehat{\beta}$	Empirical buffer-side VaR/quantile estimator used in sample R–U or RCA estimators
λ	Lagrange multiplier on the slow timescale
ϵ	Exploration parameter for $\pi_b = \epsilon$-greedy on J
T	Finite horizon
N	Number of agents in the centralised multi-agent setting
a_k	Step size of the fast Q, V Bellman recursions
γ_k	Step size of the intermediate-timescale β-update
η_k	Step size of the slow-timescale λ-update; $\eta_k \ll \gamma_k \ll a_k$ asymptotically
W	Window size, used *jointly* for the buffer $\mathcal{B}$ and Polyak averaging of (β, λ); $W = 8\,000$
$W_{\min}$	Buffer activation threshold for the dual update
$\lambda_{\max}$	Upper clip on λ; set to 200 in experiments
C_α	Single-agent CVaR budget at the operating risk level
C_α^{eff}	Centralised joint CVaR budget, $N \cdot C_{\text{base}}$
C_{base}	Per-agent CVaR budget; $C_{\text{base}} = 1.60$ throughout. Joint: $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$; single-agent calibration: $C_\alpha = C_{\text{base}}$
ξ	Centred episode-end Gaussian shock on Y in the centralised Taxi experiments; $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$
σ_ξ	Std. deviation of the episode-end Gaussian shock on Y; $\sigma_\xi = 0.10$, hence $\sigma_\xi^2 = 0.01$

Horizon convention. In Chapters 2–4, T denotes the final decision index, so decisions are taken at $t \in \{0, 1, \ldots, T\}$ and the algorithm's backwards Bellman recursion accesses its terminal boundary at $T{+}1$. In the Taxi implementation of Chapters 6–8 and Appendix A, $T = 12$ denotes the number of joint decisions, with decision columns $t \in \{0, 1, \ldots, T-1\}$ and terminal column index T; no decision is taken at the terminal column. The two conventions are consistent on the algorithmic content of Algorithm 4.1 (the count of Q-updates per episode); the distinction is one of indexing convention between the theoretical and implementation chapters.

Functions, random variables, and learner objects

P	Transition kernel of the underlying MDP
r	Per-step reward
$c, \tilde{c}$	Per-step cost; $\tilde{c}$ when evaluated on the augmented state
Y^π	Cumulative cost under policy π, $Y^\pi = \sum_t c(S_t, A_t)$; the summation range follows the *Horizon convention* below
R^π	Cumulative reward under policy π
Z_t	Augmented state (t, S_t, Y_t); the Taxi implementation additionally carries a previous-action slot at the tabular-index level (Section 8.1)
π	Generic stationary policy
π_b	Executed behaviour policy (ϵ-greedy on J)
π_g	Deployed greedy policy ($\arg\max J$); evaluated separately from π_b; the exploration-induced gap is not closed in this thesis
$V_t^\pi(z;\beta)$	Tail-probability auxiliary $\mathbb{P}^\pi(Y > \beta \mid Z_t = z)$ (*not* a generic value function in this thesis)
Q^R	Reward Q-function table
Q^C	Cost Q-function table; uses $(1/\alpha)\, V_t\, \tilde{c}$ as stage source (C1)
J_t^k	Lagrangian score $Q_t^R - \lambda_k Q_t^C$, consumed only at $\arg\max$ (C2)
$a^\star$	Shared bootstrap selector $\arg\max_a [Q_{t+1}^R - \lambda Q_{t+1}^C]$
$\mathrm{VaR}_\alpha(Y)$	Value-at-Risk of Y at tail level α; threshold above which the upper α-tail begins
$\mathrm{CVaR}_\alpha(Y)$	Conditional Value-at-Risk of Y at level α
$\mathrm{RC}_\alpha(C_i)$	Per-agent risk contribution; the RCA dual signal

The Rockafellar–Uryasev surrogate: notation policy

Two estimators of CVaR appear throughout the thesis under the same root symbol Ξ_α but distinct decorations. The convention below is used without further remark from Chapter 2 onward.

Population surrogate. The Rockafellar–Uryasev surrogate is the deterministic function

$$\Xi_\alpha(\beta; Y^\pi) \;=\; \beta \;+\; \alpha^{-1}\, \mathbb{E}^\pi\big[(Y-\beta)^+\big], \qquad \min_\beta \Xi_\alpha(\beta; Y^\pi) \;=\; \mathrm{CVaR}_\alpha(Y^\pi),$$

of two arguments: the threshold β and the distribution Y^π induced by a stationary policy π. Three forms appear:

$\Xi_\alpha(\beta; Y^\pi)$	explicit form, used whenever β is not at its minimiser or whenever β-dependence is being tracked
$\Xi_\alpha(\beta)$	context-shorthand, when π is fixed by context
Ξ_α^π	policy-shorthand, only when $\beta = \mathrm{VaR}_\alpha(Y^\pi)$, where it equals $\mathrm{CVaR}_\alpha(Y^\pi)$

The form $\Xi_\alpha^\pi(\beta)$ is not used: a shorthand carrying both a policy superscript and an explicit threshold argument conflates the two regimes and is replaced by $\Xi_\alpha(\beta; Y^\pi)$ throughout.

Empirical estimators. Two empirical estimators are constructed from the same variational identity (Eq. (2.8) of Chapter 2); they are distinguished by which policy's return distribution they sample.

$\widehat{\Xi}_\alpha(\mathcal{B})$	R–U sample mean evaluated on the set $\mathcal{B}$ at first use (Eq. (4.17)); thereafter written $\widehat{\Xi}_\alpha^{\mathcal{B}}$
$\widehat{\Xi}_\alpha^{\mathcal{B}}$	Buffer-side estimator on the rolling π_b-return buffer; the C3 dual signal, consumed by the slow λ-update
$\widehat{\Xi}_\alpha^{\text{held}}$	Held-out greedy estimator on independent π_g-rollouts; diagnostic only, never consumed by the slow update

The two empirical estimators target distinct population quantities: $\widehat{\Xi}_\alpha^{\mathcal{B}} \to \mathrm{CVaR}_\alpha(Y^{\pi_b})$ under quasi-stationarity of π_b, and $\widehat{\Xi}_\alpha^{\text{held}} \to \mathrm{CVaR}_\alpha(Y^{\pi_g})$ under independent rollout of π_g. The finite-time gap between the two is part of the empirical record of Chapter 7, not a calibration artefact.

1

Introduction

1.1 Motivation: planning under tail risk

Some plans look good on average and still go badly wrong. A taxi dispatcher routing a fleet under shifting demand cannot judge its strategy by the typical shift alone: the difference between strategies shows up in the worst few per cent of outcomes, and it is in those outcomes that calls outpace the fleet and passengers wait.

This tension — between the typical outcome and the upper α-tail of the cost distribution — recurs across operations research, control, and reinforcement learning whenever a sequential decision is taken under uncertainty. The standard framework for such problems is the Markov decision process (MDP). A *policy* π is the dispatcher's decision rule: given the current state of the system, and where relevant the time within the shift, it selects an action. Executing π induces a distribution over the cumulative cost Y^π, and the simplest summary of that distribution is its expectation $\mathbb{E}[Y^\pi]$. The expectation describes the typical shift but says nothing about how bad a small fraction of shifts can become. For the taxi dispatcher, the cumulative cost Y^π is most naturally read as the fleet's total wait time across customer requests in a shift — a quantity that is bounded on an ordinary shift but can spike severely when demand outpaces supply. We use it as the running example below.

Several measures of tail behaviour have been proposed. A *worst-case* constraint $Y^\pi \leq C$ almost surely (Wu and Durfee, 2010; Agrawal et al., 2016) demands that the wait time never exceed C; in stochastic environments where any large outcome has positive probability, this is typically unsolvable. An *expected-cost* constraint $\mathbb{E}[Y^\pi] \leq C$ (Altman, 1999) controls the average wait time across shifts but is silent on which shifts produce extreme wait times. A *chance* constraint $\mathbb{P}(Y^\pi > C) \leq \delta$ (de Nijs et al., 2017; Gautier et al., 2023a) caps the frequency δ with which wait time exceeds C, but does not distinguish between a shift a few minutes over budget and one in which wait times cascade across the fleet until many calls go unanswered.

Conditional Value-at-Risk (CVaR) at level α targets what the chance constraint leaves uncontrolled: the average cost in the worst α fraction of outcomes. For a random variable Y and $\alpha \in (0, 1)$, $\mathrm{CVaR}_\alpha(Y)$ is the expected value of Y within its worst α fraction (Rockafellar and Uryasev, 2000). For the taxi dispatcher, with $\alpha = 0.10$, the constraint $\mathrm{CVaR}_\alpha(Y^\pi) \leq C_\alpha$ asks that across the worst 10% of

shifts the average wait time stays below the tail budget C_α — targeting the severity of the upper α-tail, not the frequency with which any fixed threshold is crossed. CVaR sits between the three measures above: it is sharper than the expectation, more informative than the chance constraint, and less demanding than the worst case. It controls one tail-severity functional, and does so tractably through a convex variational representation; the formal definition and the comparison with the other constraint forms are deferred to Chapter 2.

The research question this thesis takes up. When the dispatcher has a complete model of the environment, the CVaR-constrained planning problem admits offline Lagrangian solutions (Chow and Ghavamzadeh, 2014; Haskell and Jain, 2015). When the model is unknown and the dispatcher must learn its policy from experience, the problem is one of online reinforcement learning. The canonical algorithm in this setting — introduced by Borkar and Jain (2014) — maintains a tabular Q-learner whose policy, tail threshold, and constraint multiplier evolve concurrently on three timescales, and is proved to converge in the limit to a solution of the constrained problem. The asymptotic picture is well understood. Two questions remain open: how the algorithm behaves before the limit, and whether the single-agent algorithm continues to learn reliably in finite time when the dispatcher must coordinate several taxis whose cumulative wait times share a single CVaR budget. A shared budget couples the agents through the tail of the joint cost: one taxi's overrun on a bad shift consumes capacity any taxi might have needed, so individual dispatch choices can no longer be evaluated independently. The simplest representation lifts the single-agent learner over the joint state-action space of all N taxis; whether that representation continues to support online CVaR learning as N grows is a separate question from the single-agent one. The two questions are linked: a multi-agent learner that inherits a finite-time defect from its single-agent base will inherit it at every N, and a centralised multi-agent representation that itself blocks learning will do so even when the single-agent base is repaired. This thesis takes up both questions, in this order.

1.2 From planning to online learning

Return to the taxi dispatcher. Each shift differs from the last: an evening event reshapes demand, weather closes routes, vehicles break down. The dispatcher has no advance model of which calls will arrive from where, which routes will jam, or which taxis will be available; it has only the record of past shifts under previous policies. Each new shift produces a logged trajectory of (state, action, cost, next state) tuples that the dispatcher uses to update its policy for the next shift. This is online reinforcement learning: a policy is revised after each round of interaction, using only the experience that round produced.

In MDP terms, the dispatcher observes a state S_t, takes an action A_t, receives a reward $r(S_t, A_t)$ and a cost $c(S_t, A_t)$, transitions to S_{t+1}, and repeats; the cumulative cost $Y^\pi = \sum_{t=0}^{T} c(S_t, A_t)$ over a horizon T is the random variable whose tail the CVaR constraint controls.

The canonical algorithm. Borkar and Jain (2014) introduce an online algorithm for the CVaR-constrained MDP problem. Three quantities are maintained concurrently and updated at different rates. On the *fast* timescale, an action-value table scores actions under the current constraint setting. On the *intermediate* timescale, a scalar threshold β estimates where the upper α-tail of cumulative cost begins. On the *slow* timescale, a non-negative scalar multiplier λ grows when the constraint is violated and shrinks when it is satisfied. The action-value table is indexed not only by the environment state and action but also by the time within the shift and the cumulative cost so far, so that the policy can adapt to its proximity to the budget. Under the step-size conditions of Section 2.1.6, the joint iterate converges in the limit to a solution of the CVaR-constrained problem (Borkar and Jain, 2014, Theorem 2); its finite-time behaviour and its multi-agent scaling behaviour lie outside that guarantee and are the object of this thesis.

1.3 The gap addressed by this thesis

The issue, in plain terms, is that one learned table is being asked to do two jobs at once: it scores the available actions, and it estimates whether the risk constraint is being violated. These two jobs tolerate different kinds of error, and the asymptotic guarantee (Borkar and Jain, 2014, Theorem 2) bounds neither at finite time. Making this precise requires looking inside the algorithm.

We adopt a decomposed form of the Borkar–Jain Lagrangian score that makes the reward and cost Q-functions explicit. A *Q-function* (or action-value function) under a fixed policy is a table indexed by (state, action) pairs that records, for each pair, the expected sum of per-step payoffs accumulated from the current time to the horizon T under that policy; the reward Q-function Q^R uses the per-step reward as its payoff, and the cost Q-function Q^C uses the per-step cost. Write z for the state argument — which here carries the time within the shift and the cumulative cost so far, the augmentation fixed formally in Chapter 2 (Definition 2.3) — and a for an action available at z in the sense of Section 1.2. With these conventions, the decomposition reads

$$J(z, a) = Q^R(z, a) - \lambda\, Q^C(z, a), \tag{1.1}$$

where $\lambda \geq 0$ is the constraint multiplier introduced in Section 1.2. A larger λ penalises higher-cost actions more strongly at action selection. The original work maintains a single combined action-value table whose terminal condition embeds the Lagrangian; the decomposition in (1.1) produces the same action-selection rule and makes both Q-functions visible individually. Under this decomposition we identify a finite-time conflation: the cost Q-function Q^C plays two distinct statistical roles — a *value-function* role, entering action selection only through $\arg\max_a[Q^R(z, a) - \lambda Q^C(z, a)]$, under which action-uniform perturbations cancel; and a *dual-estimator* role, in which the initial-state value $Q_0^C(z_0)$ feeds the slow multiplier update directly, with no $\arg\max$ to filter it. The two roles coincide in the asymptotic limit but not at finite k; Section 3.1 makes this precise.

The first question this thesis takes up follows. When the cost Q-function and the dual constraint estimator are produced by a single shared recursion, where does the learner fail in finite time, and can the failures be removed by structural decoupling rather than step-size tuning? The second question is the multi-agent scaling question already named in Section 1.1: once the single-agent learner is structurally decoupled, does the centralised multi-agent representation itself support risk-constrained learning as the number of agents grows?

1.4 Contributions

This thesis makes four contributions, summarised in Figure 1.1.

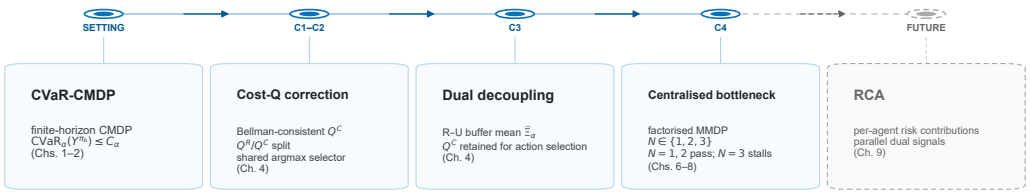

Figure 1.1. Research pipeline. C1 replaces the literal CVaR-side recursion with a Bellman-consistent recursion; C2 separates the reward and cost Q-functions into shared-selector Bellman recursions; C3 sources the multiplier signal from a sample CVaR estimator on a return buffer, retaining the cost Q-function for action selection; C4 stress-tests the resulting learner under centralisation. The diagnosis points forward to a per-agent risk-contribution decomposition (dashed, Chapter 9).

C1 — Bellman-consistent cost Q-function (Section 4.2). We correct the cost-side update of the original algorithm so that the learned cost table tracks the risk quantity it was meant to estimate, rather than a quantity that happens to lie close to it by an algebraic cancellation. Concretely, we replace the literal cost-side recursion of the ORMDP (Borkar and Jain, 2014) with a Bellman-consistent finite-horizon recursion derived from the Rockafellar–Uryasev variational form, whose initial-state fixed point equals $\mathrm{CVaR}_\alpha(Y^\pi)$ at deterministic π and the β-recursion fixed point (Lemma 3.3). Chapter 3 identifies the three places in which the original recursion departs from a textbook Bellman backwards pass, and exhibits the cancellation that hides this departure at the fixed point.

C2 — Q-function–multiplier decoupling (Section 4.3). We maintain the reward and cost Q-functions as two independent tables, so that the constraint multiplier λ enters only at action-selection time and not inside the per-step learning target. Concretely, each table runs its own backwards recursion from a per-step source term that does not depend on λ; the two tables share a single action used inside their Bellman targets and are recombined only when an action is selected. Section 4.3 shows that this construction reproduces the same selected action as the original single-table score, so a moving λ does not invalidate per-step learning targets.

C3 — Decoupled multiplier estimator (Section 4.5). We draw the multiplier's update signal from a rolling buffer of recent shift outcomes rather than reading it from the cost Q-table. The cost table is thereby freed from its second role of supplying the constraint estimate, and is used only to pick actions. The buffer-based estimator is the standard Rockafellar–Uryasev sample CVaR (Rockafellar and Uryasev, 2000); the technical statement is deferred to Section 4.5.

C4 — Centralised multi-agent diagnosis (Chapters 6–8). We stress-test the modified learner on an N-agent Taxi MMDP under centralisation, in which one learner acts over the joint state-action space of all agents. The learner clears the five-component KKT-based operational gate under the frozen centralised protocol at $N \in \{1, 2\}$ but stalls at $N = 3$ on every seed within the per-seed budget, on residual patterns that local hyperparameter adjustment does not remove. We trace the stall to two structural quantities of the joint representation: the size of the joint Q-function table, which governs how precisely each cell can be updated; and the variance of joint returns, which governs the quality of the multiplier estimator's signal. Neither is touched by C1–C3.

1.5 Position, scope, and limitations

The broader research direction motivating this work is system-level coordination under uncertainty: a planner that cannot observe everything, learn perfectly, or compute jointly should still reach decisions whose tail behaviour is acceptable. CVaR captures one component of that objective — the severity of bad outcomes — not the objective itself, and not a general safety guarantee. Risk-Contribution Allocation (Gautier et al., 2023b) extends a single centralised CVaR constraint to per-agent contributions, but does not consider the online setting. In this thesis we examine the online setting, and our centralised diagnosis indicates that scaling beyond the corrected single-agent learner requires moving the role separation we enforce centrally down to per-agent granularity; a per-agent online CVaR estimator is the candidate forward construction. This thesis stops at the centralised diagnosis; the per-agent online learner is sketched in Chapter 9 but not pursued.

The scope is correspondingly narrow. The setting is tabular and finite-horizon. Two of the three changes alter the inner-loop operator of the original three-timescale algorithm, so the limit theorem of Borkar and Jain (2014, Theorem 2) is not inherited verbatim. Section 4.6 (*Scope of the chapter's claims*) records the level at which each change acts within that proof; finite-time rates under the modified system are not quantified here. The centralised diagnosis is empirical, tabular, and restricted to a single controlled MMDP; we identify the first stalled cell under a fixed protocol rather than estimate a scaling law, with the empirical boundaries itemised in Chapter 7. Function approximation, partial observability, non-cooperative settings, and coupling structures beyond weak coupling are separate questions and are not addressed.

1.6 Thesis structure

Chapter 2 fixes the formal framework underlying Chapters 3–8 — the augmented-state finite-horizon MDP and its Bellman equation, CVaR through the Rockafellar–Uryasev variational form, the three-timescale Lagrangian relaxation, and the tower-property identity — situates the thesis against four threads of related work, and states the centralised multi-agent formulation used as the reference test bed in Chapters 6–8.

Chapter 3 develops the single-agent diagnosis. We unroll the CVaR-side recursion of the oRMDP (Borkar and Jain, 2014) under a fixed evaluation policy and identify three structural deviations from Bellman consistency whose joint algebraic interaction at the policy-evaluation fixed point leaves a bounded residual $\alpha\,\tilde{c}(z_0)$ rather than recovering the intended operator (Proposition 3.1); we then state the Bellman-consistent repair (Lemma 3.3) adopted as C1.

Chapter 4 develops C1–C3 into a coherent learner. With the repair of Chapter 3 in place, we derive a closed-form finite-time residual in the dual-estimator role of the cost Q-function (Proposition 4.2), identify its mechanism, and establish the learner's asymptotic target under standard quasi-stationarity.

Chapter 5 validates C1–C3 on a single-agent test bed across a calibrated (α, ρ) grid, identifying the range of tail levels and budgets at which the corrected learner reaches its target.

Chapter 6 fixes the centralised multi-agent test bed: the Taxi MMDP, the per-agent independent ϵ-greedy modification, the implementation stabilisations, and the experimental protocol.

Chapter 7 reports the centralised stress test at $N \in \{1, 2, 3\}$.

Chapter 8 attributes the $N = 3$ stall to the two structural quantities of the centralised representation named in contribution C4.

Chapter 9 sketches the forward direction towards per-agent Risk-Contribution Allocation (RCA) and identifies what would be needed to test it against, and potentially displace, the centralised baseline.

2

Background and Related Work

This chapter fixes the formal framework used throughout Chapters 3–8: finite-horizon MDPs with augmented state, the Bellman equation and Bellman-consistent operators, Value-at-Risk and CVaR, the Rockafellar–Uryasev variational form, the three-timescale Lagrangian relaxation, and the tower-property identity (Section 2.1). Section 2.2 situates the thesis against four threads of literature and states the centralised multi-agent formulation used as the reference test bed in Chapters 6–8.

2.1 Background

2.1.1 Finite-horizon MDPs

Definition 2.1 (Finite-horizon Markov decision process)**.** A *finite-horizon Markov decision process* is a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, c, T \rangle$ in which $\mathcal{S}$ is a finite state space, $\mathcal{A}$ a finite action space, $P(\cdot \mid s, a)$ a transition kernel on $\mathcal{S}$, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ a per-step reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ a per-step cost function, and $T \in \mathbb{N}$ a horizon.

Definition 2.2 (Markov policy)**.** A *Markov policy* is a stochastic kernel $\pi(\cdot \mid s, t)$ on $\mathcal{A}$, specifying for each $(s, t) \in \mathcal{S} \times \{0, 1, \ldots, T\}$ a probability distribution over actions. A policy is *deterministic* if $\pi(\cdot \mid s, t)$ places unit mass on a single action, in which case it is identified with a mapping $\pi : \mathcal{S} \times \{0, 1, \ldots, T\} \to \mathcal{A}$. A policy is *stationary* if $\pi(\cdot \mid s, t)$ does not depend on t, and *time-dependent* otherwise.

The policy-evaluation identities and fixed-selector unrollings used below are stated for deterministic Markov policies; randomisation appears through the ϵ-greedy behaviour policy used during learning.

Executing π from an initial state s_0 produces a random trajectory of state–action pairs $\{(S_t, A_t)\}_{t=0}^{T}$ and two cumulative random variables:

- the cumulative reward $R^\pi = \sum_{t=0}^{T} r(S_t, A_t)$, which records the shift's positive outcomes (e.g., customers delivered on time); and
- the cumulative cost $Y^\pi = \sum_{t=0}^{T} c(S_t, A_t)$, which records the fleet's cumulative wait time across customer requests in the shift.

The CVaR constraints of this thesis operate on the distribution of Y^π, not on its expectation: two policies with identical $\mathbb{E}[Y^\pi]$ can place very different mass on the

upper tail. The augmented state below makes this distribution Markovian relative to the running history.

Definition 2.3 (Augmented state)**.** The *augmented state* at time t is the triple $Z_t = (t, S_t, Y_t)$, where $Y_t = \sum_{u<t} c(S_u, A_u)$ is the cumulative cost accumulated through time $t-1$.

Since the per-step cost $c(s,a)$ is a deterministic function of the state-action pair, the cost component evolves deterministically as $Y_{t+1} = Y_t + c(S_t, A_t)$. The augmentation is not merely notational: Z_t is the state variable on which both the literal Borkar–Jain recursion and the corrected recursion of Chapter 4 are evaluated, and on which the tail-probability and cost-recursion terms of Chapters 3–4 condition (Borkar and Jain, 2014).

2.1.2 The Bellman equation

For a fixed deterministic Markov policy π and a real-valued payoff function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the *value function* of π is

$$V_t^\pi(s) = \mathbb{E}^\pi\left[\sum_{u=t}^{T} g(S_u, A_u) \,\middle|\, S_t = s\right]. \tag{2.1}$$

Definition 2.4 (Bellman equation, policy evaluation)**.** The value function V_t^π of a Markov policy π for payoff g satisfies the *Bellman equation*

$$V_t^\pi(s) = g(s, \pi(s,t)) + \sum_{s' \in \mathcal{S}} P(s' \mid s, \pi(s,t))\, V_{t+1}^\pi(s'), \tag{2.2}$$

with terminal condition $V_{T+1}^\pi(s) = 0$.

Definition 2.5 (Action-value function)**.** The *action-value function* of π for payoff g is

$$Q_t^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{u=t}^{T} g(S_u, A_u) \,\middle|\, S_t = s,\ A_t = a\right], \tag{2.3}$$

satisfying the action-value Bellman equation

$$Q_t^\pi(s,a) = g(s,a) + \sum_{s' \in \mathcal{S}} P(s' \mid s, a)\, V_{t+1}^\pi(s'), \tag{2.4}$$

with terminal condition $Q_{T+1}^\pi(s,a) = 0$ and the link $V_t^\pi(s) = Q_t^\pi\big(s, \pi(s,t)\big)$.

Definition 2.6 (Greedy and ϵ-greedy policies)**.** For an action-value function Q and a fixed tie-breaking rule, the *greedy policy* with respect to Q is the deterministic policy $\pi_Q(s,t) = \arg\max_a Q_t(s,a)$. The *$\epsilon$-greedy policy* at exploration rate $\epsilon \in (0,1)$ selects the greedy action with probability $1-\epsilon$ and an action drawn uniformly at random from $\mathcal{A}$ with probability ϵ.

Definition 2.7 (Bootstrap selector)**.** In control settings where the policy is implicit in the action-value function, the action $a^\star(s') = \arg\max_a Q_{t+1}(s', a)$ that appears inside an action-value Bellman target of the form

$$g(s,a) + \mathbb{E}\big[Q_{t+1}(S_{t+1}, a^\star(S_{t+1})) \,\big|\, S_t = s,\, A_t = a\big]$$

is called the *bootstrap selector*: the action used to bootstrap the next-state value into the target. The bootstrap selector need not coincide with the action executed by the behaviour policy during learning, nor with the action selected by the deployed greedy policy at evaluation time. Chapter 3 uses a single Lagrangian bootstrap selector $a^\star(z') = \arg\max_a \big[Q^R_{t+1}(z', a) - \lambda Q^C_{t+1}(z', a)\big]$ shared by the reward and cost Q-functions.

Definition 2.8 (Bellman-consistent operator)**.** A finite-horizon backwards operator $\mathcal{T} : V_{t+1} \mapsto V_t$ is *Bellman-consistent* for the policy π and payoff g if it reproduces the value function in the sense of Equation (2.2): applied to the true value function at $t+1$ it returns the true value function at t. The same property defined on the action-value form $Q_{t+1} \mapsto Q_t$ uses Equation (2.4).

A backwards operator that is not Bellman-consistent produces a fixed point distinct from the value function. The same recursions apply on the augmented state space Z_t of Definition 2.3 by replacing S with Z and P with the induced augmented transition kernel.

A Bellman consistency claim is an operator-level statement, distinct from a convergence claim and from a numerical-accuracy claim: it asks whether the backwards recursion has the correct fixed point when evaluated with the true next-stage value, prior to any question of whether stochastic approximation reaches that fixed point. Chapter 3 applies this criterion to the cost-side recursion of the ORMDP (Borkar and Jain, 2014) under the reward–cost decomposition adopted there, separately from the asymptotic convergence guarantee of Borkar and Jain (2014, Theorem 2). The diagnosis shows that the CVaR-side recursion is not Bellman-consistent in the sense of Definition 2.8: its fixed-point value at z_0 differs from the intended target.

2.1.3 Constrained MDPs

A constrained MDP (Altman, 1999) adds a constraint on the cumulative cost:

$$\max_\pi \mathbb{E}^\pi[R^\pi] \quad \text{s.t.} \quad \rho(Y^\pi) \le C, \tag{2.5}$$

where ρ is a scalar functional of the cumulative-cost distribution and C a budget. The classical expected-cost constraint case $\rho = \mathbb{E}$ is convex with respect to the right choice of decision variable, and admits a linear-programming solution (Altman, 1999). We work with $\rho = \mathrm{CVaR}_\alpha$ for a fixed $\alpha \in (0,1)$, a conditional-expectation functional that the standard linear-programming machinery does not handle directly (Borkar and Jain, 2014), and treat it through the augmented-state formulation of Section 2.1.1 and the Rockafellar–Uryasev variational form below.

2.1.4 Tail-risk measures: VaR and CVaR

Let Y be a real-valued random variable. Two quantities locate its upper α-tail.

Definition 2.9 (Value-at-Risk)**.** The *Value-at-Risk* of Y at tail level $\alpha \in (0, 1)$ is the threshold above which the upper α-tail begins:

$$\mathrm{VaR}_\alpha(Y) = \inf\{y \in \mathbb{R} : \mathbb{P}(Y > y) \le \alpha\}. \tag{2.6}$$

Definition 2.10 (Conditional Value-at-Risk)**.** Under continuity of the distribution of Y at the threshold $\mathrm{VaR}_\alpha(Y)$, the *Conditional Value-at-Risk* of Y at level α is the mean of Y inside its upper α-tail:

$$\mathrm{CVaR}_\alpha(Y) = \mathbb{E}[Y \mid Y > \mathrm{VaR}_\alpha(Y)]. \tag{2.7}$$

In the finite-horizon tabular setting of this thesis, Y^π may have atoms at the VaR_α threshold, in which case the tail set $\{Y > \mathrm{VaR}_\alpha(Y)\}$ in Equation (2.7) is not unique. The Rockafellar–Uryasev variational form introduced below is taken as the operative definition of CVaR throughout; it avoids the ambiguity of the conditional-tail formula under threshold atoms and is the quantity actually estimated by the algorithms of Chapters 3–4.

Figure 2.1(a) shows both quantities on a representative density: VaR_α marks where the upper α-tail starts, and CVaR_α is the centre of mass of Y restricted to that tail.

Relative to expectation, chance constraints, and worst-case constraints, CVaR is the tail-severity functional used here: it controls the mean magnitude of upper-tail outcomes and admits the variational representation that follows.

The Rockafellar–Uryasev variational form. The definition in Equation (2.7) is hard to optimise over directly: the conditioning event $\{Y > \mathrm{VaR}_\alpha(Y)\}$ depends on the distribution of Y, so when Y is itself controlled, both the integrand and the conditioning region move together. Rockafellar and Uryasev (2000) remove this dependence by recasting CVaR as the minimum of a one-dimensional convex programme.

Definition 2.11 (Rockafellar–Uryasev surrogate and threshold)**.** The *Rockafellar–Uryasev (R–U) surrogate* of Y at tail level α is the function

$$\Xi_\alpha(\beta; Y) = \beta + \alpha^{-1}\,\mathbb{E}\big[(Y - \beta)^+\big], \qquad (x)^+ = \max(x, 0). \tag{2.8}$$

The map $\beta \mapsto \Xi_\alpha(\beta; Y)$ is convex; a minimiser is a VaR_α threshold of Y, and the minimum value equals $\mathrm{CVaR}_\alpha(Y)$. We refer to $\Xi_\alpha(\beta; Y)$ as the *R–U surrogate* (or simply the *surrogate*) and to a minimising β as the *R–U threshold*.

Policy instantiation. Throughout the thesis Y is instantiated as the cumulative cost Y^π induced by a stationary policy π; the surrogate is then written $\Xi_\alpha(\beta; Y^\pi)$. The shorthands $\Xi_\alpha(\beta)$ (when π is fixed by context) and Ξ^π_α (when β is at its minimiser, where it equals $\mathrm{CVaR}_\alpha(Y^\pi)$) follow the convention recorded in the Nomenclature.

Figure 2.1(b) plots the surrogate for the same Y as panel (a). Chapter 4 builds its recursive estimators on this variational form.

2.1.5 CVaR-constrained MDPs

Definition 2.12 (CVaR-constrained MDP)**.** A *CVaR-constrained MDP* is the instantiation of the CMDP problem (2.5) with $\rho = \mathrm{CVaR}_\alpha$:

$$\pi^\star \in \arg\max_\pi \mathbb{E}^\pi[R^\pi] \quad \text{s.t.} \quad \mathrm{CVaR}_\alpha(Y^\pi) \le C_\alpha, \tag{2.9}$$

with α a fixed risk level and C_α a fixed budget.

The constraint is on the distribution that π induces on Y^π; there is no fixed reference distribution. The CVaR-optimal policy is generally non-Markovian on S_t but Markovian on the augmented state $Z_t = (t, S_t, Y_t)$ (Borkar and Jain, 2014). Conditioning on Y_t allows the policy to track its proximity to the tail; conditioning on t allows it to exploit the remaining horizon. For the taxi dispatcher, conditioning on Y_t amounts to letting the dispatch rule depend on how much of the shift's tail budget the fleet has already spent — a quantity that a state-only Markov policy on S_t cannot see.

2.1.6 Lagrangian relaxation and three-timescale stochastic approximation

Combining Equation (2.9) with the variational form (2.8) converts the constrained problem into a saddle-point problem in three variables: the policy π, the R–U threshold β, and a Lagrange multiplier $\lambda \ge 0$.

The saddle point. By the variational form (2.8), the CVaR constraint $\mathrm{CVaR}_\alpha(Y^\pi) \le C_\alpha$ of (2.9) reads $\min_\beta \Xi_\alpha(\beta; Y^\pi) \le C_\alpha$. Standard Lagrangian dualisation of the resulting constrained maximisation problem yields the saddle-point form

$$\min_{\lambda \ge 0} \max_{\pi, \beta} \Big\{ \mathbb{E}^\pi[R^\pi] + \lambda \left(C_\alpha - \Xi_\alpha(\beta; Y^\pi) \right) \Big\}, \tag{2.10}$$

in which the inner $\min_\beta$ is absorbed into the joint maximisation through the sign convention developed below. For a fixed policy π and fixed $\lambda \ge 0$, the inner β-maximisation amounts to a β-minimisation of the R–U surrogate:

$$\max_\beta \big[-\Xi_\alpha(\beta; Y^\pi) \big] = -\min_\beta \Xi_\alpha(\beta; Y^\pi) = -\mathrm{CVaR}_\alpha(Y^\pi). \tag{2.11}$$

The bracketed expression in (2.10) then reduces to $\mathbb{E}^\pi[R^\pi] + \lambda(C_\alpha - \mathrm{CVaR}_\alpha(Y^\pi))$, the standard Lagrangian of the CMDP problem with $\rho = \mathrm{CVaR}_\alpha$. The outer $\min_\lambda$ enforces the constraint through duality. The order and sign convention follow the standard Lagrangian dualisation of a maximisation problem with $\le$ constraint: $\lambda \ge 0$ is minimised in the outer problem so that constraint violation, which makes $C_\alpha - \mathrm{CVaR}_\alpha(Y^\pi)$ negative, penalises infeasible policies in the inner maximisation; β is maximised because it enters only through $-\lambda\, \Xi_\alpha(\beta; Y^\pi)$ with $\lambda \ge 0$, so that the inner β-maximisation is equivalent to the β-minimisation that the R–U form requires.

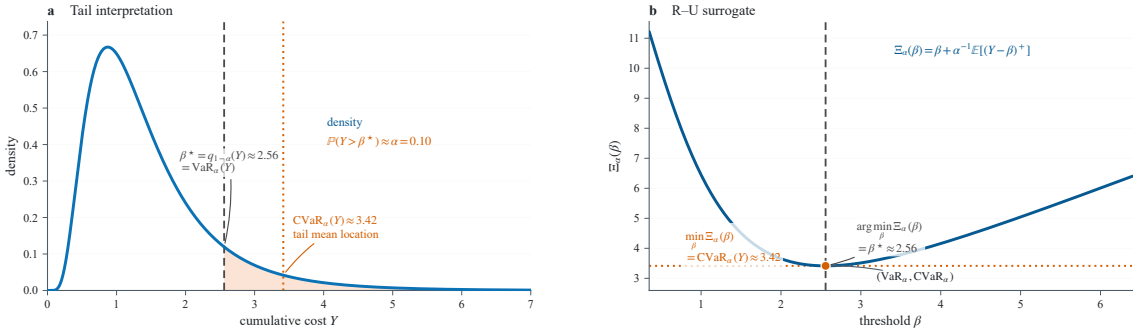

Figure 2.1. The Rockafellar–Uryasev variational form. (a) Cumulative-cost density of a representative Y, with $\mathrm{VaR}_\alpha(Y)$ marking the start of the upper α-tail and $\mathrm{CVaR}_\alpha(Y)$ marking the conditional mean inside the tail. (b) The surrogate $\Xi_\alpha(\beta; Y) = \beta + \alpha^{-1}\mathbb{E}[(Y - \beta)^+]$ as a convex function of β; a minimiser $\beta^\star$ is a VaR_α threshold of Y, and the minimum value equals $\mathrm{CVaR}_\alpha(Y)$.

Three concurrent timescales. Borkar and Jain (2014) solve (2.10) online via three updates running concurrently on a single sampled trajectory. The step sizes follow the schedule

$$a_k = k^{-a_{\mathrm{exp}}}, \qquad \gamma_k = k^{-\gamma_{\mathrm{exp}}}, \qquad \eta_k = k^{-\eta_{\mathrm{exp}}}, \\ \tfrac{1}{2} < a_{\mathrm{exp}} < \gamma_{\mathrm{exp}} < \eta_{\mathrm{exp}} \leq 1. \tag{2.12}$$

The exponent ordering induces three asymptotic rates:

- a_k drives the action-value updates on the *fast* timescale, estimating policy-evaluation quantities at the current (β, λ);
- γ_k drives a β-update on the *intermediate* timescale, tracking $\mathrm{VaR}_\alpha(Y^\pi)$ at the current policy;
- η_k drives a λ-update on the *slow* timescale, enforcing feasibility.

Under the standard multi-timescale machinery of Borkar (2008), fast updates treat slower variables as frozen and slow updates treat faster variables as having reached their stationary value (the asymmetric treatment of faster recursions with respect to slower ones referred to below as *quasi-stationarity* of the faster recursion); the joint iterate then tracks the ODE limit of Equation (2.10) and converges to the saddle-point set, with asynchronous Q-learning at frozen (β, λ) supplying the inner-loop convergence (Tsitsiklis, 1994). The convergence statement is asymptotic; it does not characterise the finite-k regime in which the faster recursions have not yet reached their stationary values.

2.1.7 The tower-property identity

Both the diagnosis in Chapter 3 and the corrected recursion in Chapter 4 rely on a single policy-evaluation identity that converts a global tail-restricted expectation into a per-stage object suitable for online stochastic approximation. Let $V_t^\pi(z) = \mathbb{P}^\pi(Y > \beta \mid Z_t = z)$ denote the conditional probability that the cumulative cost exceeds the threshold β, given the augmented state at time t under policy π. Since

time is included in Z_t, a time-dependent Markov policy on the original state space can be viewed as a stationary Markov policy on the augmented state.

Fix a stationary deterministic evaluation policy π on the augmented state space and let $\tilde{c}_t := c(S_t, \pi(Z_t))$ denote the cost at stage t along the augmented trajectory under this policy. Under this measurability convention, the expected total cost restricted to the tail event $\{Y > \beta\}$ admits a per-stage decomposition, with each $\tilde{c}_t$ weighted by the conditional tail probability $V_t^\pi(Z_t)$:

$$\mathbb{E}^\pi\left[\sum_{t=0}^{T} V_t^\pi(Z_t)\,\tilde{c}_t\right] = \mathbb{E}^\pi[Y\,\mathbb{1}\{Y > \beta\}]\,. \tag{2.13}$$

This follows by conditioning each stage product on Z_t and applying the tower property of conditional expectation (Borkar and Jain, 2014, Lemma 1). Explicitly, for each t the $\sigma(Z_t)$-measurability of $\tilde{c}_t$ combined with $V_t^\pi(Z_t) = \mathbb{E}^\pi[\mathbb{1}\{Y > \beta\} \mid Z_t]$ gives

$$\begin{aligned}\mathbb{E}^\pi\big[V_t^\pi(Z_t)\,\tilde{c}_t\big] &= \mathbb{E}^\pi\big[\tilde{c}_t \cdot \mathbb{E}^\pi[\mathbb{1}\{Y > \beta\} \mid Z_t]\big] \\ &= \mathbb{E}^\pi\big[\mathbb{E}^\pi[\tilde{c}_t\,\mathbb{1}\{Y > \beta\} \mid Z_t]\big] \;=\; \mathbb{E}^\pi\big[\tilde{c}_t\,\mathbb{1}\{Y > \beta\}\big],\end{aligned}$$

the second equality pulling the $\sigma(Z_t)$-measurable factor $\tilde{c}_t$ inside the conditional expectation and the third invoking the tower property. Summing over $t = 0, \ldots, T$ and using $\sum_{t=0}^{T} \tilde{c}_t = Y$, with $Y = Y^\pi$ under the fixed policy π, yields (2.13). We refer to (2.13) as the *tower-property identity* throughout the thesis. It is a policy-evaluation result, not an optimality one, and applies to any stationary deterministic policy whose cost at stage t, $\tilde{c}_t$, is $\sigma(Z_t)$-measurable. In this form it is the policy-evaluation identity of Borkar and Jain (2014, Lemma 1), generalised from the separable per-stage cost assumed there to any $\sigma(Z_t)$-measurable stage cost. Chapter 3 uses it to unroll the cost-side recursion at the policy-evaluation fixed point.

2.2 Related Work

The literature on risk-aware sequential decision-making is large; we restrict attention to the four threads the thesis directly inherits from or contrasts with: online CVaR-constrained Q-learning (Section 2.2.1), value-function decomposition in constrained RL (Section 2.2.2), sample-based CVaR estimation (Section 2.2.3), and weakly coupled multi-agent constrained planning (Section 2.2.4). Section 2.2.5 then states the centralised multi-agent formulation against which the corrected single-agent learner of Chapter 4 is stress-tested in Chapters 6–8.

The four threads enter the thesis along distinct axes. Online CVaR-constrained Q-learning supplies the base algorithm but no finite-time diagnosis of the cost Q-function's two simultaneous roles. Value-function decomposition supplies the algebraic language for separating reward and cost action-values but does not address the CVaR-side dual-readout role. Sample-based CVaR estimation supplies the R–U surrogate inherited by C3 but locates it inside actor-side gradient updates rather than at the dual-side readout of a value-based learner. Weakly coupled multi-agent

constrained planning supplies the per-agent risk-contribution direction that the centralised stress test of Chapters 6–8 is designed to motivate, but operates at planning time rather than online.

Non-asymptotic constrained-RL theory (Efroni et al., 2020; Ding et al., 2020) and distributional RL (Bellemare et al., 2017; Dabney et al., 2018; Lim and Malik, 2022) are adjacent but out of scope: the former proves regret or convergence-rate bounds for families of CMDP algorithms, the latter learns return distributions as an alternative tail-risk paradigm. We diagnose instead the finite-k behaviour of one specific algorithm (Borkar and Jain, 2014) and propose structural changes within its Lagrangian saddle-point framework, focused on the Q-function–multiplier conflation exposed by the reward–cost decomposition used in this thesis.

2.2.1 Online CVaR-constrained Q-learning

The algorithmic starting point is Borkar and Jain (2014), who introduce an online tabular algorithm for CVaR-constrained MDPs — abbreviated ORMDP in the sequel — combining three-timescale stochastic approximation (Borkar, 2008) with asynchronous Q-learning at frozen (β, λ) (Tsitsiklis, 1994). A notation-adapted schematic of the algorithm is given in Algorithm 2.1; we use the same step-size schedule of Equation (2.12) throughout. Borkar and Jain (2014, Theorem 2) gives an asymptotic statement that the iterates converge to the saddle-point set.

Algorithm 2.1 is a notation-adapted schematic of the update equations and terminal readouts used in Chapters 3–4, not a chronological simulator implementation: trajectory variables (X_t^k, U_t^k, Y_t^k) are sampled under an ϵ-greedy behaviour policy on J upstream of the backwards-pass updates, and the per-stage update equations are written in their operator-readout form rather than in the strict temporal order in which a simulator would observe them. The line-by-line correspondence with the original is direct: lines 4–10 of Algorithm 2.1 (the J, U, V, Q, β, λ updates) reproduce lines 3–9 of Borkar and Jain (2014, Algorithm 2, Section V, p. 2578) verbatim up to symbol renaming; the additional initialisation line 1 and the closing greedy-policy return at line 12 are stated explicitly here for completeness. In particular, the Q-update in line 7 inherits unchanged the stage source $V_t^k(z)\,\tilde{c}(z)$ (without $1/\alpha$) and the terminal-frozen bootstrap $Q_{T+1}^k(Z_{T+1}^k)$ that constitute the structural object analysed by Proposition 3.1 of Chapter 3.

In Algorithm 2.1, J is the action-selection table, V tracks the tail probability $\mathbb{P}(Y > \beta \mid Z_t = z)$, and Q supplies the scalar signal consumed by the multiplier update. Trajectory variables X_t^k, U_t^k follow the notation of Borkar and Jain (2014). The augmented-state argument throughout is $z = (t, x, y)$, with y the cumulative-cost coordinate of Definition 2.3; the V and Q terminal conditions $V_{T+1}^k(z) = \mathbb{1}\{y > \beta^k\}$ and $Q_{T+1}^k(z) = y\,\mathbb{1}\{y > \beta^k\}/\alpha$ are read off the y-coordinate of z. The finite-time issue studied in this thesis is that, under the reward–cost decomposition adopted in Chapter 3, the cost-side recursion is asked to support both action evaluation and multiplier estimation; Chapter 3 sets out the diagnosis.

The guarantee is asymptotic and analyses the value recursions at their joint fixed point. It does not address the finite-k regime in which the cost-side recursion is consumed simultaneously by action selection and by the slow multiplier update. We revisit the same algorithm under a decomposed implementation and at finite k; Chapter 3 states the resulting structural mismatch and Chapter 4 states the resulting changes. The point of departure is therefore not the saddle-point objective itself, nor the asymptotic convergence statement, but the finite-time statistical use of the cost-side recursion before the three-timescale separation has materialised.

Algorithm 2.1 oRMDP — Notation-adapted schematic of the online CVaR-constrained Q-learner of Borkar and Jain (2014).

Require: finite state space $\mathcal{X}$, action space $\mathcal{U}$; reward r, cost c; horizon T; tail level $\alpha \in (0,1)$; budget C_α; step-size schedules $a_k = k^{-a_{\text{exp}}}$, $\gamma_k = k^{-\gamma_{\text{exp}}}$, $\eta_k = k^{-\eta_{\text{exp}}}$ with $\frac{1}{2} < a_{\text{exp}} < \gamma_{\text{exp}} < \eta_{\text{exp}} \leq 1$.

Initialise J^0, V^0, Q^0; set $\beta^0 \leftarrow 0$ and $\lambda^0 \leftarrow 0$.
for $k = 1, 2, \ldots$ **do**
  for $t = T, T-1, \ldots, 0$ **do**
    $J_t^k(x,y,u) \leftarrow J_t^{k-1}(x,y,u) + a_k \mathbb{1}\{X_t^k = x,\, Y_t^k = y,\, U_t^k = u\}\big[r(x,u) + \max_{u'} J_{t+1}^k(X_{t+1}^k, Y_{t+1}^k, u') - J_t^{k-1}(x,y,u)\big]$,
    with $J_{T+1}^k(x,y) = \mathbb{1}\{Y_{T+1}^k = y\}\, \lambda^k\big(C_\alpha - \frac{y}{\alpha}\mathbb{1}\{y > \beta^k\}\big)$.
    $U_t^k \leftarrow \arg\max_u J_t^k(X_t^k, Y_t^k, u)$.
    $V_t^k(z) \leftarrow V_t^{k-1}(z) + a_k \mathbb{1}\{Z_t^k = z\}\big[V_{t+1}^k(Z_{t+1}^k) - V_t^{k-1}(z)\big]$,
    with $V_{T+1}^k(z) = \mathbb{1}\{Y_{T+1}^k = y\}\, \mathbb{1}\{y > \beta^k\}$.
    $Q_t^k(z) \leftarrow Q_t^{k-1}(z) + a_k \mathbb{1}\{Z_t^k = z\}\big[\tilde{c}(z)\, V_t^k(z) + Q_{T+1}^k(Z_{T+1}^k) - Q_t^{k-1}(z)\big]$,
    with $Q_{T+1}^k(z) = y\, V_{T+1}^k(z)/\alpha$.
  end for
  $\beta^k \leftarrow \beta^{k-1} - \gamma_k\big(\alpha - V_0^k(z_0)\big)$. ▷ intermediate timescale: track VaR_α
  $\lambda^k \leftarrow \Big(\lambda^{k-1} - \eta_k\big(C_\alpha - Q_0^k(z_0)\big)\Big)^+$. ▷ slow timescale: enforce CVaR constraint
end for
return greedy policy $\pi_g(z) = \arg\max_u J(z,u)$.

Risk-sensitive policy-gradient search (Prashanth and Fu, 2022) treats risk constraints and risk objectives through parameterised-policy gradient methods. With a gradient update against a policy parameterisation, no single recursion serves both action selection and dual estimation, so the conflation diagnosed here does not arise.

Two recent lines bear on this finite-time gap. Bastani et al. (2022) and Wang et al. (2023) establish regret bounds for CVaR-objective RL through optimism-based exploration in the augmented-state MDP, with Wang et al. (2023) achieving near-minimax-optimal rates; this path is distinct from the three-timescale SA Lagrangian framework analysed here, and the conflation diagnosis below does not contradict those rates. Muni et al. (2026) independently identify a degenerate-fixed-point pathology in the augmented-state CVaR Bellman recursion and propose an operator-level reward redistribution; C1 (Section 4.2) addresses the analogous defect at the recursion level inside the SA Lagrangian path.

2.2.2 Value-function decomposition in constrained RL

C2 builds on the decomposition of a single Lagrangian Q-function into separate reward and cost Q-functions. Russell and Zimdars (2003) introduce Q-decomposition for reinforcement-learning agents and show that linear additive structure in the per-step payoff yields a corresponding decomposition of the action-value function under a central arbitrator that combines subagent Q-values. Chapter 4 recasts this for the Lagrangian selector shared by reward and cost Q-functions, which underpins the greedy-invariance argument of that chapter.

In many Lagrangian constrained-RL implementations, reward and constraint signals are estimated separately and combined through a multiplier at policy improvement time. Tessler et al. (2019) treat reward-constrained policy optimisation as a saddle-point problem with a cost Q-function alongside the reward Q-function, combining them at policy improvement through the multiplier. Stooke et al. (2020) adopt a related Lagrangian constrained-RL architecture and replace the purely integral multiplier update by a PID-style controller to damp transient oscillations. In both cases the cost Q-function estimates expected constraint cost; the decomposition is motivated by credit-assignment tractability, and the cost Q-function does not carry the specific CVaR-side initial-state readout role diagnosed in this thesis.

These works decompose the Q-function for credit-assignment reasons in the expected-cost setting and combine the reward and cost Q-functions at policy improvement through the multiplier. We use reward–cost decomposition as a notation-adapted re-expression of the Borkar–Jain Lagrangian score, not as a claim that their algorithm explicitly maintains separate Q^R and Q^C Q-functions. The existing decomposition literature does not address the CVaR-specific question of whether the cost-side recursion can serve both as an action-value table and as the multiplier estimator at finite k; Chapter 4 gives the construction.

Thus the contribution of C2 is not the act of separating reward and cost Q-functions, which is established practice in Lagrangian constrained RL, but the functional position the decomposition occupies in the oRMDP finite-time diagnosis: it prevents a moving multiplier from entering the learning target while preserving the Lagrangian selector used for control.

2.2.3 CVaR estimators and the form underlying C3

The variational form of Rockafellar and Uryasev (2000), $\mathrm{CVaR}_\alpha(Y) = \min_\beta \Xi_\alpha(\beta; Y)$, makes sample-based CVaR estimation tractable and underlies much algorithmic work on CVaR-constrained MDPs. Chow and Ghavamzadeh (2014) develop policy-gradient methods that optimise CVaR objectives by sample-based estimation of $\Xi_\alpha(\beta; \cdot)$ on a rollout buffer and chain-rule differentiation through the trajectory distribution. Tamar et al. (2015) derive a likelihood-ratio gradient estimator for CVaR based on a conditional-expectation formula for its gradient. In both, the empirical R–U sample mean serves as a gradient signal for an actor update, evaluated and consumed episodically. These works do not place the empirical surrogate inside a value-based learner as the dual-side readout.

C3 reuses the surrogate at that position: in the terminology of Chapter 3, it replaces the (R2) dual-estimator readout of the cost Q-function with an empirical R–U surrogate, leaving the cost Q-function to serve only its (R1) value-function role through the Lagrangian selector. The construction is given in Chapter 4.

2.2.4 Weakly coupled multi-agent constrained planning

The constrained MDP framework (Altman, 1999) provides the optimisation template the thesis inherits at the single-agent level: maximise an objective subject to a scalar functional of a cost distribution. Meuleau et al. (1998) provide an early weakly coupled MDP formulation for multi-agent extensions — agents have independent local dynamics and rewards, with interaction only through a global resource constraint — and show that this structure admits decompositions that scale better than reasoning over the joint MDP. De Nijs et al. (2021) provide a taxonomy of constrained multi-agent MDP problems, comparing approaches along five dimensions: constraint strictness (soft vs. hard), agent connectivity, observability of the domain, constraint timespan (instantaneous vs. budget), and resource model (single vs. multiple).

A line of work from Gautier and collaborators develops planning-time algorithms in this space. Gautier et al. (2022) treat space–time as the shared resource for non-cooperative multi-robot path planning. Gautier et al. (2023a) introduce a multi-unit auction for chance-constrained resource allocation with a tractable MILP reduction. Gautier et al. (2023b) address the CVaR-constrained case: they formulate the risk-constrained MMDP, decompose the joint CVaR constraint into per-agent risk contributions in the sense of Tasche (1999), and update the worst reward-to-risk trade-off agent's policy through a single-agent risk-contribution-constrained planner. Gautier et al. (2023b) do not consider the online setting. Planning-time risk-contribution allocation assumes access to a model or planner capable of evaluating per-agent CVaR contributions on the joint distribution; the centralised stress test in Chapters 6–8 bears on whether an online value-based learner can generate analogous signals from sampled trajectories under centralised execution, without prior CVaR-decomposition machinery.

In risk-sensitive cooperative MARL, Shen et al. (2023) introduce an Individual-Global-Max value factorisation that aligns per-agent distorted-risk action selection with the central policy. That line targets joint-return optimisation rather than constraint enforcement at the per-agent level, and is complementary to the constraint-decomposition direction this thesis inherits from Gautier et al. (2023b). Neither line addresses the finite-time behaviour of an online CVaR-constrained Q-learner under multi-agent scaling.

2.2.5 Centralisation as the reference test bed of this thesis

The multi-agent diagnosis in Chapters 6–8 takes the centralised setting — a single learner over the joint state-action space with a single CVaR budget — as its reference test bed for the corrected single-agent learner of Chapter 4. Centralisation is the

simplest multi-agent baseline: it requires no per-agent infrastructure and preserves the role separation of the single-agent learner.

Factored MMDP. For N agents with local components $(\mathcal{S}_i, \mathcal{A}_i, P_i, r_i, c_i)$ and shared horizon T, under a weak-coupling structure (Meuleau et al., 1998; de Nijs et al., 2021) the joint transition factorises and the joint reward and cost are additive: $P^{(N)}(\cdot \mid s, a) = \prod_i P_i(\cdot \mid s_i, a_i)$, $r(s,a) = \sum_i r_i(s_i, a_i)$, $c(s,a) = \sum_i c_i(s_i, a_i)$. The joint cumulative cost is $Y^\pi = \sum_i Y_i^\pi$.

Centralised CVaR-CMDP. The centralised problem instantiates the CVaR-CMDP (2.9) on the factored MMDP:

$$\begin{aligned} \pi^\star \in \arg\max_\pi \; & \mathbb{E}^\pi[\textstyle\sum_i R_i^\pi] \\ \text{s.t. } & \mathrm{CVaR}_\alpha(\textstyle\sum_i Y_i^\pi) \le C_\alpha^{\mathrm{eff}}, \\ & C_\alpha^{\mathrm{eff}} := N \cdot C_{\mathrm{base}}. \end{aligned} \tag{2.14}$$

The per-capita scaling $C_\alpha^{\mathrm{eff}} = N \cdot C_{\mathrm{base}}$ keeps the budget per agent comparable across N; its consequences for the standardised tail margin are taken up in Chapter 8.

Why centralisation is the reference test bed. The single-agent learner of Chapter 4, defined over $Z_t = (t, S_t, Y_t)$, lifts directly to the joint state-action space, where the augmented state has size $T \cdot |\mathcal{S}|^N \cdot |\mathcal{Y}|$ and the action space has size $|\mathcal{A}|^N$. The centralised learner inherits the same role separation we enforce in the single-agent setting (Chapter 3); persistent failures under the fixed protocol of Chapters 6–8 are therefore more naturally attributed to the joint representation than to the specific single-agent mechanisms corrected in Chapter 4. The resulting stress test is reported in Chapter 7.

Bridge to the diagnosis. The definitions and references above isolate the exact object analysed in Chapter 3. Borkar and Jain (2014) supply the three-timescale online CVaR-constrained learner of Algorithm 2.1; Definition 2.3 makes the tail recursion Markovian on the augmented state; Definition 2.11 fixes the CVaR target through the R–U variational form; Definition 2.8 fixes the operator-level criterion the cost-side recursion must satisfy; and the tower-property identity (2.13) converts the global tail-restricted cost into the per-stage form on which any Bellman-consistent recursion must operate. Chapter 3 applies these five ingredients to the literal cost-side recursion of the oRMDP (Borkar and Jain, 2014) and locates the three structural deviations through which its finite-horizon operator departs from the intended one.

3

Problem Formulation and Diagnosis

Within the Lagrangian relaxation $J = Q^R - \lambda Q^C$, the cost Q-function Q^C serves two structurally distinct roles:

(R1) it scores actions through $\arg\max_a \big[Q^R(z,a) - \lambda\, Q^C(z,a)\big]$;

(R2) its initial-state value $Q_0^C(z_0)$ supplies the constraint estimate consumed by the slow λ-update.

The two roles separate at finite time for distinct structural reasons. Proposition 3.1 establishes that the cost-side recursion of the ORMDP (Borkar and Jain, 2014) violates Bellman consistency through three structural deviations that conspire, at the β-recursion fixed point, to leave the additive residual $\alpha\,\tilde{c}(z_0)$ rather than to recover the intended operator. Lemma 3.3 supplies the Bellman-consistent repair for (R1). Treatment of (R2), which survives any pure value-side correction, is deferred to Chapter 4.

3.1 Two roles of Q^C

The ORMDP (Borkar and Jain, 2014) maintains a single action-value table J with terminal condition

$$J_{T+1}(z) \;=\; \lambda\big(C_\alpha \;-\; y\,\mathbb{1}\{y > \beta\}/\alpha\big), \tag{3.1}$$

together with a separate CVaR-side recursion whose initial value $Q_0^C(z_0)$ feeds the slow λ-update. The action-dependent part of J admits the Lagrangian re-expression $Q^R - \lambda Q^C$ — a notational decomposition of the single-table score, not a claim of separate storage. Under it, Q^R and Q^C share the bootstrap selector (Definition 2.7)

$$a^\star(z') \;=\; \arg\max_a \big[Q^R_{t+1}(z',a) \;-\; \lambda\, Q^C_{t+1}(z',a)\big], \tag{3.2}$$

and recombine only at arg max. The behaviour policy π_b is ϵ-greedy on the same score (Definition 2.6); the deployed greedy policy is $\pi_g(z) = \arg\max_a[Q^R(z,a) - \lambda Q^C(z,a)]$. The exploration-induced gap between $\mathrm{CVaR}_\alpha(Y^{\pi_b})$ and $\mathrm{CVaR}_\alpha(Y^{\pi_g})$ is standard in Lagrangian constrained RL (Tessler et al., 2019; Stooke et al., 2020) and is not closed here.

The sensitivity of Q^C to estimator error differs sharply between the two roles:

(R1) Value-function role. $Q^C(z,a)$ influences the selected action only through $\arg\max_a J(z,a)$; action-uniform perturbations cancel, whereas action-dependent perturbations can flip the argmax.

(R2) Dual-estimator role. The scalar $Q_0^C(z_0)$ enters

$$\lambda \;\leftarrow\; \big(\lambda \;-\; \eta_k\big(C_\alpha - Q_0^C(z_0)\big)\big)^+$$

directly; action-uniform and action-dependent errors transmit unfiltered.

The asymptotic analysis of Borkar and Jain (2014) does not separate (R1) from (R2): under quasi-stationarity of the executed policy on the slow λ-timescale (Borkar, 2008), the recursion's initial-state value converges to $\mathrm{CVaR}_\alpha(Y^{\pi_b})$, collapsing the recursion's limit and the dual update's required input to a single scalar. At finite k they do not.

3.2 The literal cost-side update is not Bellman-consistent

Standing hypotheses.

(H1) π is stationary and deterministic on the augmented state space, so $\tilde{c}_t := c(S_t, \pi(Z_t))$ is $\sigma(Z_t)$-measurable for every t.

(H2) $\beta = \mathrm{VaR}_\alpha(Y^\pi)$ under the continuity hypothesis of Section 2.1.4; in particular $\mathbb{P}^\pi(Y > \beta) = \alpha$.

(H3) $V_0^\pi(z_0) = \alpha$ at the β-recursion fixed point. (This follows from (H2) and the augmented-state Markov property: since $Z_0 = z_0$ is deterministic, $V_0^\pi(z_0) = \mathbb{P}^\pi(Y > \beta \mid Z_0 = z_0) = \mathbb{P}^\pi(Y > \beta) = \alpha$. (H3) is recorded separately because it is the form invoked in the subsequent algebraic substitutions.)

Set $\Xi_\alpha^\pi := \Xi_\alpha(\beta; Y^\pi)$ and $\bar{c} := \max_{s,a} c(s,a)$. Under (H1)–(H3) the tower-property identity (2.13) of Section 2.1.7 applies at every t, and the decomposition $Y\,\mathbb{1}\{Y > \beta\} = (Y-\beta)^+ + \beta\,\mathbb{1}\{Y > \beta\}$ combined with Definition 2.11 and (H2) yields

$$\tfrac{1}{\alpha}\,\mathbb{E}^\pi\big[Y\,\mathbb{1}\{Y > \beta\}\big] \;=\; \tfrac{1}{\alpha}\Big[\alpha\big(\Xi_\alpha^\pi - \beta\big) \;+\; \alpha\beta\Big] \;=\; \Xi_\alpha^\pi \;=\; \mathrm{CVaR}_\alpha(Y^\pi), \tag{3.3}$$

the final equality by the minimiser property of β in Definition 2.11. Under threshold atoms the R–U variational form (Definition 2.11) remains the operative CVaR object.

The literal recursion approaches $\mathrm{CVaR}_\alpha(Y^\pi)$ for the wrong reason: freezing the terminal at Q_{T+1}^{lit} concentrates the entire $1/\alpha$-weighting in the terminal boundary, which alone supplies Ξ_α^π through (3.3); the missing $1/\alpha$ stage scaling then surfaces as a bounded residual rather than as correct telescoping.

Proposition 3.1 (Three-deviation decomposition of the literal recursion)**.** *Let*

$$\begin{aligned} Q_t^{lit}(z) \;&=\; V_t^\pi(z)\,\tilde{c}(z) \;+\; \mathbb{E}^\pi\big[Q_{T+1}^{lit}(Z_{T+1}) \,\big|\, Z_t = z\big], \\ Q_{T+1}^{lit}(z) \;&=\; \tfrac{y\,\mathbb{1}\{y>\beta\}}{\alpha}. \end{aligned} \tag{3.4}$$

Table 3.1. The three structural elements of the cost-side backwards pass at deterministic π: literal recursion of the ORMDP (Borkar and Jain, 2014) versus the Bellman-consistent operator of Definition 2.8. At $V_0^\pi(z_0) = \alpha$ the stage-scaling and bootstrap deviations jointly produce the residual $\alpha\,\tilde{c}(z_0)$, while the literal terminal independently supplies Ξ_α^π through (3.3) (Proposition 3.1).

Element	Literal Q^{lit}	Bellman-consistent Q^{corr}
Stage source	$V_t^\pi(z)\,\tilde{c}(z)$	$\frac{1}{\alpha}\,V_t^\pi(z)\,\tilde{c}(z)$
Bootstrap target	frozen at Q_{T+1}^{lit}	advanced to Q_{t+1}^{corr}
Terminal value	$y\,\mathbb{1}\{y > \beta\}/\alpha$	0

Under (H1)–(H3)*,*

$$Q_0^{lit}(z_0) \;=\; \alpha\,\tilde{c}(z_0) \;+\; \Xi_\alpha^\pi \;\in\; \big[\,\Xi_\alpha^\pi,\; \Xi_\alpha^\pi + \alpha\,\bar{c}\,\big]. \tag{3.5}$$

The recursion violates Definition 2.8: the missing $1/\alpha$ *stage scaling and the frozen bootstrap jointly generate the residual* $\alpha\,\tilde{c}(z_0)$*, while the literal terminal* $y\,\mathbb{1}\{y > \beta\}/\alpha$ *independently contributes* $(1/\alpha)\,\mathbb{E}^\pi\big[Y\,\mathbb{1}\{Y > \beta\}\big]$*, which equals* Ξ_α^π *by* (3.3)*.*

Proof. The frozen-terminal stage path does not telescope; evaluating at $t = 0$, $z = z_0$,

$$Q_0^{\text{lit}}(z_0) \;=\; V_0^\pi(z_0)\,\tilde{c}(z_0) \;+\; \mathbb{E}^\pi\big[Q_{T+1}^{\text{lit}}(Z_{T+1})\big] \tag{3.6}$$

$$\overset{\text{(H3), Def 2.3}}{=} \;\alpha\,\tilde{c}(z_0) \;+\; \tfrac{1}{\alpha}\,\mathbb{E}^\pi\big[Y^\pi\,\mathbb{1}\{Y^\pi > \beta\}\big] \tag{3.7}$$

$$\overset{(3.3)}{=} \;\alpha\,\tilde{c}(z_0) \;+\; \Xi_\alpha^\pi, \tag{3.8}$$

where (3.7) uses $Y_{T+1} = \sum_{u=0}^{T} c(S_u, A_u) = Y^\pi$ (Definition 2.3) to substitute $y = Y^\pi$ into the terminal. The interval bound follows from $0 \le \tilde{c}(z_0) \le \bar{c}$. By Definition 2.8, consistency would require $Q_0^{\text{lit}}(z_0) = \Xi_\alpha^\pi$, which holds only when $\tilde{c}(z_0) = 0$. □

The proximity of $Q_0^{\text{lit}}(z_0)$ to $\text{CVaR}_\alpha(Y^\pi)$ at the fixed point is algebraic coincidence between two of the three deviations rather than a fixed point of the intended operator: removing any single deviation in isolation destroys the apparent agreement, as the following remark records. The three deviations of Table 3.1 are not algebraically independent under Bellman consistency: bootstrap target and terminal boundary are linked through telescoping, so the repair operates on two degrees of freedom (stage-source scaling, and the bootstrap–terminal pair).

Remark 3.2 (Single-knob repairs do not suffice). Although Table 3.1 lists three structural deviations, Bellman consistency changes the bootstrap and terminal boundary jointly: an advanced bootstrap must be paired with a zero terminal boundary, since otherwise telescoping leaves a phantom contribution at $T + 1$. The relevant repairs therefore read as two independent knobs — stage scaling, and the bootstrap–

terminal pair — and each single-knob repair yields a distinct, non-CVaR value at z_0:

$$\begin{aligned} &\text{(i) } \tfrac{1}{\alpha} V_t^\pi \tilde{c} \text{ with frozen terminal:} \\ &\qquad Q_0^{(\mathrm{i})}(z_0) \;=\; \tilde{c}(z_0) \,+\, \Xi_\alpha^\pi, \\ &\text{(ii) } V_t^\pi \tilde{c} \text{ with advanced bootstrap and zero terminal:} \\ &\qquad Q_0^{(\mathrm{ii})}(z_0) \;=\; \alpha\, \Xi_\alpha^\pi. \end{aligned} \tag{3.9}$$

Only joint repair of both knobs reproduces Ξ_α^π exactly (Lemma 3.3); the surface proximity of the literal recursion to $\mathrm{CVaR}_\alpha(Y^\pi)$ thus requires the simultaneous presence of the missing $1/\alpha$ stage scaling and the frozen bootstrap. The interval $[\,\Xi_\alpha^\pi,\, \Xi_\alpha^\pi + \alpha\bar{c}\,]$ in (3.5) bounds the resulting residual over $\tilde{c}(z_0) \in [0, \bar{c}]$.

Lemma 3.3 (Corrected backwards-pass identity)**.** *Let*

$$\begin{aligned} Q_t^{corr}(z) \;&=\; \tfrac{1}{\alpha}\, V_t^\pi(z)\, \tilde{c}(z) \;+\; \mathbb{E}^\pi\big[Q_{t+1}^{corr}(Z_{t+1}) \,\big|\, Z_t = z\big], \\ Q_{T+1}^{corr}(z) \;&\equiv\; 0. \end{aligned} \tag{3.10}$$

Under (H1)–(H3)*,*

$$Q_0^{corr}(z_0) \;=\; \Xi_\alpha^\pi \;=\; \mathrm{CVaR}_\alpha(Y^\pi). \tag{3.11}$$

Proof. With $Q_{T+1}^{\mathrm{corr}} \equiv 0$, downward iteration from $t = T$ telescopes the stage terms:

$$Q_t^{\mathrm{corr}}(z) \;=\; \tfrac{1}{\alpha}\, \mathbb{E}^\pi\Big[\sum_{s=t}^{T} V_s^\pi(Z_s)\, \tilde{c}_s \,\Big|\, Z_t = z\Big]. \tag{3.12}$$

At $z = z_0$, identity (2.13) supplies $\mathbb{E}^\pi\big[\sum_{t=0}^{T} V_t^\pi(Z_t)\, \tilde{c}_t\big] = \mathbb{E}^\pi\big[Y\, \mathbb{1}\{Y > \beta\}\big]$, whence

$$Q_0^{\mathrm{corr}}(z_0) \;=\; \tfrac{1}{\alpha}\, \mathbb{E}^\pi\big[Y\, \mathbb{1}\{Y > \beta\}\big] \;\overset{(3.3)}{=}\; \Xi_\alpha^\pi. \tag{3.13}$$

□

The corrected recursion is adopted as C1 in Chapter 4; the empirical contrast with Q^{lit} appears as "paper literal" and "C1 Bellman-consistent" in Figure 5.1(a). Treatment of (R2), together with its closed-form residual (Proposition 4.2, Section 4.4), follows in Chapter 4.

4

Online CVaR Q-Learner with Q-Function–Multiplier Decoupling

Chapter 3 identified two finite-time failure modes of the literal CVaR-side recursion of Borkar and Jain (2014), each tied to a distinct role that the recursion plays. In its cost Q-function role, the literal stage source, bootstrap target, and terminal boundary do not jointly define a Bellman-consistent finite-horizon evaluation: the recursion's fixed point lies close to Ξ_α^π only by algebraic cancellation between two opposing structural errors, not by Bellman consistency (Section 3.2). In its dual-estimator role, the scalar Q_0^C inherits a multiplicative coupling to the fast-timescale V recursion that survives in finite time even after the cost Q-function itself has been repaired; the precise statement of this coupling is given in Section 4.4, once the required notation is in place.

This chapter builds a learner that addresses both modes in response to the diagnosis above. We propose three changes that retain the same asymptotic target as the original oRMDP system: one is a correction to the Bellman operator, and two are design choices that separate roles the literal recursion was carrying jointly. C1 (Section 4.2) replaces the literal cost-side update with a Bellman-consistent one. C2 (Section 4.3) runs the reward and cost Q-functions as separate Bellman recursions sharing a common bootstrap selector, recombined only at action selection. C3 (Sections 4.4–4.5) sources the dual signal from a rolling buffer of episodic returns rather than from Q_0^C itself. Section 4.1 fixes the standing assumptions and states the limiting saddle set the C1+C2+C3 system targets; Sections 4.2–4.5 develop each change in turn; Section 4.6 assembles them into Algorithm 4.1, revisits the three changes in the context of the assembled learner, and states the scope of the chapter's claims relative to the oRMDP three-timescale theorem.

4.1 Setting and the design principle

The diagnostic separation of Chapter 3 is taken as given. Under the design principle of this chapter, Q^C is retained only as a cost Q-function (C1, Section 4.2); the Lagrangian score $J = Q^R - \lambda Q^C$ is decomposed so that λ enters only through a shared bootstrap selector (C2, Section 4.3); and the dual signal is supplied separately by an empirical Rockafellar–Uryasev estimator $\widehat{\Xi}_\alpha$ on a rolling buffer of behaviour-policy returns (C3, Section 4.5), so that the scalar $Q_0^C(z_0)$ no longer serves as the dual

estimator. After C1+C2+C3, Q^C is *the cost Q-function alone*, $\widehat{\Xi}_\alpha$ is *the dual estimator alone*, and J is the action-selection score that combines the two Q-functions at arg max.

Standing assumptions for the chapter. The augmented state $Z_t = (t, S_t, Y_t)$ and the tower-property identity (2.13) are as fixed in Chapter 2. Throughout this chapter:

1. Costs and rewards are bounded and the horizon T is finite;
2. All policy-evaluation identities are stated at fixed (β, λ) and at a stationary policy π;
3. The bootstrap selector $a^\star$ is a deterministic mapping under the relevant fixed (β, λ). Under generic $Q^R - \lambda Q^C$ the arg max is unique and the issue is immaterial; when degenerate ties occur on a measure-zero set we resolve them by a fixed total order on $\mathcal{A}$, so $a^\star$ remains deterministic throughout;
4. Asynchronous Bellman identities follow the standard visitation hypotheses of Tsitsiklis (1994) (every state-action pair visited infinitely often, with the standard step-size summability);
5. Quasi-stationarity of π_b on the slow λ-timescale (Borkar, 2008) is invoked only for the limiting projected dual equilibrium and the limiting target of the buffer estimator; finite-time identities are stated under fixed (β, λ) without it;
6. Fixed-policy identities (tower identity (2.13), Proposition 4.2) hold under the deterministic selector $a^\star$ at fixed (β, λ); π denotes the deterministic stationary policy induced by $a^\star$, and we abbreviate $V_0^\pi(z_0; \beta)$ as V_0 and $\Xi_\alpha(\beta; Y^\pi)$ as $\Xi_\alpha(\beta)$ when π is unambiguous. Buffer-side statements (C3) concern π_b and reintroduce it explicitly.

The backwards Bellman-update convention — bootstrap values at index t are read within-sweep from index $t{+}1$, and within index t, V_t^k is updated before $Q_t^{C,k}$ — is fixed by Algorithm 4.1 and inherited by all identities below. Finite-time stabilisations introduced under fixed compute budgets are itemised in Section 4.6 and kept outside the C1+C2+C3 content.

Limiting target of the system. Four assumptions support the limit identification used through Sections 4.2–4.5; each is paired below with the change it supports.

(A1) The *asynchronous-Q visitation hypothesis* of Tsitsiklis (1994): every reachable state–action pair in $\mathcal{S} \times \mathcal{A}$ is visited infinitely often along the iteration. This is supplied by ϵ-greedy behaviour on J^k over the π_b-reachable subset of the augmented state–action space and supports the asynchronous-Q convergence of C1 on the modified operator.

(A2) The *three-timescale separation* $a_k \gg \gamma_k \gg \eta_k$ of Borkar (2008), with the Robbins–Monro summability $\sum_k a_k = \infty$, $\sum_k a_k^2 < \infty$ and analogously for γ_k, η_k. It supports C2 at every fixed (β, λ).

(A3) The *quantile-regularity* assumption of Rockafellar and Uryasev (2000): continuity of the return c.d.f. $F_{Y^{\pi_b}}$ at $\mathrm{VaR}_\alpha(Y^{\pi_b})$ for every stationary π_b along the iteration path.

(A4) *Buffer mixing* on a scale strictly faster than η_k, $W\eta_k \to 0$, so that the buffer estimator tracks the behaviour-policy distribution at the rate the dual update reads it. It supports C3 through sample-mean consistency on the buffer-mixing scale.

Under (A1)–(A4) the C1+C2+C3 system targets the same limiting saddle set as the original oRMDP formulation,

$$\left\{ \begin{array}{l} (\beta^\star, \lambda^\star, \pi^\star) : V_0^{\pi^\star}(\beta^\star) = \alpha, \\ \qquad\qquad\quad \Xi_\alpha(\beta^\star; Y^{\pi^\star}) \le C_\alpha, \\ \qquad\qquad\quad \lambda^\star\big(\Xi_\alpha(\beta^\star; Y^{\pi^\star}) - C_\alpha\big) = 0, \\ \qquad\qquad\quad \lambda^\star \ge 0 \end{array} \right\}. \tag{4.1}$$

A full convergence theorem and finite-time rates for the modified system are not established here; the substantive claims of this chapter are the structural identities of Propositions 4.1–4.2 together with the limit-set identification above.

4.2 C1: a Bellman-consistent cost Q-function

This section develops the Bellman-consistent recursion called for at the close of Section 3.2. The tail-probability auxiliary

$$V_t^\pi(z;\beta) = \mathbb{P}^\pi(Y > \beta \mid Z_t = z) \tag{4.2}$$

satisfies the backward recursion $V_t^\pi(z;\beta) = \sum_{z'} P^\pi(z' \mid z)\, V_{t+1}^\pi(z';\beta)$ under any stationary π, with terminal $V_{T+1}^\pi(z;\beta) = \mathbb{1}\{Y > \beta\}$ (Borkar and Jain, 2014). Write $\tilde{c}(z,a)$ for the per-step cost, indexed jointly by state and action. Under the deterministic selector $a^\star(\cdot)$ fixed in Section 4.1, the executed action obeys $A_t = a^\star(Z_t)$ on the trajectory; the realised one-step cost $\tilde{c}_t = \tilde{c}(Z_t, a^\star(Z_t))$ is therefore $\sigma(Z_t)$-measurable, and we write this realised cost as $\tilde{c}(Z_t)$ in trajectory expressions, in line with Chapter 3. Define the tail-weighted stage source

$$g_t(z,a;\beta) = \frac{1}{\alpha}\, V_t^\pi(z;\beta)\, \tilde{c}(z,a). \tag{4.3}$$

The factor $1/\alpha$ is forced by the variational form (2.8). Under the deterministic selector $\tilde{c}_t$ is $\sigma(Z_t)$-measurable, so the stage product $V_t^\pi(Z_t;\beta)\,\tilde{c}_t$ is $\sigma(Z_t)$-measurable as well; the tower-property identity (2.13) then telescopes its trajectory aggregate to

$$\frac{1}{\alpha}\,\mathbb{E}^\pi\left[\sum_{t=0}^{T} V_t^\pi(Z_t;\beta)\,\tilde{c}_t\right] = \frac{1}{\alpha}\,\mathbb{E}^\pi[Y\,\mathbb{1}\{Y > \beta\}]\,, \tag{4.4}$$

which equals $\Xi_\alpha(\beta; Y^\pi)$ on the β-recursion fixed point $V_0^\pi(\beta) = \alpha$ — equivalently $\beta = \mathrm{VaR}_\alpha(Y^\pi)$, where it further reduces to $\mathrm{CVaR}_\alpha(Y^\pi)$ — and departs from Ξ_α by an

explicit β-weighted residual otherwise (Section 4.4). Any other constant multiplier on the stage source $V_t^\pi(z;\beta)\,\tilde{c}(z,a)$ misses this fixed-point identification at $V_0^\pi(\beta) = \alpha$.

The recursion. The corrected cost Q-function update is

$$\begin{aligned} Q_t^{C,k}(z,a) = Q_t^{C,k-1}(z,a) \;+\; a_k\,\mathbb{1}\{Z_t^k{=}z,\; A_t^k{=}a\} \\ \times\big[\, g_t(z,a;\beta_k) \;+\; Q_{t+1}^{C,k}(Z_{t+1}^k, a^\star) \;-\; Q_t^{C,k-1}(z,a)\,\big], \end{aligned} \tag{4.5}$$

with terminal $Q_{T+1}^{C,k} \equiv 0$ and bootstrap selector $a^\star$ specified in Section 4.3. Three structural changes inside C1 distinguish (4.5) from the literal oRMDP recursion (Borkar and Jain, 2014), each addressing one of the three deviations identified in Section 3.2: the stage source carries the $1/\alpha$ scale, the bootstrap target advances with t, and the terminal is zero, removing the double-counting otherwise injected by $y\,\mathbb{1}\{y > \beta\}/\alpha$ at the boundary.

Analysis vs. implementation. In (4.5) the stage source $g_t(z,a;\beta_k)$ is written with the policy-evaluation tail probability $V_t^\pi(z;\beta_k)$ for the analysis of this section; in the SA implementation (Algorithm 4.1, line 10), the same recursion runs with the within-sweep iterate $V_t^k(z)$ substituted for $V_t^\pi(z;\beta_k)$, both V and Q^C being updated on the fast timescale a_k. The fast-timescale fluctuation of V_0^k around its β-recursion target α enters the slow-timescale readout Q_0^C through a closed-form factor quantified in Corollary 4.3 of Section 4.4.

Fixed-point identity. Fix a stationary π, a scalar β, and a deterministic selector $a^\star$ induced by π. Equation (4.5) is the asynchronous Robbins–Monro iteration on the linear backward Bellman policy-evaluation operator with stage source $g_t(\cdot;\beta)$ and zero terminal; under the standing assumptions of Section 4.1, the iterates converge in sup-norm (Tsitsiklis, 1994) to its unique fixed point, which by backward induction admits the closed form

$$Q_t^C(z,a) = \mathbb{E}^\pi\left[\sum_{\tau=t}^{T} \frac{1}{\alpha}\, V_\tau^\pi(Z_\tau;\beta)\,\tilde{c}_\tau \,\middle|\, Z_t = z,\; A_t = a\right]. \tag{4.6}$$

At the initial state,

$$Q_0^C(z_0;\pi,\beta) = \frac{1}{\alpha}\,\mathbb{E}^\pi\left[\sum_{t=0}^{T} V_t^\pi(Z_t;\beta)\,\tilde{c}_t\right]; \tag{4.7}$$

combining (4.7) with the tower-property identity (2.13) (Borkar and Jain, 2014, Lemma 1) gives

$$Q_0^C(z_0;\pi,\beta) = \frac{1}{\alpha}\,\mathbb{E}^\pi[Y\,\mathbb{1}\{Y > \beta\}]\,. \tag{4.8}$$

Equation (4.8) is the structural target for the cost Q-function; Section 4.4 unfolds it into the closed form that distinguishes asymptotic from finite-time behaviour.

4.3 C2: reward/cost decomposition and shared selector

Section 3.1 fixed the Q^R/Q^C representation; this section establishes its structural justification. The full assembled learner appears as Algorithm 4.1 in Section 4.6; what is established here is the linear-functional identity behind C2, on which the greedy invariance of Proposition 4.1 below depends.

Under a deterministic continuation $a^\star$, define the *stage-source-to-value functional*

$$\mathcal{Q}[p]_t(z,a) \;=\; \mathbb{E}\left[\sum_{\tau=t}^{T} p_\tau(Z_\tau, A_\tau)\,\middle|\, Z_t{=}z, A_t{=}a\right], \tag{4.9}$$

which sends a per-step payoff $p_t(z,a)$ to the policy-evaluation value at (t,z,a) along the trajectory whose first action is a and whose subsequent actions follow $A_\tau{=}a^\star(Z_\tau)$. $\mathcal{Q}[p]$ is the closed-form fixed point obtained by iterating the finite-horizon Bellman policy-evaluation operator with stage source p and continuation $a^\star$ from the zero terminal $Q_{T+1}\equiv 0$, and the map $p\mapsto\mathcal{Q}[p]$ is *linear* on the space of bounded per-step payoffs. Concretely, for any two bounded payoffs $p^{(1)}, p^{(2)}$ and any scalars c_1, c_2, the linearity identity

$$\mathcal{Q}\big[c_1\,p^{(1)} + c_2\,p^{(2)}\big] \;=\; c_1\,\mathcal{Q}\big[p^{(1)}\big] \;+\; c_2\,\mathcal{Q}\big[p^{(2)}\big]$$

holds pointwise: taking the expectation under one common trajectory law commutes with linear combinations of the per-step payoff. C2 exploits this linearity by maintaining $Q^R{=}\mathcal{Q}[r]$ and $Q^C{=}\mathcal{Q}[g(\cdot;\beta)]$ as two Bellman recursions on λ-independent stage sources, recombined only at $\arg\max$; setting $(c_1,c_2)=(1,-\lambda)$ in the identity above gives exactly the combination $Q^R-\lambda Q^C=\mathcal{Q}[r-\lambda g]$ used in Proposition 4.1.

Define the Lagrangian score and the shared bootstrap selector

$$J_t^k(z,a) = Q_t^{R,k}(z,a) - \lambda_k\,Q_t^{C,k}(z,a), \tag{4.10}$$

$$a^\star(z') \in \arg\max_a \Big[Q_{t+1}^{R,k}(z',a) - \lambda_k\,Q_{t+1}^{C,k}(z',a)\Big]. \tag{4.11}$$

The reward update is

$$Q_t^{R,k}(z,a) \;\mathrel{+}=\; a_k\Big[r_t + Q_{t+1}^{R,k}(z',a^\star) - Q_t^{R,k}(z,a)\Big], \tag{4.12}$$

the cost update is (4.5) with the same $a^\star$, and both access within-sweep $t{+}1$ values. The behaviour policy is $\pi_b = \epsilon$-greedy on J^k; the deployed greedy policy $\pi_g = \arg\max J^k$ uses the pure $\arg\max$, with the exploration–evaluation gap as in Chapter 3.

The four lines that constitute C2, each derived above, are collected here for reference:

$$\begin{cases} \textbf{(selector)} & a^\star(z') \in \arg\max_a\big[Q_{t+1}^{R,k}(z',a) - \lambda_k\,Q_{t+1}^{C,k}(z',a)\big], \\ \textbf{(reward)} & Q_t^{R,k}(z,a) \mathrel{+}= a_k\Big[r_t + Q_{t+1}^{R,k}(z',a^\star) - Q_t^{R,k}(z,a)\Big], \\ \textbf{(cost)} & Q_t^{C,k}(z,a) \mathrel{+}= a_k\Big[g_t(z,a;\beta_k) + Q_{t+1}^{C,k}(z',a^\star) - Q_t^{C,k}(z,a)\Big], \\ \textbf{(score)} & J_t^k(z,a) = Q_t^{R,k}(z,a) - \lambda_k\,Q_t^{C,k}(z,a). \end{cases} \tag{4.13}$$

The multiplier λ_k enters only through (selector) and (score); the per-step stage sources r_t and $g_t(\cdot;\beta_k)$ in (reward) and (cost) carry no λ_k dependence. The cost stage source $g_t(\cdot;\beta_k)$ is the C1 source (4.3); the zero terminal $Q^{C,k}_{T+1} \equiv 0$ of C1 is inherited.

The shared-selector requirement. Selectors must agree across the two Q-functions. Linearity in (4.9) is a property of a *single* functional defined under a *single* continuation rule. Under independent selectors $a^\star_R \neq a^\star_C$, the reward and cost recursions roll out under different action sequences from every state and therefore measure expectations under different trajectory distributions, producing distinct functionals $\mathcal{Q}_R$ and $\mathcal{Q}_C$; their combination $\mathcal{Q}_R[r] - \lambda\,\mathcal{Q}_C[g]$ admits no representation $\mathcal{Q}[r - \lambda g]$ under any single continuation, and the linearity identity used in Proposition 4.1 becomes unavailable. The shared $a^\star$ collapses the two functionals to one and reinstates $p \mapsto \mathcal{Q}[p]$ as a linear map on a common probability space.

λ enters only through the bootstrap selector. The per-step stage sources r_t and $g_t(\cdot;\beta_k)$ carry no dependence on λ_k; the multiplier enters the Bellman target solely through the selector $a^\star$ — that is, through *which* neighbouring Q-value is accessed at the bootstrap, not through the value of the stage source itself. This confines the λ-coupling to the projection's continuation rule and leaves the per-step accumulation λ-free; under the three-timescale separation $a_k \gg \gamma_k \gg \eta_k$ (Borkar, 2008) the residual selector coupling operates under quasi-stationary λ. The contrast object is the combined single-recursion Lagrangian J^{combined} of Proposition 4.1, which stores J directly with per-step target $\rho_t(\cdot;\beta,\lambda) = r_t - \lambda\, g_t(\cdot;\beta)$, in which λ rescales the cost component at every Bellman step: one Bellman recursion must track a target whose magnitude itself moves on the slow timescale. C2 confines that movement to the bootstrap selector at each Bellman read, rather than allowing λ to rescale the stage source itself.

Proposition 4.1 (Greedy invariance under a shared selector)**.** *Fix $\lambda \geq 0$ and a deterministic selector $a^\star$. Let Q^R, Q^C denote the policy-evaluation fixed points of* (4.12) *and* (4.5) *under $a^\star$, and let J^{combined} denote the fixed point of the same recursion applied to the combined stage source $\rho_t(z,a;\beta) = r_t - \lambda\, g_t(z,a;\beta)$ under the same $a^\star$. Then $Q^R_t(z,a) - \lambda\, Q^C_t(z,a) = J^{\text{combined}}_t(z,a)$ pointwise, and the induced greedy policies coincide.*

Proof. By (4.9), $Q^R = \mathcal{Q}[r]$, $Q^C = \mathcal{Q}[g(\cdot;\beta)]$, and $J^{\text{combined}} = \mathcal{Q}[r - \lambda g(\cdot;\beta)]$ under the common $a^\star$ and the shared zero terminal. Linearity of $p \mapsto \mathcal{Q}[p]$ gives $\mathcal{Q}[r - \lambda g] = \mathcal{Q}[r] - \lambda\,\mathcal{Q}[g] = Q^R - \lambda Q^C$ pointwise; equal action-value functions induce equal arg max sets. □

Scope. Proposition 4.1 is a representation-equivalence result at fixed λ and fixed $a^\star$, not a convergence statement for the moving three-timescale system. It extends Q-decomposition (Russell and Zimdars, 2003) from the Bellman *optimality* operator — non-linear in the per-step payoff, where decomposition requires component maximisers to agree — to the finite-horizon policy-*evaluation* operator under a

fixed deterministic selector, where linearity of (4.9) delivers pointwise value equality, a stronger statement than equality of greedy sets alone. Embedded in the full three-timescale system, the per-step argument inherits the inner-loop convergence of Tsitsiklis (1994) at frozen (β, λ) and the outer timescales of Borkar (2008).

4.4 From cost Q-function correctness to dual-side noise

Proposition 4.2 below establishes the closed-form residual of the dual-estimator failure mode previewed at the close of Section 3.1. C1 establishes Q^C as a Bellman-consistent cost Q-function; it does not establish that Q_0^C should serve as the scalar constraint estimate consumed by the slow λ-update. The closed form below explains why.

Proposition 4.2 (Closed-form residual at the cost Q-function fixed point)**.** *Let π be a stationary deterministic policy on the augmented state space — equivalently, the deterministic bootstrap selector held fixed during policy evaluation — and let β be a fixed scalar. Let $V_0^\pi(\beta) = \mathbb{P}^\pi(Y > \beta)$ and recall $\Xi_\alpha(\beta; Y^\pi) = \beta + \alpha^{-1}\mathbb{E}^\pi[(Y-\beta)^+]$. At the policy-evaluation fixed point of (4.5) under π and β,*

$$Q_0^C(z_0; \pi, \beta) = \Xi_\alpha(\beta; Y^\pi) \; - \; \beta\left(1 - \frac{V_0^\pi(\beta)}{\alpha}\right). \tag{4.14}$$

Proof. Telescoping (4.5) along the trajectory under π with $Q_{T+1}^C \equiv 0$ gives the policy-evaluation fixed point in expanded form,

$$Q_0^C(z_0; \pi, \beta) = \frac{1}{\alpha}\,\mathbb{E}^\pi\left[\sum_{t=0}^{T} V_t^\pi(Z_t; \beta)\,\tilde{c}(Z_t)\right].$$

For each t, $\tilde{c}(Z_t)$ is $\sigma(Z_t)$-measurable under the deterministic selector (Section 4.2), and $V_t^\pi(Z_t; \beta) = \mathbb{E}^\pi[\,\mathbb{1}\{Y > \beta\} \mid Z_t]$ by (4.2). The tower property gives

$$\mathbb{E}^\pi[V_t^\pi(Z_t; \beta)\,\tilde{c}(Z_t)] = \mathbb{E}^\pi[\tilde{c}(Z_t)\,\mathbb{E}^\pi[\mathbb{1}\{Y > \beta\} \mid Z_t]] = \mathbb{E}^\pi[\tilde{c}(Z_t)\,\mathbb{1}\{Y > \beta\}]\,.$$

Summing over t and using $\sum_{t=0}^{T} \tilde{c}(Z_t) = Y$ (Borkar and Jain, 2014, Lemma 1) yields

$$Q_0^C(z_0; \pi, \beta) = \frac{1}{\alpha}\,\mathbb{E}^\pi[Y\,\mathbb{1}\{Y > \beta\}]\,.$$

The decomposition $Y\,\mathbb{1}\{Y > \beta\} = (Y-\beta)^+ + \beta\,\mathbb{1}\{Y > \beta\}$ gives

$$Q_0^C(z_0; \pi, \beta) = \frac{1}{\alpha}\,\mathbb{E}^\pi\big[(Y-\beta)^+\big] + \frac{\beta}{\alpha}\,\mathbb{P}^\pi(Y > \beta);$$

the first term equals $\Xi_\alpha(\beta; Y^\pi) - \beta$ by (2.8) and the second equals $\beta\,V_0^\pi(\beta)/\alpha$. Combining,

$$\begin{aligned} Q_0^C(z_0; \pi, \beta) &= \Xi_\alpha(\beta; Y^\pi) - \beta \; + \; \frac{\beta\,V_0^\pi(\beta)}{\alpha} \\ &= \Xi_\alpha(\beta; Y^\pi) \; - \; \beta\left(1 - \frac{V_0^\pi(\beta)}{\alpha}\right). \end{aligned}$$

□

The corollary below is a finite-k rewriting of the deterministic identity in Proposition 4.2, obtained by substituting the SA iterate $V_0^{(k)}$ for the deterministic value $V_0^{\pi}(\beta)$ at frozen (π, β_k); subsequent use of (4.15) under slowly moving π_b is a local finite-time diagnostic, not a fixed-point statement.

Corollary 4.3 (Finite-time amplification scale)**.** *Let $V_0^{(k)} = \alpha + \delta_V^{(k)}$ denote the fast-timescale iterate around its β-recursion target. At any fixed β_k and stationary deterministic policy π in the sense of Proposition 4.2, the closed-form readout* (4.14) *of Q_0^C at z_0 admits the equivalent form*

$$Q_0^{C,(k)} \;=\; \Xi_\alpha(\beta_k; Y^\pi) \;+\; \frac{\beta_k\, \delta_V^{(k)}}{\alpha}. \tag{4.15}$$

The statement is the algebraic re-expression of Proposition 4.2 obtained by writing $V_0^{\pi}(\beta)$ as $\alpha + \delta_V^{(k)}$; its interpretation as a numerical diagnostic on the actual SA iterate requires the fast-timescale recursion to be near its policy-evaluation value, a heuristic regime separate from the fixed-point statement of Proposition 4.2.

Proof. Substitute $V_0^{(k)} = \alpha + \delta_V^{(k)}$ into (4.14):

$$-\beta_k\left(1 - \frac{V_0^{(k)}}{\alpha}\right) \;=\; -\beta_k\left(1 - 1 - \frac{\delta_V^{(k)}}{\alpha}\right) \;=\; \frac{\beta_k\, \delta_V^{(k)}}{\alpha}. \qquad \square$$

What the corollary says numerically. Equation (4.15) is an algebraic identity: the slow-timescale readout $Q_0^{C,(k)}$ equals its target $\Xi_\alpha(\beta_k; Y^\pi)$ plus the fast-timescale error $\delta_V^{(k)}$ scaled by the operating-point factor β_k/α. The factor is determined by the operating point alone, not by the step-size schedule. At the single-agent calibration of Chapter 5, $\alpha = 0.10$ and the post-gate Polyak (rolling-window) mean $\bar{\beta} \approx 1.44$ give $\bar{\beta}/\alpha \approx 14$, so a fast-timescale deviation $\delta_V \sim \alpha$ — well inside the asymptotic-tightness target of the V recursion — produces a Q_0^C deviation of magnitude $\beta_k\, \delta_V/\alpha \approx 14\, \delta_V$, i.e. roughly fourteen times the fast-timescale fluctuation itself (numerically ≈ 1.4 in cost units when $\delta_V \sim \alpha = 0.10$). This is the structural reason an asymptotically aligned cost Q-function cannot, on its own, serve as the slow-timescale dual estimator. The $\bar{\beta}/\alpha$ values realised across the centralised stress test of Chapter 7 are reported there.

Asymptotic alignment, finite-time amplification. At the β-recursion fixed point $V_0^{\pi}(\beta) = \alpha$ the parenthesised factor in (4.14) is zero and $Q_0^C = \Xi_\alpha(\beta; Y^\pi)$; combined with $\beta = \mathrm{VaR}_\alpha(Y^\pi)$ at the same fixed point and the (A3) continuity of F_{Y^π} at VaR_α, this gives $Q_0^C = \mathrm{CVaR}_\alpha(Y^\pi)$: the repaired cost Q-function is asymptotically aligned with the constraint quantity. At finite k, with $a_k \gg \gamma_k \gg \eta_k$ (Borkar and Jain, 2014; Borkar, 2008), $V_0^{(k)}(\beta_k)$ oscillates around α. Substituting $V_0^{(k)} = \alpha + \delta_V^{(k)}$ and reading Y^π as the behaviour-policy return Y^{π_b} gives the local diagnostic

$$Q_0^{C,(k)} \;\approx\; \Xi_\alpha(\beta_k; Y^{\pi_b}) \;+\; \frac{\beta_k\, \delta_V^{(k)}}{\alpha}\,: \tag{4.16}$$

the fast-timescale fluctuation enters the dual gradient $Q_0^{C,(k)} - C_\alpha$ multiplied by β_k/α before reaching the slow update, with no amplitude-reducing operation between the two timescales. The mechanism is finite-time variance amplification through a closed-form factor on an otherwise correct cost Q-function; the remedy lies on the dual side, with the dual signal sourced from a separate estimator that does not inherit this amplification (Section 4.5).

4.5 C3: a Rockafellar–Uryasev buffer estimator

This section introduces the separate dual estimator motivated by the closed-form readout established in Section 4.4 (Proposition 4.2, Corollary 4.3). Maintain a rolling buffer $\mathcal{B}$ of the most recent episodic returns generated under the executed behaviour policy π_b, with capacity W. At each λ-update time, form the empirical Rockafellar–Uryasev estimator

$$\widehat{\Xi}_\alpha(\mathcal{B}) = \widehat{\beta} + \frac{1}{\alpha\,|\mathcal{B}|} \sum_{Y_i \in \mathcal{B}} (Y_i - \widehat{\beta})^+, \qquad \widehat{\beta} = \widehat{\mathrm{VaR}}_\alpha(\mathcal{B}), \tag{4.17}$$

the variational form (2.8) evaluated at the plug-in VaR_α threshold of $\mathcal{B}$ in the convention of (2.6), equivalently the standard sample CVaR_α at level α (Rockafellar and Uryasev, 2000; Chow and Ghavamzadeh, 2014; Tamar et al., 2015). The empirical threshold $\widehat{\beta}$ is computed from $\mathcal{B}$ alone, distinct from the recursive iterate β_k that underlies the C1-sourced cost estimator. The dual update is

$$\lambda_{k+1} = \Big[\, \lambda_k + \eta_k \Big(\widehat{\Xi}_\alpha(\mathcal{B}_k) - C_\alpha \Big) \,\Big]_+ . \tag{4.18}$$

Limiting target and dual equilibrium. Under quasi-stationarity of π_b on the slow λ-timescale (Borkar, 2008), π_b is approximately fixed on the buffer's mixing scale. In the ideal large-buffer limit, where the empirical return distribution on $\mathcal{B}$ approximates the return distribution induced by the current behaviour policy, (4.17) targets $\mathrm{CVaR}_\alpha(Y^{\pi_b})$. By Proposition 4.2 and the β-recursion fixed point, the C1 readout obeys $Q_0^C \to \mathrm{CVaR}_\alpha(Y^{\pi_b})$ in the same limit. The two estimators target the same limiting quantity, and the substitution leaves the limiting projected dual equilibrium condition unchanged, namely $\lambda_\infty \geq 0$, $\mathrm{CVaR}_\alpha(Y^{\pi_b}) - C_\alpha \leq 0$, and $\lambda_\infty \big(\mathrm{CVaR}_\alpha(Y^{\pi_b}) - C_\alpha\big) = 0$.[1]

Finite-time interpretation. The two estimators differ at finite time. The Q_0^C-sourced signal carries the V_0/α residual of (4.16), a multiplicative coupling to the fast timescale. The buffer-sourced signal (4.17) is a sample mean of episodic returns. With π_b treated as fixed within each episode, the returns $\{Y_i\}_{i \leq |\mathcal{B}|}$ are independent draws across episodes by the Markov property; on the buffer-mixing scale of (A4), where π_b moves slowly relative to W, they are also approximately identically distributed.

[1] On the unshocked return distribution treated in this chapter; under the episode-end smoothing shock ξ of Section 6.2 the two targets agree up to the $O(\sigma_\xi/\alpha)$ cross-tail residual of (6.3).

The Rockafellar–Uryasev variational form is stationary in β at the population VaR_α, so by the envelope theorem the plug-in $\widehat{\beta}$ contributes only at second order to the asymptotic variance. The leading $O(|\mathcal{B}|^{-1})$ term equals the oracle CLT variance $\mathrm{Var}((Y-\beta_\alpha)^+)/(\alpha^2|\mathcal{B}|)$ (Trindade et al., 2007); the buffer signal therefore does not inherit the V_0/α amplification.

Under bounded returns $|Y| \le M$ a McDiarmid argument with bounded-differences constant $2M/(\alpha|\mathcal{B}|)$ (Brown, 2007) yields

$$\mathbb{P}\big(\,|\widehat{\Xi}_\alpha(\mathcal{B}) - \mathbb{E}\widehat{\Xi}_\alpha(\mathcal{B})| \ge t\,\big) \;\le\; 2\exp\!\left(-\frac{|\mathcal{B}|\,\alpha^2\,t^2}{2\,M^2}\right). \tag{4.19}$$

This applies directly under the bounded smoothing alternatives of Chapter 6 ($\xi \sim U[-\delta,\delta]$, clipped Gaussian); under the unbounded Gaussian shock used in the experiments it serves as a high-probability diagnostic after truncation at exponentially small tail mass. The plug-in bias is $O((|\mathcal{B}|\alpha)^{-1})$ (Trindade et al., 2007), additive and free of V_0/α coupling, so the structural removal of the amplification mechanism of Corollary 4.3 is intact independently of this finite-$|\mathcal{B}|$ correction.

At finite W with a moving π_b, $\widehat{\Xi}_\alpha(\mathcal{B})$ is a *local behaviour-policy estimator*: it targets the CVaR of the empirical return distribution generated by recent values of π_b on the buffer's mixing scale, and carries finite-W bias by design. Under departures from buffer-scale quasi-stationarity of π_b the bias picks up a component bounded by the worst pairwise step-level total variation of π_b across the window (Section 8.3). The substitution removes the V_0/α amplification of (4.16); the residual finite-sample variance is bounded by (4.19).

The cost Q-function and the dual estimator are now sourced from different objects (contribution C3); Q^C continues to drive greedy action selection through $J = Q^R - \lambda Q^C$.

4.6 The assembled algorithm

The three changes combine into a single online learner. Algorithm 4.1 assembles in one episode loop the Bellman-consistent cost Q-function of Section 4.2 (C1), the Q^R/Q^C decomposition with shared bootstrap selector of Section 4.3 (C2), and the Rockafellar–Uryasev buffer estimator of Section 4.5 (C3). The action-value tables Q^R and Q^C are indexed by (t,z,a) with $z=(t,s,y)$; the tail-probability auxiliary V by (t,z); the buffer $\mathcal{B}$ holds episodic returns under the executed behaviour policy π_b.

Mathematical core and finite-time stabilisations. Lines 3–12 implement C1+C2 with the objects of Sections 4.2–4.3; lines 13–17 implement C3 with the estimator of Section 4.5 (lines 18–20 cover the buffer-activation else branch). The mathematical content of the algorithm is exhausted by these lines together with the step-size ordering of Borkar (2008). Implementation stabilisations introduced to keep finite-time runs well-behaved under fixed compute budgets — a cascaded warm-up, sub-sampled dual-update cadence, a dual-gradient soft clip, rolling-window Polyak

averaging used for evaluation only, and slack-driven recovery — are documented in detail in Section 6.5. None alters the C1+C2+C3 Bellman recursions or the C3 buffer estimator; they act on finite-time training dynamics, read-out variance, or the realised dual-ascent trajectory.

Episodic return Y^k. The episodic return Y^k defined at line 3 of Algorithm 4.1 is the rollout sum supplied by the environment, consumed at line 4 for the terminal indicator $V_{T+1}^k = \mathbb{1}\{Y^k > \beta_k\}$ and at line 13 for the buffer append. The centralised stress-test environment of Chapter 6 (Section 6.2) adds an episode-end smoothing shock ξ to this sum before it enters lines 4 and 13, with the cost-side stage source at line 10 unaffected; the algorithmic content of Algorithm 4.1 is otherwise unchanged.

Algorithm 4.1 Online CVaR Q-Learner with Q-Function–Multiplier Decoupling.

Require: risk level $\alpha \in (0,1)$; CVaR budget C_α (under the centralised lift of Chapter 6, C_α is replaced by $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ in line 17; see Section 6.1); horizon T; step-size schedules $\{a_k\}, \{\gamma_k\}, \{\eta_k\}$ with $a_k \gg \gamma_k \gg \eta_k$; buffer capacity W and activation threshold $W_{\min}$; dual clip $\lambda_{\max}$; exploration rate ϵ; initial threshold β_0.

Ensure: reward Q-function Q^R, cost Q-function Q^C, dual variable λ; deployed greedy policy $\pi_g(z) = \arg\max_a \big[Q^R(z,a) - \lambda\, Q^C(z,a)\big]$.

1: $Q^R, Q^C, V \leftarrow 0$; $\beta_1 \leftarrow \beta_0$; $\lambda_1 \leftarrow 0$; $\mathcal{B} \leftarrow \varnothing$
2: **for** $k = 1, 2, \ldots$ **do**
▷ *Rollout under $\pi_b = \epsilon$-greedy on $J_t = Q_t^R - \lambda_k Q_t^C$.*
3: Sample $Z_0^k \to Z_{T+1}^k$; record $\{A_t^k, \tilde{c}_t, r_t\}_{t=0}^T$ and $Y^k = \sum_{t=0}^T \tilde{c}_t$
4: $V_{T+1}^k \leftarrow \mathbb{1}\{Y^k > \beta_k\}$, $Q_{T+1}^{R,k} \leftarrow 0$, $Q_{T+1}^{C,k} \leftarrow 0$
▷ *Backwards Bellman update on the fast timescale a_k.*
5: **for** $t = T, T-1, \ldots, 0$ **do**
6: $(z, a, z') \leftarrow (Z_t^k, A_t^k, Z_{t+1}^k)$
7: $a^\star \leftarrow \arg\max_{a'} \big[Q_{t+1}^{R,k}(z', a') - \lambda_k Q_{t+1}^{C,k}(z', a')\big]$ ▷ shared selector (C2)
8: $V_t^k(z) \mathrel{+}= a_k \big[V_{t+1}^k(z') - V_t^k(z)\big]$
9: $Q_t^{R,k}(z,a) \mathrel{+}= a_k \big[r_t + Q_{t+1}^{R,k}(z', a^\star) - Q_t^{R,k}(z,a)\big]$ ▷ reward Q-function
10: $Q_t^{C,k}(z,a) \mathrel{+}= a_k \big[\alpha^{-1} V_t^k(z)\, \tilde{c}(z,a) + Q_{t+1}^{C,k}(z', a^\star) - Q_t^{C,k}(z,a)\big]$ ▷ cost Q (C1)
11: **end for**
▷ *Threshold update on the intermediate timescale γ_k.*
12: $\beta_{k+1} \leftarrow \beta_k - \gamma_k \big(\alpha - V_0^k(z_0)\big)$
▷ *Dual update on the slow timescale η_k.*
13: Append Y^k to $\mathcal{B}$ (rolling capacity W)
14: **if** $|\mathcal{B}| \geq W_{\min}$ **then**
15: $\widehat{\beta}_k \leftarrow \widehat{\mathrm{VaR}}_\alpha(\mathcal{B})$
16: $\widehat{\Xi}_\alpha(\mathcal{B}) \leftarrow \widehat{\beta}_k + \dfrac{1}{\alpha\,|\mathcal{B}|} \sum_{Y_i \in \mathcal{B}} (Y_i - \widehat{\beta}_k)^+$ ▷ R–U estimator
17: $\lambda_{k+1} \leftarrow \Pi_{[0,\lambda_{\max}]}\big(\lambda_k + \eta_k \big[\widehat{\Xi}_\alpha(\mathcal{B}) - C_\alpha\big]\big)$ ▷ projected dual ascent (C3)
18: **else**
19: $\lambda_{k+1} \leftarrow \lambda_k$
20: **end if**
21: **end for**
22: **return** Q^R, Q^C, λ

Notation conventions for Algorithm 4.1. Five conventions used in the pseudocode below require flagging here. (i) The horizon T denotes the final decision index, with the terminal boundary placed at $t{=}T{+}1$; together with the backwards Bellman-update convention of Section 4.1, this fixes the time indexing for all bootstrap reads. (ii) Two thresholds appear and must not be conflated: the learner's intermediate-timescale iterate β_k, used in the V_{T+1} indicator at line 4 and updated at line 12; and the buffer-side empirical VaR_α, written $\widehat{\beta}$ (the per-episode instantiation of the estimator in Eq. (4.17)), used only inside the Rockafellar–Uryasev sample mean at line 15. The hat is the disambiguator. (iii) The within-loop binding $(z, a, z') \leftarrow (Z_t^k, A_t^k, Z_{t+1}^k)$ at line 6 confines all subsequent reads and updates within the inner loop to the realised augmented cell of the rollout; this implicit-binding form is equivalent to the explicit indicator $\mathbb{1}\{Z_t^k = z, A_t^k = a\}$ used in Algorithm 2.1, and the two notational styles describe the same asynchronous-SA semantics. (iv) The deployed greedy policy π_g in the Ensure clause is defined with the same λ symbol as the operational Bellman path; in the experimental reports of Chapters 6–8, π_g is evaluated under the Polyak-averaged $\bar{\lambda}$ of Section 6.5, a finite-time stabilisation outside the C1+C2+C3 content. (v) The compound-assignment operator $+=$ used in lines 8–10 abbreviates the stochastic-approximation update $x \leftarrow x + a_k \left[\,\cdot\,\right]$ in standard form (compare lines 4, 6, 7 of Algorithm 2.1, which write the same recursion out fully). The step-size schedule and the specific Taxi-MDP instantiation are fixed in Sections 6.2–6.6; the five-component KKT-based operational gate used to declare convergence in Chapters 5–7 is defined in Section 5.3, and it is not part of Algorithm 4.1's specification.

The three changes in retrospect. With Algorithm 4.1 on the page, the three changes act on the policy-evaluation fixed point itself. C1 moves the inner-loop fixed point from the literal recursion's value in $[\Xi_\alpha^\pi,\ \Xi_\alpha^\pi + \alpha\bar{c}]$ (3.5) — sustained by the cancellation of two opposing structural errors — to the Bellman-consistent readout of (4.8), whose residual is given by Proposition 4.2. C2 imposes the shared bootstrap selector of Proposition 4.1, which makes the per-step Bellman targets λ-free in their stage sources. C3 replaces $Q_0^C(z_0)$, with its multiplicative V_0/α coupling (4.16), by a sample-mean Rockafellar–Uryasev estimator carrying $O(W^{-1})$ variance and no fast-timescale amplification. The fixed points are determined by C1, C2, and C3; the trajectory to each is determined by $(a_k, \gamma_k, \eta_k, \lambda_{\max}, W)$.

Scope of the chapter's claims. Because C1 changes the inner-loop Bellman operator and C2 separates the action-value representation, the three-timescale theorem of Borkar and Jain (2014, Theorem 2) is not inherited verbatim. The three changes act at distinct levels of that proof: C1 substitutes the inner-loop operator (asynchronous-Q convergence (Tsitsiklis, 1994) applies to the corrected operator at frozen (β, λ)); C2 substitutes the inner-loop projection (Proposition 4.1 preserves greedy invariance at every fixed λ); C3 alters the slow-timescale driver — the only level at which the oRMDP proof ceases to apply verbatim — and adds the buffer-mixing condition $W\eta_k \to 0$ of (A4), under which the limiting projected dual equilibrium is preserved. A full convergence theorem and finite-time rates are not established here.

Table 4.1 records the structural differences between the literal CVaR-side recursion of the oRMDP (Borkar and Jain, 2014) and the learner of this chapter, grouped by the change targeting each row. Under quasi-stationarity of π_b (Borkar, 2008), both schemes target the same limiting constraint quantity $\mathrm{CVaR}_\alpha(Y^{\pi_b})$ at z_0; the entries record finite-time-structural differences only, located in the Bellman consistency of the cost recursion and in the architecture of the dual signal. Analytical support is in the cited sections: Section 3.2 and Equations (3.5)–(3.11) for the literal recursion's proximity-by-cancellation; Section 4.5 for the oracle first-order asymptotic of the Rockafellar–Uryasev sample mean (Brown, 2007; Trindade et al., 2007); and Proposition 4.2 for the closed-form residual tying the two.

Table 4.1. Structural changes from the literal oRMDP recursion to the C1+C2+C3 learner of this chapter. C1 is a correction to the Bellman operator; C2 and C3 are design choices that separate the roles the literal recursion was carrying jointly. Analytical content and citations supporting each row are given in the preceding paragraph and in the cited sections.

Component	**Literal** oRMDP	**This chapter (C1+C2+C3)**
Cost Q-function recursion — C1 (Section 4.2)		
Stage source	$V_t\,\tilde{c}$	$\alpha^{-1}\,V_t\,\tilde{c}$
Bootstrap target	Q_{T+1} (terminal-frozen)	Q_{t+1} (time-shifted, within-sweep)
Terminal boundary	$y\,\mathbb{1}\{y>\beta\}/\alpha$	$Q^C_{T+1}\equiv 0$
Fixed point at z_0	not Bellman-consistent (Section 3.2)	$\alpha^{-1}\mathbb{E}^\pi[Y\,\mathbb{1}\{Y>\beta\}]$ (Eq. 4.8); equals Ξ^π_α at $\beta{=}\mathrm{VaR}_\alpha(Y^\pi)$
Action-value structure and bootstrap selector — C2 (Section 4.3)		
Q-function representation	single Lagrangian score J	separate Q^R, Q^C
Bootstrap selector	implicit in literal score	shared $a^\star\in\arg\max_a[Q^R_{t+1}-\lambda Q^C_{t+1}]$
Bellman-target dependence on λ	λ in single-table terminal J_{T+1}; per-step source r	λ-independent per-step sources r_t, g_t; λ enters only via $a^\star$ (Prop. 4.1)
Dual estimator — C3 (Sections 4.4–4.5)		
Dual signal	$Q^C_0(z_0)$	R–U sample mean $\widehat{\Xi}_\alpha(\mathcal{B}_k)$ (Eq. 4.17)
Finite-time noise mechanism	multiplicative coupling $\beta\big(1-V_0^\pi/\alpha\big)$ to fast V recursion	sample-mean variance $\mathrm{Var}\big((Y-\beta_\alpha)^+\big)/(\alpha^2\lvert\mathcal{B}\rvert)$; no V_0/α amplification
Role of Q^C	conflated: Q-function and dual estimator	cost Q-function only

5

Single-Agent Validation

Chapter 4 produced three structural claims: a Bellman-consistent tail-weighted cost Q-function (Section 4.2); a Q^R/Q^C decomposition whose shared selector preserves the greedy policy of the combined Lagrangian Q-function at any $\lambda \geq 0$ (Proposition 4.1); and a sample-mean dual signal without the V_0/α amplification of the Q_0^C readout, leaving the limiting projected dual equilibrium condition unchanged (Proposition 4.2, Section 4.5). This chapter validates each claim on a single-agent test bed, fixing the empirical ground from which Chapter 6 departs into the centralised lift.

The chapter is structured as an ablation study of the C1+C2+C3 specification followed by two finite-time characterisations. Section 5.2 is the ablation: it applies Section 3.2's unrolling diagnostic at an operating point off the β-recursion fixed point, separating the C1 (Bellman-consistent) recursion from the paper-literal variant, and contrasting the dual-signal effect of Q_0^C versus the C3 buffer estimator under fixed $\beta \neq \mathrm{VaR}_\alpha$. Sections 5.3 and 5.4 characterise the operating envelope of the assembled learner: the first reports five-seed dispersion of the operational gate at a fixed calibration point; the second stress-tests Proposition 4.2 across a calibrated (α, ρ) grid and examines the empirical scaling of the finite-window residual against the $1/\alpha$ envelope the closed form predicts. Section 5.5 hands off to Chapter 6.

5.1 The single-agent test bed and its multi-agent extension

Per-agent MDP. The single-agent test bed of this chapter, and the per-agent template lifted to N agents in Chapter 6, is the same finite-horizon time-indexed MDP $\mathcal{M}_0 = \langle \mathcal{S}_0, \mathcal{A}_0, (P_{0,t}), (r_{0,t}), (c_{0,t}), T \rangle$. The subscript 0 marks the *per-agent template* common to all agents in the multi-agent lift of Chapter 6 — not an agent index — and is kept distinct from agent-specific instantiations (s_i, a_i) and the joint tags $\mathcal{S}^{(N)}, \mathcal{A}^{(N)}$ used there. The components are:

- Local lane state space $\mathcal{S}_0 = \{0, 1, 2\}$, where lanes 0 and 2 are shoulders and lane 1 is the centre; $|\mathcal{S}_0| = 3$. The column is a shared finite-horizon time index, carried separately from the lane state. Initial lane state $s_0 = 1$ at column 0.
- Local action space $\mathcal{A}_0 = \{\texttt{up}, \texttt{stay}, \texttt{down}\}$, $|\mathcal{A}_0| = 3$.

- Local transition $P_{0,t}$. The shared column/time index advances deterministically from t to $t+1$ until the finite horizon terminates. The lane evolves under a hazard-band-conditioned kernel: outside the hazard band the intended lane is reached deterministically; inside the hazard band the centre lane is sticky and the shoulder lanes are biased towards the centre. Boundary clipping at $s \in \{0, 2\}$ saturates the lane index at the boundary and the state remains non-terminal. The full kernel is in Appendix A.
- Local reward $r_{0,t}(s, a)$: a per-step lane-indexed reward keyed to the post-action intended lane, with the centre lane the unique reward maximiser at every step. The terminal step pays the centre-lane reward in full and zero on the shoulders, so ending in the centre is the unique terminal reward maximiser. Numerical values in Appendix A.
- Local per-step cost $c_{0,t}(s, a) = c_t^{\mathrm{haz}}(s) + c^{\mathrm{sw}}(a)$, with c^{haz} a column-indexed hazard ramp on the centre lane within the central column band (zero on the shoulders; maximum value 0.40; full ramp in Appendix A) and $c^{\mathrm{sw}}(a) = 0.10 \cdot \mathbb{1}\{a \neq \texttt{stay}\}$; per-step cost bounded by $\bar{c} = 0.50$.
- Horizon $T = 12$ decisions. Decision columns are $t \in \{0, 1, \ldots, T-1\}$ and the terminal column index is T. Per-step quantities accrue at every decision, including the terminal reward step; the per-agent state has no absorbing goal.

Physical setting and operational criterion. The single-agent test bed instantiates $\mathcal{M}_0$ as a single-agent CVaR-CMDP with risk level α and CVaR budget C_α; the episodic return is $Y = \sum_{t=0}^{T-1} \tilde{c}_t$. Throughout this chapter the held-out evaluator estimates $\mathrm{CVaR}_\alpha(Y^{\pi_g})$ under the deployed greedy policy π_g, while the algorithmic dual update is driven by the behaviour-policy buffer estimator $\widehat{\Xi}_\alpha^{\mathcal{B}}$ of Section 4.5. Operational success is reported on the deployed-greedy feasibility criterion $\mathrm{CVaR}_\alpha(Y^{\pi_g}) \leq C_\alpha$ at the calibrated budget C_α specified in Section 5.3.

Relation to the multi-agent test bed. Chapter 6 (Section 6.2) lifts the per-agent template $\mathcal{M}_0$ defined above to $N{>}1$ agents on the joint state–action space $\mathcal{S}^{(N)} = \mathcal{S}_0^N$, $\mathcal{A}^{(N)} = \mathcal{A}_0^N$, with a shared CVaR budget $C_\alpha^{\mathrm{eff}} = N \cdot C_{\mathrm{base}}$ and an episode-end smoothing shock $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ applied to the joint return before it enters the buffer (Section 6.2); the per-agent cost structure $c_{0,t}$ is preserved without modification. The single-agent test bed of this chapter corresponds to the $N{=}1$ instance of that lift, with ξ retained for bit-identical estimator behaviour across chapters. The buffer estimator $\widehat{\Xi}_\alpha^{\mathcal{B}}$, the per-cell visit statistics, and the held-out CVaR evaluator are therefore the same objects in this chapter as in Chapter 7's $N{=}1$ cell; the two chapters differ only in the diagnostic question each asks — this chapter validates the architecture, Chapter 7 stresses it across N.

5.2 Validation of C1 and C3: the Q_0^C diagnostic

Figure 5.1 validates two predictions, but they are not the same prediction: panel (a) is about the Q-function side and panel (b) is about the constraint side. Reading them separately is essential, because the central question raised by the diagnostic of Chapter 3 — does the literal recursion produce a learner that respects the CVaR constraint at finite time — is answered on the dual side, not on the Q-function side.

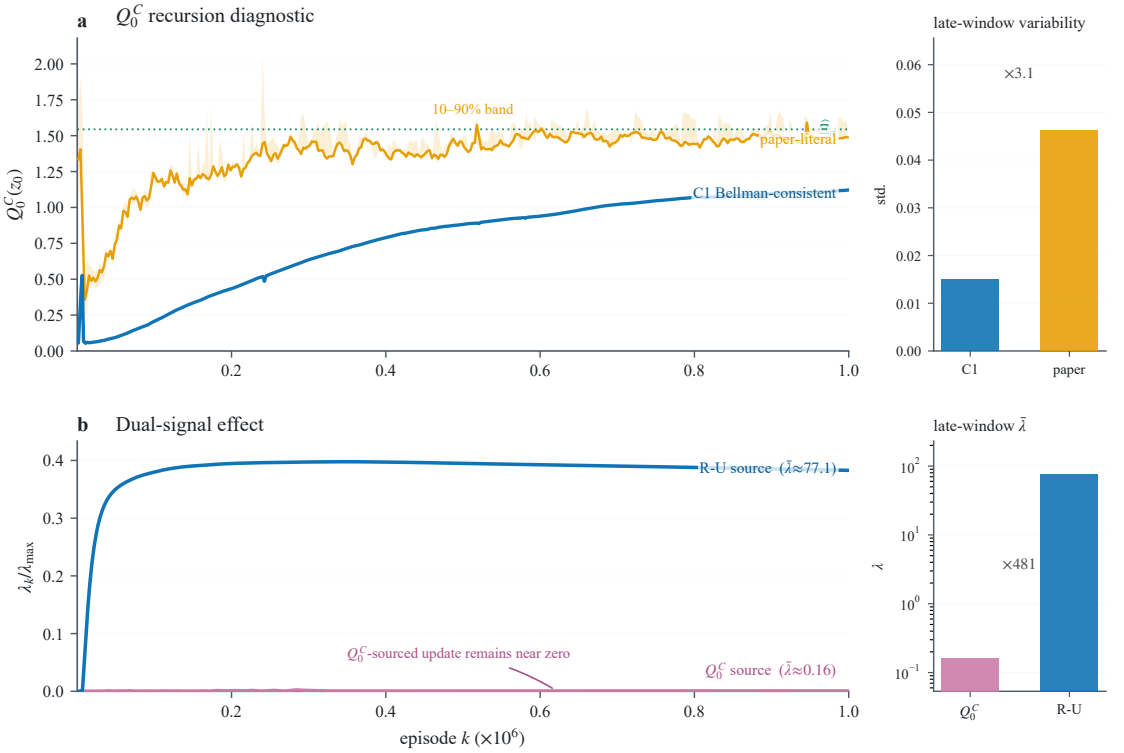

Figure 5.1. Single-agent Q_0^C diagnostic at $\alpha = 0.10$, $C_\alpha = 1.60$, fixed (β, π_b) off the β-recursion fixed point ($V_0^{\pi_b}(\beta) < \alpha$). **(a)** Median and 10–90% band of $Q_0(z_0, a^\star)$ for two CVaR-side recursions inside the C1+C2 architecture: C1 Bellman-consistent (4.5) (blue) and paper-literal (3.5) (orange; the oRMDP (Borkar and Jain, 2014) verbatim). Green dotted: $\Xi_\alpha(\beta; Y^{\pi_b})$ reference. Right bar: late-window $\sigma(Q_0)$ ratio paper-literal/C1 $\approx 3.1\times$. **(b)** $\lambda^k/\lambda_{\max}$ with dual signal sourced from Q_0^C (pink, $\bar{\lambda} \approx 0.16$) versus the R–U buffer estimator (4.17) (blue, $\bar{\lambda} \approx 77.1$); right bar shows the $\approx 481\times$ ratio.

Q-function side (panel a). Panel (a) applies Section 3.2's unrolling diagnostic at an operating point off the β-recursion fixed point ($V_0^{\pi_b} < \alpha$). The Bellman-consistent C1 recursion settles strictly below $\Xi_\alpha(\beta; Y^{\pi_b})$ by the closed-form residual of Proposition 4.2, as predicted. The paper-literal recursion shares C1's residual exactly and adds only the small absolute offset of (3.5), here within sampling dispersion of the C1 trace. Mean-level agreement is not Bellman consistency: the $\approx 3.1\times$ gap in late-window $\sigma(Q_0)$ between paper-literal and C1 shows that the literal recursion's two cancelling errors agree only on average, while their finite-time fluctuations remain wider on every window. This is what the diagnostic of Chapter 3 predicts at the Q-function level.

Constraint side (panel b). Panel (b) answers directly the question raised by the diagnostic of Section 3.1: does the literal recursion enforce the constraint? At the same operating point, the multiplier λ^k is monitored under two dual-signal sources. With Q_0^C as the dual signal — the oRMDP default — the closed-form residual $\beta(1 - V_0^{\pi_b}/\alpha)$ of Proposition 4.2 amplifies fast Q_0^C fluctuations into the slow dual gradient; λ never escapes the neighbourhood of zero (late-window $\bar{\lambda} \approx 0.16$), so the constraint exerts essentially no pressure on action selection, and a learner so configured does not enforce the CVaR budget — the underlying observation behind the diagnostic of Section 3.1. With the C3 buffer estimator as the dual signal, the same limiting projected dual equilibrium condition is targeted (Section 4.5) without inheriting the amplification: λ ascends within the same iteration budget to $\bar{\lambda} \approx 77.1$, a $\approx 481\times$ ratio — the dual-side empirical fingerprint of the β/α amplification scale of Corollary 4.3 — restoring constraint pressure on the greedy direction. The Q-function-side fluctuations of panel (a) are therefore not merely cosmetic: routed through the dual mechanism, they collapse the multiplier and disable the constraint, while the C3 substitution restores it. Panel (a) validates C1; panel (b) validates C3.

5.3 Validation of the operational gate: five-seed dispersion

What the gate measures, and why it appears here. Section 5.2 established C1 and C3 against the unrolling diagnostic at a fixed off-fixed-point operating point. This section turns to a different question: whether the assembled C1+C2+C3 learner reaches the saddle set (4.1) of Section 4.1 under the moving schedule of Algorithm 4.1. The saddle set is characterised by the four Karush–Kuhn–Tucker (KKT) conditions for the CMDP Lagrangian: stationarity in the primal threshold β, primal feasibility $\Xi_\alpha \leq C_\alpha$, dual feasibility $\lambda \geq 0$, and complementary slackness $\lambda(\Xi_\alpha - C_\alpha) = 0$ (Borkar and Jain, 2014). The operational gate of this thesis is a five-component finite-time KKT residual diagnostic that places explicit tolerances on the corresponding sample quantities: β-stationarity $|\bar{V}_0 - \alpha|$, feasibility $\widehat{\Xi}_\alpha - C_\alpha$, complementarity $\lambda(\widehat{\Xi}_\alpha - C_\alpha)$, λ-drift over a rolling window, and the sup-norm Bellman residual on the C1 cost Q-function. The residual is read on Polyak-averaged iterates over a late-window sample. The gate is the same instrument used in Chapter 7 for the multi-agent comparison and in Section 5.4 for the (α, ρ) stress test; its appearance here at $N{=}1$ supplies the single-agent baseline those chapters reference. The five residuals listed above are the single-agent predecessors of the G_1–G_5 gate labels formally defined in Section 7.1; Chapter 6 fixes the centralised tolerances and the two-tier scaling convention, which differ from the present single-agent bands and are not directly comparable as numerical certificates.

Figure 5.2 reports the residual traces and the per-seed episodes-to-gate. Every seed clears the five-component stop within budget; episodes-to-gate are spread across seeds by about $2.2\times$ between fastest and slowest. The dispersion fixes the single-agent reference for the centralised stress test of Chapter 7: a single budget covers

every single-agent seed comfortably, and the gate is the instrument with which the joint-dimension breakdown at $N{=}3$ will be located.

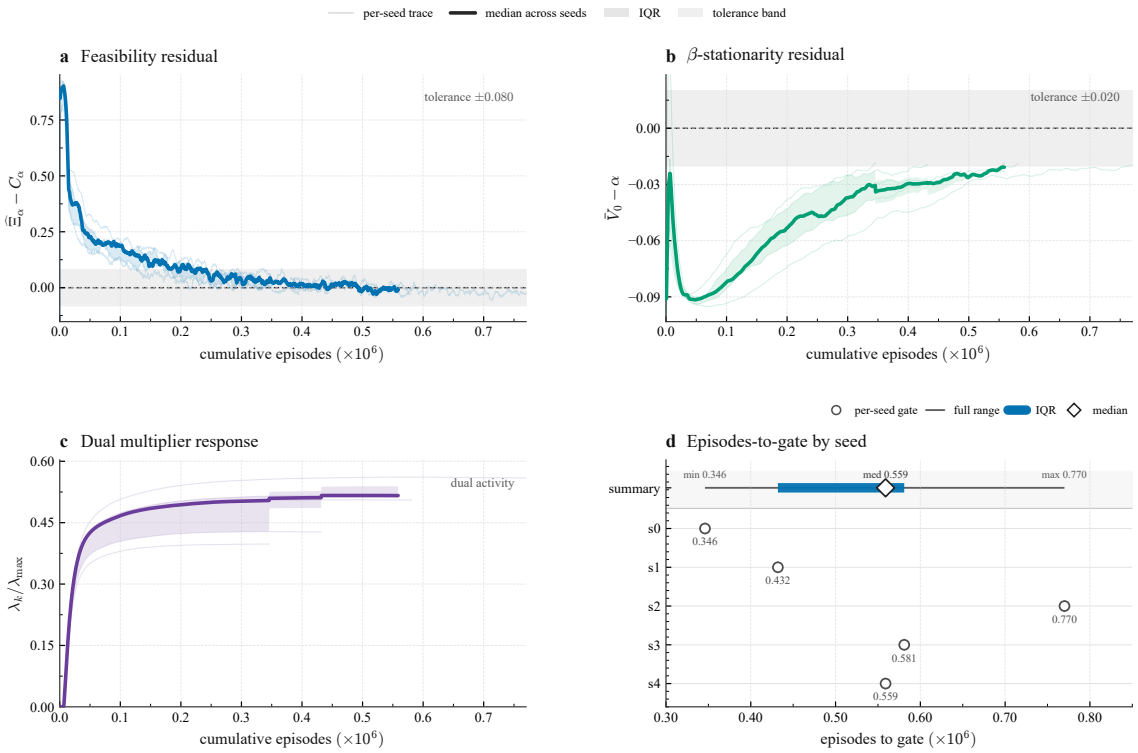

Figure 5.2. Single-agent KKT residuals and seed dispersion ($\alpha = 0.10$, $C_\alpha = 1.60$, five seeds). (a) Feasibility $\widehat{\Xi}_\alpha - C_\alpha$ vs. cumulative episodes, ± 0.080 band; (b) β-stationarity $\bar{V}_0 - \alpha$, ± 0.020 band; (c) dual multiplier $\lambda^k/\lambda_{\max}$; (d) episodes-to-gate per seed (min/median/max = 0.346/0.559/0.770, $\times 10^6$). Per-seed traces faint, five-seed median bold, IQR shaded. The ± 0.020 band on $\bar{V}_0 - \alpha$ in panel (b) is tighter than the centralised stress-test value $\tau_V = 0.05$ of Chapter 6 (Section 6.6), so episodes-to-gate are not directly comparable across the two chapters.

5.4 Validation of Proposition 4.2: the operating envelope across the (α, ρ) grid

Beyond gate clearance at the $(\alpha, C_\alpha){=}(0.10, 1.60)$ calibration point of Section 5.3, this section extends validation across a calibrated (α, ρ) grid to test the $1/\alpha$ envelope predicted by Proposition 4.2. The motivation is operational: the deployed greedy policy's held-out CVaR was observed to miss the calibrated budget by a small amount independent of the schedule. The closed-form residual of Proposition 4.2 identifies the mechanism: the fast-timescale fluctuation δ_V enters the Q_0^C readout multiplied by β/α, and even with C3's buffer estimator removing the Q_0^C signal from the dual update itself, the same factor governs how cleanly the late-window (β, V_0) trajectory projects onto $\Xi_\alpha^{\pi_b}$.

The (α, ρ) grid in Figure 5.3 was then constructed to ask a sharper question: if the residual's $1/\alpha$ amplification is real, the median relative residual should collapse onto a single constant in α-rescaled units across cells, with cell-level departures structured by where the underlying continuity and tail-sample assumptions of the closed form

are most strained. Five seeds per cell on a 3×3 design gives 45 runs — enough to detect the collapse against the predicted envelope at the median without claiming a scaling slope on a single α axis.

The budget C_α is positioned at fixed relative pressure ρ between the safe-policy and risky-policy reference CVaR_α levels, so the same constraint pressure is exerted across cells and the comparison is decoupled from variation in absolute margin.

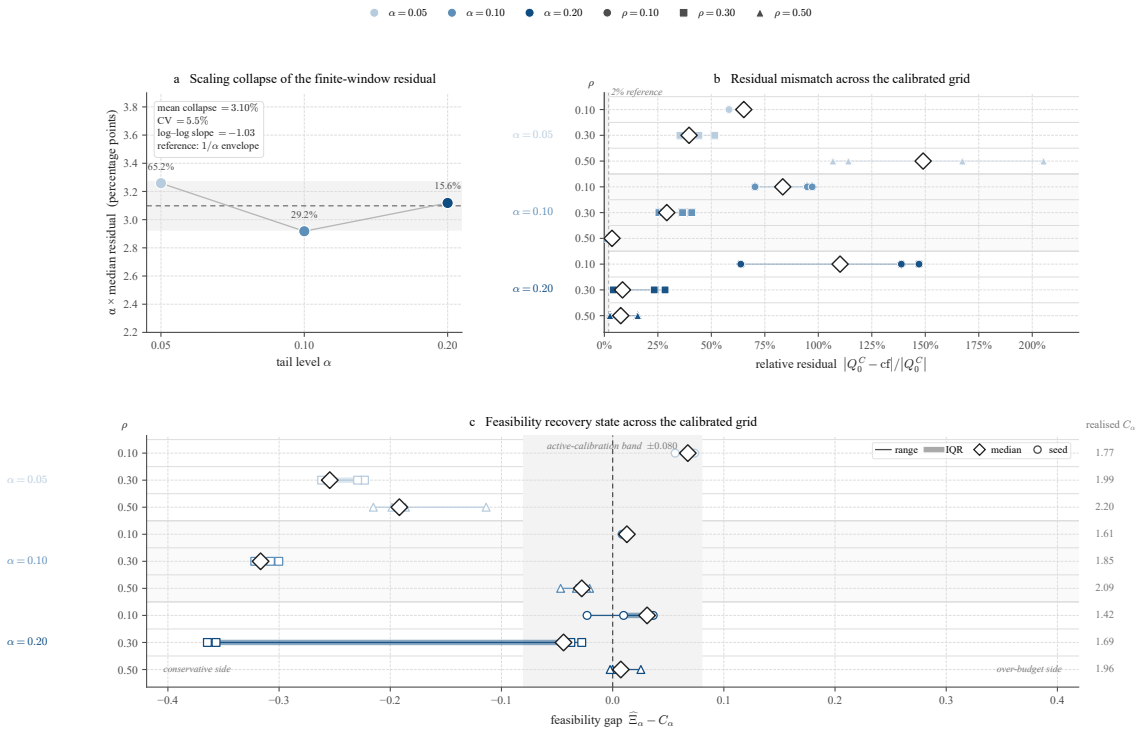

Figure 5.3. α-stratified operating envelope of the C1+C2+C3 specification across the calibrated (α, ρ) grid; 3×3 design, five seeds per cell (45 runs). For each $\alpha \in \{0.05, 0.10, 0.20\}$, C_α is positioned at fixed relative pressure $\rho \in \{0.10, 0.30, 0.50\}$ between the safe-policy and risky-policy reference CVaR_α levels; ρ=0.10 is the tightest cell and ρ=0.50 the loosest. Realised C_α: $\{1.77, 1.99, 2.20\}$ at α=0.05, $\{1.61, 1.85, 2.09\}$ at α=0.10, $\{1.42, 1.69, 1.96\}$ at α=0.20, reported in cell labels of panels (b)–(c). **(a)** Scaling collapse of the finite-window residual: per-α medians $\alpha \cdot \mathrm{median}(|Q_0^C - \mathrm{cf}|/|Q_0^C|) \in \{3.26, 2.92, 3.12\}\%$, coefficient of variation 5.5% (sample, Bessel-corrected), mean 3.10%, log–log slope -1.03; dashed horizontal marks the empirical collapse constant. **(b)** Per-cell relative residual dispersed by seed; ±2% vertical reference shown as a tightness band. **(c)** Feasibility gap $\widehat{\Xi}_\alpha - C_\alpha$ at the gate with ±0.08 active-calibration band. Here cf denotes the closed-form prediction from Proposition 4.2.

Key takeaways from Figure 5.3. Three findings, one per panel. (i) The per-α median residuals collapse onto a single constant in α-rescaled units, with a three-point log–log slope of -1.03; this is consistent with, rather than an estimate of, the slope -1 predicted by the V_0/α factor of Proposition 4.2, leaving a protocol-specific collapse constant as the only free parameter. (ii) Cell-level departures track two routes predicted by (4.14): β-instability under point-mass-structured returns at low ρ (most extreme at α=0.20, ρ=0.10), and tail rarefaction at low α when the effective tail sample count $W\alpha$ thins the buffer estimator (most extreme at α=0.05, ρ=0.50). (iii) Feasibility at the gate sits within the ±0.08 active-calibration band at α=0.20; at

lower α the $1/\alpha$-amplified cost Q-function fluctuations perturb the greedy direction of $J = Q^R - \lambda Q^C$ and bias the late-window operating point conservatively (the trained policy uses less than the calibrated budget).

The operating envelope of the C1+C2+C3 specification is therefore characterised at the calibrated protocol: the asymptotic identity of Proposition 4.2 is recovered at the median to within a constant 3.10% in α-rescaled units, feasibility within the active-calibration band is achieved on every cell at α=0.20, and at α=0.05 the V_0/α amplification saturates the envelope with a conservative bias of order $0.1\, C_\alpha$.

The α axis here plays the same diagnostic role, in the single-agent setting, that the N axis plays in the centralised lift of Chapter 8: both expose finite-budget stress on the C1+C2+C3 envelope, though through different mechanisms — here through the V_0/α amplification and the finite-$W\alpha$ tail-sample count, there through joint table volume and joint-return variance. Routes to reducing the envelope — a buffer-side dual estimator with an explicit finite-$W\alpha$ correction, or the per-agent decomposition of Chapter 9 — are taken up in the chapters indicated.

5.5 Forward to the centralised lift

Chapter 6 fixes the centralised multi-agent test bed; Chapter 7 stress-tests $N \in \{1, 2, 3\}$ and locates the breakdown at $N = 3$; Chapter 8 attributes it to two structural quantities of the centralised representation; Chapter 9 examines the per-agent decomposition the attribution points to. With the single-agent Q-function–multiplier coupling resolved by Chapters 3–4 and the operating envelope of the C1+C2+C3 learner characterised in this chapter, the $N = 3$ breakdown is interpreted, under the experimental controls of Chapters 6–8, as evidence for a centralised-representation limitation.

6

Centralised Multi-Agent Setting and Stress-Test Protocol

We apply the C1+C2+C3 learner of Chapter 4 as a controlled centralised stress test: the single-agent specification of Chapter 5 is lifted, without architectural modification, into the joint representation of the N-agent factored MMDP, and run on $N \in \{1, 2, 3\}$ under one frozen stress-test protocol. No new analytical claim is introduced. The scaling comparison of Chapters 7–8 rests on within-protocol invariance: hyperparameters, gate tolerances, and per-seed budget are calibrated at $N = 1$ and reused without retuning across $N \in \{1, 2, 3\}$ and across seeds. Any failure of Algorithm 4.1 reported under this protocol is therefore interpreted as a property of the centralised representation, with the single-agent Q-function–multiplier confound of Chapter 3 removed and per-N retuning excluded as an explanation.

6.1 The centralised CVaR-CMDP

The two failures diagnosed in Chapter 3 are single-agent. To test whether the changes developed in Chapter 4 survive in a multi-agent setting, we use as immediate reference the centralised representation formalised in Section 2.2.5: a single learner acting over the joint state-action space of the N-agent factored MMDP, under a single CVaR budget,

$$\begin{aligned} \pi^\star \in \arg\max_{\pi} \; & \mathbb{E}^\pi[\textstyle\sum_i R_i^\pi] \\ \text{s.t. } & \mathrm{CVaR}_\alpha(\textstyle\sum_i Y_i^\pi) \le C_\alpha^{\text{eff}}, \\ & C_\alpha^{\text{eff}} := N \cdot C_{\text{base}}, \end{aligned} \tag{6.1}$$

restating Equation (2.14) of Section 2.2.5 for direct reference in this chapter. The corrected single-agent learner of Chapter 4, defined over $Z_t = (t, S_t, Y_t)$, lifts directly to the joint space: the augmented state has size $T \cdot |\mathcal{S}|^N \cdot |\mathcal{Y}|$ and the joint action has size $|\mathcal{A}|^N$.

As argued in Section 2.2.5, this lift inherits the role separation enforced by C1–C3 in full, so a persistent failure pattern under the frozen protocol is more naturally attributed to the joint representation than to the single-agent mechanisms repaired in Sections 3.2 and 4.4. Two outcomes are possible. If the centralised lift clears the operational gate at every tested N, then role separation is sufficient for this stress

test and the centralised representation remains adequate within the tested range. If the gate fails at some N despite role separation, the result points away from the repaired single-agent conflation mechanism and toward the joint representation. The standard confounders — hyperparameters, exploration schedule, per-seed budget — are held fixed across N, and seed-level dispersion is reported, so a stall at a specific N admits a representation-level attribution rather than a tuning explanation. Chapters 7–8 carry out the stress test and interpret its outcome.

6.2 Environment: the Taxi MMDP

Indexing convention. From this chapter onward, the horizon symbol T denotes the *number* of joint decisions (with $T = 12$ in the instantiation below), decision columns are $t \in \{0, 1, \dots, T-1\}$, and the terminal column index T carries no decision. This convention differs from Chapters 2–4, where T is the final decision index and decisions are taken at $t \in \{0, 1, \dots, T\}$; the two conventions are consistent on the per-episode count of Q-updates in Algorithm 4.1, and the difference is one of indexing only.

Per-agent MDP (recalled from Section 5.1). Each agent $i \in \{1, \dots, N\}$ executes the per-agent template $\mathcal{M}_0 = \langle \mathcal{S}_0, \mathcal{A}_0, (P_{0,t}), (r_{0,t}), (c_{0,t}), T \rangle$ defined in Section 5.1 of Chapter 5, with $|\mathcal{S}_0| = |\mathcal{A}_0| = 3$, $\bar{c} = 0.50$, $T = 12$, and the centre-lane initial state $s_{i,0} = 1$ at column 0 shared across agents; the subscript 0 marks the per-agent template (not an agent index), as established there. The full per-step reward, hazard, and transition schedules are reproduced in Appendix A. The lift to N agents below uses $\mathcal{M}_0$ verbatim.

Joint MMDP. The N-agent MMDP is the centralised lift, with joint state space $\mathcal{S}^{(N)} = \mathcal{S}_0^N$ and joint action space $\mathcal{A}^{(N)} = \mathcal{A}_0^N$, hence $|\mathcal{S}^{(N)}| = 3^N$ and $|\mathcal{A}^{(N)}| = 3^N$. Trajectories factorise:

$$\begin{aligned} P_t^{(N)}(s' \mid s, a) &= \prod_{i=1}^{N} P_{0,t}(s_i' \mid s_i, a_i), \\ R_t^{(N)}(s, a) = \sum_{i=1}^{N} r_{0,t}(s_i, a_i), &\qquad C_t^{(N)}(s, a) = \sum_{i=1}^{N} c_{0,t}(s_i, a_i). \end{aligned} \tag{6.2}$$

The joint cumulative cost is $Y_T^{(N)} = \sum_{t=0}^{T-1} C_t^{(N)}(s_t, a_t)$. The CVaR constraint of (6.1) is enforced at the joint budget $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$, $C_{\text{base}} = 1.60$ — the *per-capita budget convention* used throughout, under which the per-agent allowance $C_\alpha^{\text{eff}}/N = C_{\text{base}}$ is held constant across N.

Episode-end smoothing shock. A scalar $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ with $\sigma_\xi = 0.10$ is sampled once per episode and added to the rolled-out return: the smoothed joint cumulative cost is $\tilde{Y} := Y_T^{(N)} + \xi$, with sample $Y^k = Y_T^{(N),k} + \xi^k$ available to the algorithm. The buffer is populated by Y^k and the terminal indicator reads

$V_T = \mathbb{1}\{Y^k > \beta\}$. The discrete return c.d.f. $F_{Y_T^{(N)}}$ carries atoms at its support points; convolving with $\mathcal{N}(0, \sigma_\xi^2)$ produces a smooth Lebesgue density uniformly bounded by $1/(\sigma_\xi\sqrt{2\pi})$, so the smoothed return c.d.f. $F_{\tilde{Y}}$ is continuous (in fact C^∞) everywhere. This satisfies the bounded-density premise of Borkar and Jain (2014, Proposition 2), which yields the Lipschitz-continuity of $h(\beta) = \mathbb{P}(\tilde{Y} > \beta)$ required by the β-recursion convergence of Borkar and Jain (2014, Lemma 2), and a fortiori the continuity of $F_{\tilde{Y}}$ at $\mathrm{VaR}_\alpha(\tilde{Y})$ that the per-agent risk-contribution estimator of Section 9.2 needs for population-level consistency. The Gaussian shock has unbounded support, so we do not use it to support any bounded-increment SA convergence claim; bounded alternatives ($\xi \sim U[-\delta, \delta]$, clipped Gaussian) supply the same continuity inside their support.

The shock enters the V terminal indicator $V_T = \mathbb{1}\{\tilde{Y} > \beta\}$ on the smoothed return $\tilde{Y}$ but *not* the cost-side stage source: under the deterministic bootstrap selector $a^\star$ the cost at stage t, $\tilde{c}(Z_t) = C_t^{(N)}(s_t, a^\star(z_t))$, is $\sigma(Z_t)$-measurable in the sense of Chapter 2, and ξ does not contribute to $\sum_t \tilde{c}_t = Y_T^{(N)}$. Backward propagation preserves the terminal, so the C1 cost Q-function's fixed-point identification under the tower identity at $a^\star$ is evaluated on the cross-tail event of the unshocked aggregate weighted by the smoothed-tail indicator,

$$\begin{aligned} Q_0^C(z_0; \beta) \;&\to\; \alpha^{-1}\,\mathbb{E}^{a^\star}\big[Y_T^{(N)}\,\mathbb{1}\{\tilde{Y} > \beta\}\big] \\ &=\; \mathrm{CVaR}_\alpha(\tilde{Y}^{a^\star}) \;-\; \alpha^{-1}\,\mathbb{E}^{a^\star}\big[\xi\,\mathbb{1}\{\tilde{Y} > \beta\}\big] \end{aligned} \tag{6.3}$$

at the β-recursion fixed point $V_0^{a^\star}(\beta) = \alpha$, using $Y_T^{(N)} = \tilde{Y} - \xi$ and the standard Rockafellar–Uryasev decomposition on the first term. The residual $\alpha^{-1}\mathbb{E}[\xi\,\mathbb{1}\{\tilde{Y} > \beta\}]$ is of order $O(\sigma_\xi/\alpha)$ at the operating σ_ξ and vanishes as $\sigma_\xi \to 0$. The rate follows from the closed form $\sigma_\xi^2\, f_{\tilde{Y}}(\beta)/\alpha$ — supplied by Stein's identity for the Gaussian shock[1] — together with $f_{\tilde{Y}}(\beta) = O(\sigma_\xi^{-1})$, which holds because the underlying $Y_T^{(N)}$ is atomic on the Taxi MMDP. The C3 buffer estimator, by contrast, computes its threshold from the buffer itself: an empirical $(1-\alpha)$-quantile $\widehat{\beta}$ is computed from the rolling buffer of π_b-returns and $\widehat{\Xi}_\alpha^{\mathcal{B}} = \widehat{\beta} + \alpha^{-1}\,\overline{(Y - \widehat{\beta})_+}$ is the Rockafellar–Uryasev sample CVaR. In the large-buffer quasi-stationary limit $\widehat{\beta} \to \mathrm{VaR}_\alpha^{\pi_b}(\tilde{Y})$ and $\widehat{\Xi}_\alpha^{\mathcal{B}} \to \mathrm{CVaR}_\alpha^{\pi_b}(\tilde{Y})$ on $\tilde{Y}$ by the sample-mean LLN. The C1 cost Q-function and the C3 buffer estimator therefore target $\mathrm{CVaR}_\alpha(\tilde{Y})$ up to two finite residuals — the $O(\sigma_\xi/\alpha)$ cross-tail term on the Q-function side and the π_b–$a^\star$ gap at finite ϵ (Section 6.4) — and align with $\mathrm{CVaR}_\alpha(Y_T^{(N),\,\pi_b})$ on the original unshocked variable in the joint $\sigma_\xi \to 0$, $\epsilon \to 0$ limit.

To compare the two against a single threshold b, evaluate both at that fixed b. The buffer estimator's fixed-threshold target is $\Xi_\alpha(b; \tilde{Y}^{\pi_b})$, and the Q-function-side

[1] Equivalently, by integration by parts on $u\,\phi_{\sigma_\xi}(u) = -\sigma_\xi^2\,\phi'_{\sigma_\xi}(u)$ conditional on $Y_T^{(N)}$, with ϕ_{σ_ξ} the Gaussian-shock density.

fixed-threshold limit is $Q_0^C(z_0; a^\star, b) = \alpha^{-1}\mathbb{E}^{a^\star}[Y_T^{(N)}\mathbb{1}\{\tilde{Y} > b\}]$ from (6.3). The gap decomposes as

$$\Xi_\alpha(b; \tilde{Y}^{\pi_b}) - Q_0^C(z_0; a^\star, b) = \underbrace{\Xi_\alpha(b; \tilde{Y}^{\pi_b}) - \Xi_\alpha(b; \tilde{Y}^{a^\star})}_{\text{behaviour–greedy gap}} + \underbrace{\left[\Xi_\alpha(b; \tilde{Y}^{a^\star}) - Q_0^C(z_0; a^\star, b)\right]}_{\text{closed-form residual + cross-tail}}. \tag{6.4}$$

The second term has two components. The first is the fixed-threshold closed-form residual $b\,(1 - V_0^{a^\star}(b)/\alpha)$ of Proposition 4.2 with $\tilde{Y}$ substituted for Y throughout (which uses only the $\sigma(Z_t)$-measurability of $\tilde{c}$ at $a^\star$ together with the tower property, both preserved here); it vanishes at the β-recursion fixed point $V_0^{a^\star}(b) = \alpha$. The second is the cross-tail residual $\alpha^{-1}\mathbb{E}^{a^\star}[\xi\,\mathbb{1}\{\tilde{Y} > b\}]$ identified in (6.3), which remains as an $O(\sigma_\xi/\alpha)$ smoothing slack and vanishes as $\sigma_\xi \to 0$. The behaviour–greedy gap absorbs the joint-action deviation of π_b from $a^\star$ under the ϵ-greedy product policy; no asymptotic order claim in either σ_ξ or ϵ is made. This separated gap is what the C3 role separation absorbs: $\widehat{\Xi}_\alpha^{\mathcal{B}}$ supplies the dual feasibility signal and Q^C feeds the greedy selector. Because ξ is shared at the joint-return level rather than drawn per agent, it contributes a constant $\sigma_\xi^2 = 0.01$ to $\mathrm{Var}(\tilde{Y})$ independent of N, so any N-scaling reported in Chapter 8 cannot be attributed to ξ.

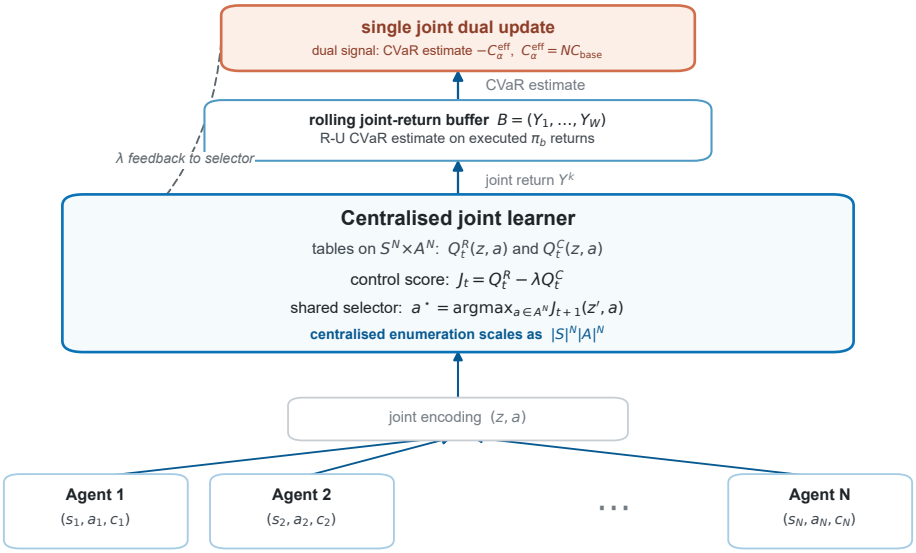

Figure 6.1. Centralised online CVaR Q-learner. Algorithm 4.1 lifted to the joint state–action representation: per-agent outputs (s_i, a_i, c_i) are encoded into joint (z, a), on which a single learner maintains Q_t^R, Q_t^C conceptually over $\mathcal{S}^{(N)} \times \mathcal{A}^{(N)}$ with shared selector $a^\star = \arg\max_a \big(Q_{t+1}^R(z, a) - \lambda Q_{t+1}^C(z, a)\big)$. The implemented table further indexes time t and accumulated cost y in the augmented state $z = (t, s, y)$ of Chapter 2; the implementation shape used in Chapter 8's mechanism (1) reading is $T \cdot |S|^N \cdot |Y| \cdot (|A|^N + 1) \cdot |A|^N$. The C3 R–U buffer-mean estimator supplies the dual signal against $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$; λ enters the Bellman path only through the selector.

Relation to the Chapter 4 equality. The exact C1 identity $Q_0^C(z_0;\pi,\beta) = \Xi_\alpha(\beta;Y^\pi)$ is on the unsmoothed Y^π; the $O(\sigma_\xi/\alpha)$ cross-tail slack in (6.3) is an implementation-level finite residual, not part of any convergence claim.

Choice of scale. The MMDP is sized so that $N = 1$ clears the gate with headroom under the per-seed budget, $N = 2$ remains feasible, and $N = 3$ exhibits a decisive break at the same budget. The construction is a controlled stress test, not a scalability benchmark; whatever drives the break is the object Chapter 8 examines.

Algorithm 6.1 Centralised online CVaR Q-learner on the N-agent factored MMDP. Starred ($\star$) lines mark departures from the ORMDP of Borkar and Jain (2014): the C1, C2, C3 modifications of Chapter 4 and the per-agent independent rollout of Section 6.3.

Require: risk level $\alpha \in (0,1)$; joint CVaR budget $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ (Eq. (6.1)); horizon T; step-size schedules $\{a_k\} \gg \{\gamma_k\} \gg \{\eta_k\}$; buffer capacity W, activation $W_{\min}$; dual clip $\lambda_{\max}$; exploration rate ϵ; initial threshold β_0; number of agents N.

Ensure: Q^R, Q^C, λ on the joint table; deployed greedy policy $\pi_g(z) = \arg\max_a [Q^R(z,a) - \lambda Q^C(z,a)]$.

$Q^R, Q^C, V \leftarrow 0$; $\beta_1 \leftarrow \beta_0$; $\lambda_1 \leftarrow 0$; $\mathcal{B} \leftarrow \varnothing$
for $k = 1, 2, \ldots$ **do**
  ▷ *Rollout: per-agent independent ϵ-greedy (Section 6.3).*
  $\star$ At each t, compute $a_t^\star = \arg\max_a [Q_t^{R,k} - \lambda_k Q_t^{C,k}](Z_t^k, a)$ and decompose $a_t^\star = (a_{t,1}^\star, \ldots, a_{t,N}^\star)$
  $\star$ For each agent i, set $A_{t,i}^k = a_{t,i}^\star$ w.p. $1-\epsilon$, otherwise draw uniformly on $\mathcal{A}_0$; let $A_t^k = (A_{t,1}^k, \ldots, A_{t,N}^k)$ ▷ per-agent ϵ-greedy
  Roll out $Z_0^k \to Z_T^k$; record $\{A_t^k, \tilde{c}_t, r_t\}_{t=0}^{T-1}$ and $Y^k = \sum_t \tilde{c}_t + \xi^k$, $\xi^k \sim \mathcal{N}(0, \sigma_\xi^2)$
  $V_T^k \leftarrow \mathbb{1}\{Y^k > \beta_k\}$; $Q_T^{R,k} \leftarrow 0$; $Q_T^{C,k} \leftarrow 0$
  ▷ *Backwards Bellman update on the fast timescale a_k.*
  for $t = T-1, \ldots, 0$ **do**
    $(z, a, z') \leftarrow (Z_t^k, A_t^k, Z_{t+1}^k)$
    $\star$ $a^\star \leftarrow \arg\max_{a'} \big[Q_{t+1}^{R,k}(z', a') - \lambda_k Q_{t+1}^{C,k}(z', a')\big]$ ▷ shared selector (C2)
    $V_t^k(z)$ += $a_k \big[V_{t+1}^k(z') - V_t^k(z)\big]$
    $\star$ $Q_t^{R,k}(z,a)$ += $a_k \big[r_t + Q_{t+1}^{R,k}(z', a^\star) - Q_t^{R,k}(z,a)\big]$ ▷ reward Q-function (C2)
    $\star$ $Q_t^{C,k}(z,a)$ += $a_k \big[\alpha^{-1} V_t^k(z)\, \tilde{c}(z,a) + Q_{t+1}^{C,k}(z', a^\star) - Q_t^{C,k}(z,a)\big]$ ▷ cost Q (C1)
  end for
  ▷ *Threshold update on the intermediate timescale γ_k.*
  $\beta_{k+1} \leftarrow \beta_k - \gamma_k \big(\alpha - V_0^k(z_0)\big)$
  ▷ *Dual update on the slow timescale η_k.*
  Append Y^k to $\mathcal{B}$ (rolling capacity W)
  if $|\mathcal{B}| \geq W_{\min}$ **then**
    $\star$ $\widehat{\beta}_k \leftarrow \widehat{\mathrm{VaR}}_\alpha(\mathcal{B})$; $\widehat{\Xi}_\alpha^{\mathcal{B}} \leftarrow \widehat{\beta}_k + \dfrac{1}{\alpha|\mathcal{B}|} \sum_{Y_j \in \mathcal{B}} (Y_j - \widehat{\beta}_k)^+$ ▷ R–U estimator (C3)
    $\star$ $\lambda_{k+1} \leftarrow \Pi_{[0,\lambda_{\max}]}\big(\lambda_k + \eta_k [\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}}]\big)$ ▷ projected dual ascent on C_α^{eff}
  else
    $\lambda_{k+1} \leftarrow \lambda_k$
  end if
end for
return Q^R, Q^C, λ

Pseudocode and departures from Borkar–Jain. Algorithm 6.1 states the centralised lift in pseudocode parallel to the literal oRMDP of Borkar and Jain (2014); starred ($\star$) lines mark departures from that recursion, unstarred lines coincide with it. The starred lines fall into four classes: the three structural changes of Chapter 4 (C1, C2, C3) and the rollout-policy change of the present chapter (per-agent independent ϵ-greedy of Section 6.3).

Reading the stars. The four classes appear in execution order through one episode. (i) The two rollout-policy lines compute and sample the joint action via per-agent independent ϵ-greedy of (6.7), replacing the joint ϵ-greedy (6.6) of Borkar and Jain (2014). (ii) The shared-selector line and the separate Q^R recursion on a λ-independent stage source together implement C2: a single $a^\star = \arg\max_a [Q^R - \lambda Q^C]$ feeds both Bellman recursions in place of the literal scheme's single combined table. (iii) The cost Q-function recursion implements C1: the stage source $\alpha^{-1} V_t^k(z)\, \tilde{c}(z,a)$ derived in Section 4.2 replaces the literal cost-side update whose fixed point lies close to Ξ_α only by algebraic cancellation (Chapter 3). (iv) The dual-update pair on the slow timescale implements C3: the R–U buffer estimator $\widehat{\Xi}_\alpha^{\mathcal{B}}$ on $\mathcal{B}$ supplies the dual signal in place of the Q_0^C readout, and the projection is onto $[0, \lambda_{\max}]$ at the joint per-capita budget $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$. Unstarred lines are common to both algorithms: the rollout mechanics, the augmented-state $z = (t, s, y)$, the β update on the intermediate timescale, the λ clipping mechanism, and the three-timescale ordering itself.

6.3 Exploration: per-agent independent ϵ-greedy

Definition: Hamming distance. For two joint actions $a, a^\star \in \mathcal{A}^{(N)}$ written componentwise as $a = (a_1, \ldots, a_N)$ and $a^\star = (a_1^\star, \ldots, a_N^\star)$,

$$d(a, a^\star) \;:=\; \big|\{i \in \{1, \ldots, N\} : a_i \neq a_i^\star\}\big| \;\in\; \{0, 1, \ldots, N\} \tag{6.5}$$

is the number of agent slots in which a and $a^\star$ disagree. A *Hamming-d deviation from* $a^\star$ is a joint action with $d(a, a^\star) = d$; the Hamming-1 deviations are the $N(|\mathcal{A}_0| - 1)$ joint actions reached by changing exactly one agent's slot.

Joint ϵ-greedy (Borkar–Jain reference). Borkar and Jain (2014, Theorem 2) treats the rollout policy as the joint ϵ-greedy

$$\pi_b^{\text{joint}}(a \mid z) \;=\; (1 - \epsilon)\, \mathbb{1}\{a = a^\star(z)\} \;+\; \frac{\epsilon}{|\mathcal{A}^{(N)}|}, \tag{6.6}$$

i.e. the greedy joint action $a^\star(z) = \arg\max_a J_t(z, a)$ is played with probability $1 - \epsilon + \epsilon/|\mathcal{A}^{(N)}|$ and the remaining ϵ-mass is spread uniformly over the full joint action set $\mathcal{A}^{(N)}$ of size 3^N. A specific off-greedy joint action therefore receives probability $\epsilon/|\mathcal{A}^{(N)}| = \epsilon/3^N$ — $\epsilon/27$ at $N = 3$. The Hamming-1 successors of $a^\star$ that patch the cells the bootstrap accesses therefore decay in N as a vanishing fraction of the deviation budget.

Per-agent variant. We replace (6.6) by per-agent independent ϵ-greedy: at each rollout step we read the joint greedy action $a^\star = \arg\max_a J_t(z, a)$ and decompose it into its per-agent components $a^\star = (a_1^\star, \ldots, a_N^\star)$ (the joint action is exactly the tuple of per-agent actions, so the decomposition is the identity); each agent i then independently samples $a_i^\star$ with probability $1-\epsilon$ and otherwise draws uniformly from $\mathcal{A}_0$; the resulting tuple $(a_1, \ldots, a_N)$ is the executed joint action.

Proposition 6.1 (Hamming-distance coverage law)**.** *Under the per-agent independent ϵ-greedy above, for any $a \in \mathcal{A}^{(N)}$ at Hamming distance $d(a, a^\star) \in \{0, \ldots, N\}$,*

$$\Pr(a \mid a^\star) = \prod_{i=1}^{N} p_i(a_i \mid a_i^\star), \quad p_i(a_i \mid a_i^\star) = \begin{cases} 1 - \epsilon + \epsilon/|\mathcal{A}_0|, & a_i = a_i^\star, \\ \epsilon/|\mathcal{A}_0|, & a_i \neq a_i^\star. \end{cases} \tag{6.7}$$

With $|\mathcal{A}_i| \equiv |\mathcal{A}_0|$, $\Pr(a \mid a^\star) = (\epsilon/|\mathcal{A}_0|)^d (1 - \epsilon + \epsilon/|\mathcal{A}_0|)^{N-d}$.

Proof. Per-agent draws are independent conditional on $a^\star$, so the joint probability factors agent-wise. Each agent's marginal combines a $1-\epsilon$ greedy branch and an ϵ uniform branch, giving $1 - \epsilon + \epsilon/|\mathcal{A}_0|$ at the greedy component and $\epsilon/|\mathcal{A}_0|$ at any non-greedy component. □

Action support on the reachable set. Every joint action has positive probability conditional on the greedy joint action, hence positive execution probability at any reachable joint state.[2] This is what the asynchronous Q-learning convergence theorem of Tsitsiklis (1994) requires on the action side: positive probability on every action at every state the trajectory revisits infinitely often. Two coverage properties that asynchronous-SA *precision* (as opposed to almost-sure convergence in the limit) needs are *not* implied. First, the product-form rule does not guarantee uniform or finite-budget coverage of the joint state–action table $\mathcal{S}^{(N)} \times \mathcal{A}^{(N)}$: the per-cell visit count under a fixed per-seed transition budget shrinks as the joint table grows, and the visited fraction is concentrated on the trajectories the executed π_b generates rather than spread uniformly. Second, it does not produce visitation of cells the joint dynamics make unreachable from the shared initial state (s_0, π_b). Both points are properties of the joint representation independent of which ϵ-greedy variant is used and are taken up quantitatively in Chapter 8; the local-rate gain made by the product-form rule is recorded next.

Local coverage. A specific Hamming-1 deviation carries probability $(\epsilon/|\mathcal{A}_0|)(1 - \epsilon + \epsilon/|\mathcal{A}_0|)^{N-1}$, which still decays in N — geometrically at base $1 - \epsilon + \epsilon/|\mathcal{A}_0| < 1$ — but much more slowly than the joint ϵ-greedy probability $\epsilon/|\mathcal{A}^{(N)}|$ assigned to the same action: the ratio of the two rates widens exponentially in N. At the small $N \in \{1, 2, 3\}$ of the centralised stress test the product-form rule therefore preserves substantially more local Hamming-1 coverage than joint ϵ-greedy. This is the locality the bootstrap accesses.

[2] Specifically, the per-agent factor $\Pr(a_i \mid a_i^\star) \in \{\epsilon/|\mathcal{A}_0|, 1 - \epsilon + \epsilon/|\mathcal{A}_0|\}$ is strictly positive for every $a_i \in \mathcal{A}_0$, so the joint product $\prod_i \Pr(a_i \mid a_i^\star)$ in (6.7) is strictly positive for every $a \in \mathcal{A}^{(N)}$.

Limitation. The product-form rule restores per-agent rate, not joint rate. Made precise: each per-agent factor in (6.7) is $\Pr(a_i \mid a_i^\star) \geq \epsilon/|\mathcal{A}_0|$, the single-agent ϵ-greedy rate of Borkar and Jain (2014) at $N = 1$. The joint probability assigned to a Hamming-d deviation is the product of d such factors at non-greedy slots and $N - d$ factors at greedy slots, so a specific Hamming-d joint action receives

$$(\epsilon/|\mathcal{A}_0|)^d \, (1 - \epsilon + \epsilon/|\mathcal{A}_0|)^{N-d},$$

which decays geometrically in d. The full-deviation joint action ($d = N$) therefore receives $(\epsilon/|\mathcal{A}_0|)^N$, exponentially smaller than the joint ϵ-greedy assignment $\epsilon/|\mathcal{A}^{(N)}|$ to the same action. The relevant cell for centralised asynchronous Q-learning is a joint state–action cell, and the rule does not restore uniform coverage of $\mathcal{S}^{(N)} \times \mathcal{A}^{(N)}$: it patches local Bellman propagation around $a^\star$ at the per-agent action granularity, not the joint table volume; the joint-table volume is a representational quantity addressed in Chapter 8.

6.4 Behaviour distribution: what transfers from the convergence proof

Per-agent independent ϵ-greedy is not the joint ϵ-greedy on which Borkar and Jain (2014, Theorem 2) rests; it induces a different stationary π_b with deviation mass concentrated on Hamming-1 successors as (6.7) makes precise. The exploration rule therefore changes the sampling law that generates transitions, not the Bellman object the recursion evaluates.

The Bellman targets of Algorithm 4.1 remain well-defined under any stationary π_b: (4.2) is a definition under the executed policy. The tower identity (2.13), by contrast, is invoked at the deterministic bootstrap selector $a^\star$ — distinct from the sampling policy π_b — under which the cost at stage t, $\tilde{c}(Z_t) = c(S_t, a^\star(Z_t))$, is $\sigma(Z_t)$-measurable in the sense of Chapter 2; this is independent of whether π_b is the joint ϵ-greedy of Borkar and Jain (2014) or the product-form variant of Section 6.3. The Chapter 4 fixed-point identifications therefore transfer — Q_0^C to (6.3) and the buffer estimator to $\mathrm{CVaR}_\alpha^{\pi_b}(\tilde{Y})$ in the large-buffer limit. Proposition 4.2's closed-form residual, with $\tilde{Y}$ substituted for Y, uses only the $\sigma(Z_t)$-measurability of $\tilde{c}$ at $a^\star$ together with the tower property and transfers to the centralised lift on the same closed form, supplemented by the $O(\sigma_\xi/\alpha)$ cross-tail residual of (6.3). What does not transfer verbatim is the three-timescale convergence proof of Borkar and Jain (2014, Section V): that proof treats the exploration rule as a fixed parameter of the asynchronous noise structure, which the product-form variant alters. We treat $N = 1, 2$ gate satisfaction across five seeds as empirical evidence that the variant is operable under the frozen protocol, not as a new convergence theorem.

Constraint target and deployment gap. Chapters 7–8 report $\widehat{\Xi}_\alpha^{\mathcal{B}}$ on the smoothed return under π_b, consistent with the standard ϵ-greedy constrained-RL convention (Tessler et al., 2019; Stooke et al., 2020). Greedy-policy quantities under $\pi_g = \arg\max J$ are reported separately at the Polyak-averaged operating point.

The product-form π_b of Section 6.3 carries an N-dependent behaviour–greedy gap: by Proposition 6.1, the per-step joint-action disagreement between π_b and π_g is $1-(1-\epsilon+\epsilon/|\mathcal{A}_0|)^N$, growing from ≈ 0.033 at $N=1$ to ≈ 0.097 at $N=3$ under the protocol's $\epsilon=0.05$, $|\mathcal{A}_0|=3$. No uniform π_b-to-π_g CVaR transfer bound is claimed under the product-form π_b.

6.5 Stabilisations for finite-time runs

Algorithm 4.1's asymptotic guarantees do not prescribe what an early-training joint table looks like before Q^C has any information at most cells. We introduce five operational stabilisations that address this; each is a finite-time mechanism on π_b, the C1+C2 propagation, or the slow C3 update, and none enters the convergence claim. Table 6.1 records each by the path it perturbs and confirms all five are fixed at the $N=1$ centralised calibration of this chapter and reused without retuning across N.

Table 6.1. Finite-time stabilisations used in Algorithm 4.1 runs throughout Chapters 7–8. *Path* distinguishes the operational Bellman update from the gate read-out and reported greedy-policy evaluation. All five are fixed at the $N=1$ centralised calibration of this chapter and reused across $N\in\{1,2,3\}$ without retuning.

Mechanism	**Variable**	**Path**	**Fixed across N**	**Activation**
Cascaded warm-up	β, λ early transient	operational	yes	First $\sim$5 000 ep. freeze on (β,λ); further $\sim$1 500 ep. freeze on λ alone.
Sub-sampled dual cadence	λ frequency, η_k indexing	operational	yes	Dual fires every 5 episodes; η_k indexed by update count.
Dual-gradient soft clip	per-update λ increment	operational	yes	Each $\eta_k(\widehat{\Xi}_\alpha^{\mathcal{B}}-C_\alpha^{\text{eff}})$ increment soft-clipped at magnitude 1.5 before projection onto $[0,\lambda_{\max}]$.
Rolling-window evaluation averages, $W=8\,000$	$\bar\beta$, $\bar\lambda$, $\bar V_0$, and reported evaluation traces	evaluation only	yes	Window-mean used for the gate criterion $\lvert\bar V_0-\alpha\rvert$ and reported greedy CVaR; not in the Bellman update.
Slack-driven recovery	λ-update counter (always); λ iterate (gated on slack sign)	operational	yes	*Trigger:* post-warmup, sustained slack $\widehat{\Xi}_\alpha^{\mathcal{B}}-C_\alpha^{\text{eff}}$ over the Polyak window. *Action 1:* Kushner–Yin reset of the η_k counter. *Action 2 (over-feasible branch only):* one-shot downward kick on λ.

Rationale. The cascaded warm-up prevents (β,λ) from reacting to a near-empty joint table before (Q^R,Q^C,V) populate enough cells to yield a non-trivial selector. The sub-sampled dual cadence matches η_k to the rate at which $\widehat{\Xi}_\alpha^{\mathcal{B}}$ actually moves: between consecutive updates the buffer gains ~ 5 new π_b-returns, so episode-indexed η_k would decay before mixing. The dual-gradient soft clip bounds single-update excursions in λ during the cascaded warm-up; projection onto $[0,\lambda_{\max}]$ is unchanged, so the SA fixed point is not altered. Polyak averaging is confined to the evaluation path because averaging (β,λ) in the Bellman update would conflict with the $a_k \gg \gamma_k \gg \eta_k$ separation Algorithm 4.1 relies on. Slack-driven recovery counters a finite-time lag in the dual signal: in an overly conservative regime the buffer is dominated by feasible low-cost episodes and the ordinary negative dual gradient

may be too weak or too delayed to reduce λ; the recovery resets the η_k counter to re-energise the slow update, and on the over-feasible branch applies a one-shot downward kick on λ. The η_k-counter reset alters the effective step-size schedule and the downward kick is a state-dependent intervention, so the long-run impact of recovery on the projected dual equilibrium is outside the scope of this thesis; the C1+C2+C3 Bellman recursions and the buffer estimator are unchanged, with only the dual-ascent trajectory of λ_k affected.

At $N = 3$ the slack-driven recovery's empirical firing pattern is reported in Chapter 7 once the per-cell record is in place; the construction in this chapter does not anticipate that record.

6.6 Frozen protocol after $N = 1$ calibration

The sweep is $N \in \{1, 2, 3\}$, five seeds per cell, hyperparameters fixed across N (Table 6.2). *All hyperparameters, gate tolerances, per-seed budget, and non-convergence rule are calibrated on the $N = 1$ centralised setting of this chapter and then frozen for $N = 2$ and $N = 3$; no N-specific retuning is used.* This is referred to throughout as the *frozen-protocol convention* of Chapter 6. The per-seed budget is 6×10^6 episodes; seeds that have not satisfied the five-component KKT-based operational gate within budget are recorded as non-convergent and reported, not discarded. Per cell each run records episodes-to-gate, achieved $\widehat{\Xi}_\alpha^{\mathcal{B}}$ vs. C_α^{eff}, the converged (β, λ), and the trajectory-level statistics consumed by Chapter 8.

Gate convention. The single-agent validation of Chapter 5 and the centralised stopping rule of this chapter are separately calibrated finite-time diagnostics: the former validates the corrected architecture, the latter is the frozen stop for the N-sweep under fixed compute. They are not used as directly comparable numerical certificates; the scaling comparison requires invariance *within* the centralised protocol — the same gate, hyperparameters, budget, and logging rule across $N \in \{1, 2, 3\}$ and across all seeds. The Chapter 6 gate is two-tier because two structurally distinct categories of finite-time error live on incompatible scales. *SA-convergence components* (G_1, G_4, G_5) test whether the iterates have stopped moving on quantities bounded independently of N ($V_0 \in [0, 1]$ at target α; $\bar{\lambda}$-drift floored by the Polyak window; the Bellman residual on Q^R, Q^C, V), and take N-invariant absolute tolerances. *Saddle-quality components* (G_2, G_3) test feasibility of the SA fixed-point policy: both the evaluation CVaR and the saddle gap scale with $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$, while the sampling SE on the evaluation CVaR scales as $\sqrt{N}$ under the factored MMDP, so a relative tolerance of 5% of C_α^{eff} yields a margin-to-noise ratio that *grows* with N — the unit-correct choice; G_3 uses 10% for complementary slackness, matching Tamar et al. (2015) and Chow and Ghavamzadeh (2014). The specific tolerances are recorded in Table 6.2; the formal G_1–G_5 definitions and the $N = 1$ bit-identity of G_2 to the Chapter 5 single-agent feasibility band are stated in Section 7.1 alongside the empirical record. The remaining gate components are calibrated separately for the centralised stress test; in particular, β-stationarity here uses $\tau_V = 0.05$ rather than

the tighter ±0.020 band of Chapter 5, so episodes-to-gate values are not directly comparable across the two chapters.

Table 6.2. Hyperparameters for the centralised stress test of Chapters 7–8, calibrated on the $N = 1$ centralised setting of this chapter and then fixed across N. $C_{\text{base}} = 1.60$, $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$. All thresholds in this table belong to the centralised stress-test protocol and are fixed across N and seed; they are not a relaxation or tightening of the Chapter 5 single-agent validation gate.

Parameter	Value
Risk level α	0.10 for the centralised stress test; $\{0.05, 0.10, 0.20\}$ on the calibrated grid of Figure 5.3.
Environment	N-agent Taxi MMDP; 3 lanes with shared 12-step column/time index; centre-lane start $s_{i,0} = 1$ at column 0; horizon $T = 12$ joint decisions; hazard-band-conditioned lane kernel; no per-agent goal absorption. Full per-agent reward, hazard, and transition schedules in Appendix A.
Joint state / action spaces	$\lvert\mathcal{S}^{(N)}\rvert = 3^N$, $\lvert\mathcal{A}^{(N)}\rvert = 3^N$.
N sweep	$\{1, 2, 3\}$; five seeds per cell; $N = 3$ recorded as the first demonstrated stall cell.
Per-seed budget	6×10^6 episodes.
Step-size exponents $(a_{\text{exp}}, \gamma_{\text{exp}}, \eta_{\text{exp}})$	(0.60, 0.70, 0.75); $0.5 < a_{\text{exp}} < \gamma_{\text{exp}} < \eta_{\text{exp}} \leq 1$ (Borkar, 2008).
Step-size scales (a_0, γ_0, η_0) and offset	(0.30, 0.10, 1.80) with unit offset, i.e. $a_k = a_0/(1+k)^{a_{\text{exp}}}$ and analogously for γ_k, η_k. The single-agent Chapter 5 calibration uses the same exponents and offset with $\eta_0 = 3.90$; the centralised stress-test value $\eta_0 = 1.80$ is fixed across $N \in \{1, 2, 3\}$.
Cascaded warm-up	(β, λ) frozen $\sim$ 5 000 ep.; λ alone frozen $\sim$ 1 500 ep.; full three-timescale recursion thereafter.
Dual update cadence	Every 5 episodes; η_k indexed by update count; per-update gradient soft-clipped at magnitude 1.5.
Dual signal source	R–U buffer-mean $\widehat{\Xi}_\alpha^{\mathcal{B}}$ (Eq. (4.17)); $W_{\min} = 512$.
Buffer / Polyak window W	$W = 8\,000$, used jointly as buffer capacity and Polyak window for $(\bar{\beta}, \bar{\lambda}, \bar{V}_0)$.
Dual upper clip $\lambda_{\max}$	200.
Exploration	ϵ-greedy, $\epsilon = 0.05$; per-agent independent for $N \geq 2$ (Section 6.3, Proposition 6.1).
Polyak averaging usage	Greedy evaluation only; not in operational Bellman path.
Convergence gate	Five-component KKT-based operational gate (G_1–G_5), two-tier (see *Gate convention* above), on rolling window W. *Cat. A (N-invariant absolute):* $\mathbf{G}_1$ (β-stationary) $\lvert\bar{V}_0 - \alpha\rvert \leq 0.05$; $\mathbf{G}_4$ (dual stationary) $\bar{\lambda}$-drift over $W \leq 0.10$ with $\bar{\lambda} < \lambda_{\max}$; $\mathbf{G}_5$ (Bellman-stability diagnostic, single threshold across N) residual on $\{Q^R, Q^C, V\} \leq 0.06$. *Cat. B (relative to C_α^{eff}):* $\mathbf{G}_2$ (primal feasibility, one-sided) $\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}} \leq 0.05\, C_\alpha^{\text{eff}}$; $\mathbf{G}_3$ (complementary slackness, active branch $\bar{\lambda} \geq 5.0$) $\lvert\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}}\rvert \leq 0.10\, C_\alpha^{\text{eff}}$. Streak 5, log every 1 000 episodes.
Calibration discipline	Hyperparameters, gate tolerances, per-seed budget, non-convergence rule fixed at the $N = 1$ centralised calibration of this chapter; reused without retuning for $N = 2, 3$. Failed seeds reported, not discarded.
Cost smoothing	Episode-end shock $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, $\sigma_\xi = 0.10$; added to Y^k for buffer / terminal indicator only; not in Q^C stage source (Section 6.2).
Max per-step cost per agent	0.50 (hazard 0.40, lane-switch 0.10; additive).

Budget and seed count. Under the fixed compute budget we prioritise within-cell seed agreement over extending a smaller number of runs. Five seeds are run *within* each $N \in \{1, 2, 3\}$ cell, and the seed index is not aligned across N: $\mathcal{S}^{(N)} \times \mathcal{A}^{(N)}$ differs across N, so a common seed value would not produce comparable trajectories. Seed-level comparison is therefore made within a cell; cross-N comparison is made on five-seed aggregates. The protocol distinguishes a systematic stall from seed

pathology rather than estimating an episode budget at which N=3 would eventually clear.

We do not sweep $N \geq 4$. Larger lifts were considered during preliminary scale selection but excluded once N=3 produced a seed-consistent, budget-limited stall under the frozen protocol. Both Chapter 8 axes, $|\mathcal{S}^{(N)}|\cdot|\mathcal{A}^{(N)}|$ and $\mathrm{Var}(Y)$, grow monotonically in N, so additional cells would be uniformly stalled rather than informative about a scaling exponent or about which mechanism dominates. A slope fitted to one or two further budget-constrained cells would be underpowered. We therefore present N=3 as the first demonstrated stall cell; the resolved scaling slope is deferred to the per-agent decomposition, whose axes scale with $|\mathcal{S}_0|\cdot|\mathcal{A}_0|$ and $\mathrm{Var}(C_i)$.

Limitations. Generality to other weakly coupled MMDPs rests on the analytic form of the two mechanisms in Chapter 8; environment sweeps and function approximation are out of scope; the five-seed protocol distinguishes stall from seed pathology but does not measure an N-scaling exponent; and the per-capita budget convention $C_\alpha^{\mathrm{eff}} = N \cdot C_{\mathrm{base}}$ holds the joint cost mean on the same scale across N but does not hold standardised tail tightness constant under weak per-agent correlation, a residual Chapter 8 discusses.

7

Empirical Results

The C1+C2+C3 learner of Chapter 4 is a single-agent fix on three separable failure modes of the original CVaR-side recursion. This chapter asks the next question: does the fix remain operational once the learner is lifted, without further modification, into the centralised representation of Chapter 6? The answer is finite-time and protocol-specific. Under the frozen protocol, $N = 1$ and $N = 2$ reach the operational gate within budget on every seed; $N = 3$ does not, and fails on the same axes by margins that close on a single operating point across seeds. The chapter does not argue that risk-aware multi-agent learning scales — it constructs the empirical object that Chapter 8 explains.

What we report is structured around two requirements: the protocol must be tight enough that a failure can be attributed to the joint representation rather than to per-N tuning, and the empirical record at $N = 3$ must support the two structural facts that Chapter 8 is required to explain — and no more. We therefore separate the per-cell narrative from the bridge to the mechanism reading, and bound the record's scope explicitly before handing it off.

7.1 Operational gate, signals, and saturation diagnostics

The gate is a finite-time operational stop, not a KKT condition of the original problem. The KKT conditions are the first-order optimality conditions for a constrained problem: primal feasibility, dual feasibility ($\lambda \geq 0$), complementary slackness, and Lagrangian stationarity. The Lagrangian relaxation (4.10) has dual $\lambda \geq 0$ and no upper bound; the clip $\lambda_{\max} = 200$ is a finite-time stabilisation introduced in Section 6.5 to keep the slow update on its prescribed timescale during the cascaded warm-up. The five quantities monitored at termination correspond to the residuals that vanish at a KKT point of the unclipped problem; we evaluate them in finite time against fixed tolerances and add the clip-saturation indicator $\bar{\lambda}/\lambda_{\max}$. We refer to the composite as the *five-component KKT-based operational gate* (*gate* for short throughout this chapter); the term is a working name for a finite-time diagnostic specific to this thesis rather than a name from the literature. Informally — and this is the reading to keep in mind — gate clearance is the finite-time check that the deployed policy is approximately feasible *and* the iterates have stopped moving on quantities that would be stationary at an unclipped saddle. It is not a KKT

stop in the textbook sense: λ is clipped, the gate is evaluated on Polyak-averaged window quantities, and tolerances are non-zero. The writing reflects this distinction throughout.

The five components. The gate components G_1–G_5 and their tolerances are listed in Table 6.2 (Section 6.6, *Gate convention*); we reproduce the structure here only to anchor the saturation diagnostic that follows. On a per-seed Polyak window of width $W = 8\,000$ used in greedy-policy evaluation (Section 6.5), the five components split into two categories with structurally distinct tolerance scales. *Category A* comprises the three SA-convergence components G_1 (β-stationarity, $|\bar{V}_0(z_0) - \alpha| \le \tau_V = 0.05$), G_4 (dual stationarity, $|\Delta\bar{\lambda}_W| \le \tau_\lambda = 0.10$ with $\bar{\lambda} < \lambda_{\max}$), and G_5 (Bellman residual on $\{Q^R, Q^C, V\}$, $\le \tau_B = 0.06$), each on a quantity bounded independently of N and therefore on an N-invariant absolute tolerance. *Category B* comprises the two saddle-quality components G_2 (one-sided primal feasibility, $\widehat{\Xi}^{\mathcal{B}}_\alpha - C^{\text{eff}}_\alpha \le \tau_F$) and G_3 (active-branch complementary slackness, $|\widehat{\Xi}^{\mathcal{B}}_\alpha - C^{\text{eff}}_\alpha| \le \tau_C$ under $\bar{\lambda} \ge 5$), each with a relative tolerance scaled to C^{eff}_α: $\tau_F = 0.05\, C^{\text{eff}}_\alpha$ and $\tau_C = 0.10\, C^{\text{eff}}_\alpha$, matching the unit of Y and the CVaR-MDP literature (Tamar et al., 2015; Chow and Ghavamzadeh, 2014). In G_4, $\Delta\bar{\lambda}_W$ denotes the Polyak-window range $\max_{t\in[k-W+1,\,k]} \bar{\lambda}_t - \min_{t\in[k-W+1,\,k]} \bar{\lambda}_t$, which is non-negative by construction and renders the absolute-value bars redundant; they are kept for notational symmetry with the $|\bar{V}_0 - \alpha|$ form of G_1.

A seed stops when all five conditions hold for five consecutive log decisions. At $N = 1$, $\tau_F = 0.080$ recovers the absolute single-agent value of Chapter 5, so the multi-agent gate is bit-identical to the single-agent feasibility band at $N = 1$ and relaxes for $N \ge 2$ only in the unit-correct direction (Section 6.6).

Why these five. Each component carries information the others miss: feasibility alone clears a clip-frozen $\bar{\lambda}$ that maintains feasibility only by truncation; SA stationarity alone clears a stationary but infeasible iterate; complementarity without the staleness checks clears a saddle reached only by oscillation. The two-tier split applies these conditions in the units in which each underlying quantity lives.

Two signals, two distinct estimators of CVaR. The constraint is enforced on the executed behaviour policy π_b. The slow update consumes the buffer-mean estimator (4.17), $\widehat{\Xi}^{\mathcal{B}}_\alpha$, computed on the rolling π_b-return buffer $\mathcal{B}$ of capacity W. The greedy-policy evaluation computes $\widehat{\Xi}^{\text{held}}_\alpha$ on independent rollouts of $\pi_g = \arg\max J$. The two estimators are consistent for different policies: $\widehat{\Xi}^{\mathcal{B}}_\alpha$ targets $\Xi^{\pi_b}_\alpha$ and $\widehat{\Xi}^{\text{held}}_\alpha$ targets $\Xi^{\pi_g}_\alpha$. Their finite-time gap reflects two sources, not mutually exclusive. The first is the behaviour–greedy gap $\Xi^{\pi_b}_\alpha - \Xi^{\pi_g}_\alpha$: under per-agent independent ϵ-greedy the per-step joint-action disagreement is $1 - (1 - \epsilon + \epsilon/|\mathcal{A}_0|)^N$ by Proposition 6.1, growing with N. The second is the π_b-drift bias of the buffer mean when π_b is not quasi-stationary on the window scale, bounded above by the total-variation argument of Section 8.3. The two estimators remain consistent for their respective policy targets in the limit (Section 4.4); in finite time the gap is the empirical record, and the present protocol does not separate the two sources. The gate component

G_2 is evaluated on $\widehat{\Xi}_\alpha^{\mathcal{B}}$, since that is the signal the slow update operates on; $\widehat{\Xi}_\alpha^{\text{held}}$ is diagnostic only, reported alongside in Figure 7.1b. Conflating the two would silently substitute a held-out summary for the training-time constraint.

The saturation diagnostic that replaces λ-drift. Once $\bar{\lambda}$ reaches the upper clip, the increment $\Delta\bar{\lambda}_W$ measured on the Polyak window is suppressed by truncation rather than by stationarity, so G_4 reduces to a numerical floor that is not informative about convergence. We therefore monitor, alongside G_4, the un-clipped dual proposal

$$u_k \;:=\; \bar{\lambda}_{k-1} \;+\; \eta_k \left(\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}}\right), \tag{7.1}$$

and the clip-saturation indicator $\bar{\lambda}/\lambda_{\max}$. When the guard is at unity and $u_k > \lambda_{\max}$ persistently, the slow update is requesting a larger penalty than the clip permits: the dual is not stationary, it is censored. This is the correct reading at $N = 3$. We continue to report G_4 for completeness, but treat it as uninformative once the clip-saturation indicator reaches unity, and infer the dual side from u_k, the clip-saturation indicator, and G_3 jointly.

Per-seed gate vs. five-seed aggregates. G_1–G_5 are evaluated per-seed on running operational tolerances. The five-seed summaries below are post-termination diagnostics: post-gate windows for $N = 1, 2$ and end-of-budget windows for $N = 3$. The two CVaR estimators agree within the five-seed dispersion in the small-N cells but are reported separately so that the signal each summarises is unambiguous.

7.2 Result summary

Table 7.1 records the per-cell outcome on the diagnostic axes that Chapter 8 carries forward: seeds gated, episodes-to-gate, post-termination operating point $(\bar{\beta}, \bar{\lambda})$, the buffer feasibility consumed by the slow update, the held-out feasibility on π_g, the β-stationarity residual, the maximum Bellman residual on the cost Q-function, and the count of recovery activations invoked by the over-feasibility procedure of Section 6.5. The table is the chapter's central numerical record; the graphical complement is Figure 7.1, which plots episodes-to-gate dispersion (panel a, top), the time-resolved gate residuals at $N = 3$ (panel a, middle), the buffer-versus-held-out feasibility on the slow update's signal (panel b), and the end-of-budget operating points in the (β, λ) plane (panel c). Read the table and the figure together: the table fixes the end-of-budget numbers and the figure fixes the trajectories that produced them.

N	gated /5	episodes-to-gate median (range, 10^6)	$(\bar{\beta}, \bar{\lambda})$ median	$\widehat{\Xi}_\alpha^{\mathcal{B}}/C_\alpha^{\text{eff}}$ median	$\widehat{\Xi}_\alpha^{\text{held}}/C_\alpha^{\text{eff}}$ median	$\lvert\bar{V}_0 - \alpha\rvert/\tau_V$ median	Bellman residual/τ_B median	recov. range
1	5	0.36 (0.22–0.78)	(1.44, 53.6)	1.03	1.04	0.7	<0.1	0
2	5	0.72 (0.65–0.89)	(2.90, 88.8)	1.03	1.05	0.9	0.3	0
3	0	≥ 6.0 *censored*	(4.33, 200)	1.16	1.04	1.5–1.8	4.5–5.6	62–74

Table 7.1. Five-seed summary at $\alpha = 0.10$, $C_{\text{base}} = 1.60$, per-seed budget 6×10^6 episodes. Episodes-to-gate is the per-seed gate firing time; the $N = 3$ row is censored at the per-seed budget. $\bar{\beta}, \bar{\lambda}$ are Polyak-averaged values on the reporting window: post-gate for $N = 1, 2$ and end-of-budget for $N = 3$. $\widehat{\Xi}_\alpha^{\mathcal{B}}$ is the buffer-mean estimator on the post-gate window for $N = 1, 2$ and on the end-of-budget window for $N = 3$; this is the signal consumed by the slow update. $\widehat{\Xi}_\alpha^{\text{held}}$ is held-out greedy evaluation on π_g. Both feasibility columns are reported relative to budget; the upper feasibility tolerance is $\widehat{\Xi}_\alpha^{\mathcal{B}}/C_\alpha^{\text{eff}} \leq 1.05$, and the shaded $\pm5\%$ band in Figure 7.1b is diagnostic. The Bellman residual column reports the residual reduced uniformly across N: per seed, $\max\{\|\Delta Q^R\|_\infty, \|\Delta Q^C\|_\infty, \|\Delta V\|_\infty\}/\tau_B$ is taken as a median over the last 20% of logged decisions on the archived diagnostic path; cells then report the five-seed median ($N = 1, 2$) or the per-seed range ($N = 3$, censored).

7.3 Gate satisfaction at $N = 1$ and $N = 2$

The numerical record for this section is the $N = 1$ and $N = 2$ rows of Table 7.1; the prose that follows is anchored to those rows.

Every seed at $N = 1$ and $N = 2$ clears the operational gate within the per-seed budget. The two cells return the same five-seed picture qualitatively, and three features are worth recording for the contrast with $N = 3$.

Episodes-to-gate. The five-seed median is 0.36×10^6 at $N = 1$ and 0.72×10^6 at $N = 2$ (Table 7.1, Figure 7.1a, top). Each median lies an order of magnitude inside the per-seed budget; the within-cell range is tight on every seed. Between $N = 1$ and $N = 2$, the observed median roughly doubles, against a centralised budget that has doubled, a joint Q^C table whose implemented augmented size has grown by a factor of forty (Section 8.1, under the implementation shape of Figure 6.1), and a standardised budget-above-mean margin[1] $(C_\alpha^{\text{eff}} - \mathbb{E}[Y])/\sqrt{\text{Var}(Y)}$ that widens by $\sqrt{2}$ under the convention $C_\alpha^{\text{eff}} = N\, C_{\text{base}}$ (Section 7.6 for the regime). The two-cell observation does not establish a scaling law; it records that, at this budget and on these two cells, the $N = 2$ cell costs about twice the $N = 1$ cell to gate, despite the constraint being statistically looser in standard-deviation units.

Feasibility on the signal that drives the slow update. The buffer feasibility, which is the signal $\widehat{\Xi}_\alpha^{\mathcal{B}}$ the slow update operates on, lies within the relative band on both cells at end of budget: ratio 1.03 at $N = 1$ and 1.03 at $N = 2$ against the gate's 5 % band (Table 7.1). Per-seed gate decisions are made on running operational tol-

[1] We use "standardised" in the elementary z-score sense throughout this thesis: a margin scaled by the joint return's standard deviation. The quantity $(C_\alpha^{\text{eff}} - \mathbb{E}[Y])/\sqrt{\text{Var}(Y)}$ expresses budget headroom in units of $\sqrt{\text{Var}(Y)}$ and is the natural standardisation for comparing constraint tightness across cells with different $\text{Var}(Y)$; it is not a defined term of the CVaR-MDP literature, and we cite none for the standardisation itself.

erances and are individually inside the band on every seed; the post-gate summaries reported here are evaluated on later windows and may differ slightly from the exact stop-window values. Held-out greedy evaluation gives 1.04 at $N = 1$ and 1.05 at $N = 2$; the $N = 2$ cell lies at the band edge and $N = 1$ just inside. The buffer-versus-held-out gaps at these cells are ≤ 0.02 relative, small in absolute units; under product-form ϵ-greedy the behaviour–greedy gap alone is bounded by the per-step joint-disagreement $1-(1-\epsilon+\epsilon/|\mathcal{A}_0|)^N$ of Proposition 6.1, which grows with N and accounts for the magnitude observed here without requiring a π_b-drift attribution at $N \in \{1, 2\}$.

Operating point and seed agreement. The endpoint duals are $(\bar{\beta}, \bar{\lambda}) = (1.44, 53.6)$ at $N = 1$ and $(2.90, 88.8)$ at $N = 2$ (Figure 7.1c). $\bar{\beta}$ scales linearly per capita ($\bar{\beta}/N \approx 1.44$ at both cells), tracking the centralised budget per agent without pathological inflation; $\bar{\lambda}$ moves with the budget but stays well below $\lambda_{\max} = 200$, so the clip-saturation indicator $\bar{\lambda}/\lambda_{\max}$ never approaches unity. Per-seed scatter is tight relative to the between-cell separation, with $\bar{\lambda}/\lambda_{\max}$ spread at a few per cent across seeds, and the per-seed Polyak-window drift $|\Delta\bar{\lambda}_W|$ lies comfortably inside $\tau_\lambda = 0.10$. Algorithm 4.1 returns reproducibly to a tightly clustered operating point at this scale; most of the seed-level dispersion is on the trajectory by which the cluster is reached, while the cluster itself is tight.

7.4 $N = 3$: a seed-consistent stall on every gate axis

The numerical record for this section is the $N = 3$ row of Table 7.1; the trajectory-level record is the middle panel of Figure 7.1a. The two together fix the stall pattern this section describes.

The operational gate fires on no seed at $N = 3$ within the per-seed budget. Termination is at the budget boundary on every seed; the end-of-budget operating point is the same within seed dispersion; each G_1, G_2, G_3, G_5 residual is missed in the same direction by a margin of the same order. G_4 is uninformative under saturation, and is interpreted against u_k and the clip-saturation indicator as set out in Section 7.1. The cell is reported as a censored stall at the budget boundary; the simultaneity of the missed components and the order-of-magnitude separation from $N \in \{1, 2\}$ are signatures of a regime change that a scaling fit through three points would smooth over. The components are taken in turn; the seed-level coherence is collected at the end.

The dual saturates and stays saturated. $\bar{\lambda}$ climbs through the cascaded warm-up and reaches $\lambda_{\max} = 200$ near 5×10^5 episodes. From there on, on every seed, it does not leave the boundary for the remaining 5.5×10^6 episodes. The clip-saturation indicator $\bar{\lambda}/\lambda_{\max}$ (Figure 7.1a, middle) remains at unity over this window. The unclipped proposal (7.1) stays above $\lambda_{\max}$ throughout the saturation window: what we observe is censoring of the dual trajectory, not stationarity of the unclipped

dynamics. The slack-driven recovery procedure of Section 6.5 fires 62–74 times per seed at $N = 3$, against zero firings at $N \in \{1, 2\}$. Throughout the saturation window $\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}} > 0$, so the conditional downward-kick branch is inert and the firings are η_k-counter resets that re-energise ascent rather than detach $\bar{\lambda}$ from the upper clip. The recoveries do exactly what the procedure was designed for in this regime; they are not a near-convergence event. The signal that would release $\bar{\lambda}$ from the boundary is the same window-to-window fluctuation in $\widehat{\Xi}_\alpha^{\mathcal{B}}$ that drives the recoveries themselves. The active-branch complementarity component G_3 is missed by roughly 1.6× its tolerance (end-of-budget median ≈ 0.76 against $0.10\, C_\alpha^{\text{eff}} = 0.48$, Table 7.1); the corroborating textbook KKT complementarity product $\bar{\lambda} \cdot |\widehat{\Xi}_\alpha^{\mathcal{B}} - C_\alpha^{\text{eff}}|$ stands at order 10^2 at $\bar{\lambda} = \lambda_{\max}$, so the saddle's first-order complementarity condition is far from satisfied. G_4 is uninformative under saturation.

The cost Q-function remains above tolerance along a non-monotone trace. The C1 recursion (4.5) on the joint table does not approach its finite-horizon Bellman fixed point along a controlled trajectory at this budget; the Chapter 4 fixed-point identification at z_0 characterises the target the recursion is failing to attain. The maximum Bellman residual on (Q^C, Q^R, V) traces a non-monotone path (Figure 7.1a, middle, orange): a warm-up peak above $12\,\tau_B$, descent through the post-saturation phase to within tolerance briefly, re-rise through the second half of the budget to settle at 4.5–5.6 τ_B on the late tail. The mid-run dip is not gate satisfaction: the operational gate requires simultaneous control of all five normalised residuals on a streak, and a transient single-component dip while the other informative residuals remain outside tolerance is not one. The dip co-occurs with $\bar{\lambda}$ reaching saturation, after which the bootstrap selector $a^\star$ is driven by a near-fixed penalty rather than a responsive dual signal, and reverses as the executed π_b continues to drift over a cost Q-function whose residual remains above tolerance.

The dual signal carries an infeasibility bias on the signal that matters. The buffer-mean estimator $\widehat{\Xi}_\alpha^{\mathcal{B}}$ on the $W = 8\,000$ rolling window stays above $C_\alpha^{\text{eff}} = 4.80$ throughout the post-warm-up phase on every seed, with five-seed median end-of-budget ratio 1.16 (Table 7.1); in absolute terms the buffer CVaR lies near 5.56, about 0.76 above the budget. The held-out evaluation on π_g lies at ratio 1.04, just inside the 5 % band: the deployed greedy policy is close to feasibility, while the buffer that drives the slow update reports π_b at sixteen per cent above budget. Two mechanisms, not mutually exclusive, can produce a buffer-versus-held-out gap of this magnitude. The first is the standard behaviour–greedy gap: under per-agent independent ϵ-greedy with $\epsilon = 0.05$ and $|\mathcal{A}_0| = 3$, the per-step joint-action disagreement probability is $1 - (1 - \epsilon + \epsilon/|\mathcal{A}_0|)^N \approx 0.097$ at $N = 3$ by Proposition 6.1, about three times the corresponding $N = 1$ disagreement probability ≈ 0.033 under the same product-form rule. A simulation-lemma bound on $\Xi_\alpha^{\pi_b} - \Xi_\alpha^{\pi_g}$ in a bounded-cost MMDP is large enough at this disagreement probability that the behaviour–greedy gap cannot be ruled out as a contributor. The second is the π_b-drift bias of the buffer mean on a non-quasi-stationary window under the saturated dual — the mechanism Chapter 8 examines directly through the total-variation bound on the buffer estimator (Section 8.3). The present protocol does not separate the two contributions

empirically; we report the total gap and the two mechanisms that can produce it rather than a numerical decomposition. The structural reading is the same under either attribution: at $N = 3$ the signal that drives the slow update reports infeasibility with a stationary signature across all five seeds, while the deployment signal lies near the band edge. The signal capable of releasing λ from the upper clip is the buffer-side dual gradient itself; under the protocol and at $N = 3$, that signal does not become reliably negative within budget.

Tail probability misses target throughout the run. The β-subsystem has not equilibrated. The Polyak-averaged tail probability $\bar{V}_0(z_0)$ does not stabilise inside the gate band: the normalised residual $|\bar{V}_0 - \alpha|/\tau_V$ floats above unity through most of the post-warm-up phase, dips into tolerance briefly as the dual reaches the boundary, and settles at 1.5–1.8 on the median by end of budget (Figure 7.1a, middle, green trace). The un-averaged V_0 swings on consecutive rolling windows by an order of magnitude: the β-recursion receives an inconsistent V_0 across the run, and Polyak averaging smooths the trace without bringing its level inside the band. The endpoint $\bar{\beta} = 4.33$ in Figure 7.1c has per-capita value $\bar{\beta}/N \approx 1.44$, agreeing with $N \in \{1, 2\}$; this apparent linearity is a window average over an unequilibrated trace rather than a VaR_α readout, and $\bar{\beta}$ reaches a numerically plausible level only as the average of a wide-amplitude oscillation.

Seed-level coherence. The four time-resolved residuals in Figure 7.1a (middle) lie on the same qualitative trajectory across seeds. The seed-to-seed range in each end-of-budget residual is small relative to its own median: for the Bellman residual, range $\approx 1.2\,\tau_B$ around a median of $\approx 5.2\,\tau_B$ (relative range ≈ 0.22); for $|\bar{V}_0 - \alpha|/\tau_V$, range ≈ 0.3 around 1.8; for $\bar{\lambda}$, all five seeds at 200 to numerical precision. The end-of-budget $(\bar{\beta}, \bar{\lambda})$ scatter at $N = 3$ in Figure 7.1c forms a single tight cluster on the $\lambda_{\max}$ line, and the achieved $\widehat{\Xi}_\alpha^{\mathcal{B}}$ forms a similarly tight cluster above the gate band. Under five independent seeds every run terminates at a tightly clustered operating point with a consistent residual pattern on every axis: this is consistent with structural rejection rather than seed-level variability. The operational hypothesis under test — that C1+C2+C3 lifts unmodified into the centralised representation — is not supported at $N = 3$ under this protocol.

The N=1 calibration suggested that the gate-clearance budget would grow with N in a monotone, tractable way — larger joint table, longer warm-up, eventual convergence. The five-seed picture at N=3 contradicts that expectation. The five runs do not differ on whether the gate is missed but on by how much, on the same axes, in the same direction; the cluster on the $\lambda_{\max}$ line in Figure 7.1c is what makes the bottleneck question of Chapter 8 unavoidable. Whatever is failing at N=3 is failing in a way that additional compute on the same representation will not relocate, and the diagnosis that follows attributes the failure to quantities of the joint representation rather than of the algorithm.

7.5 Two empirical facts handed to Chapter 8

The record of Sections 7.3 and 7.4 yields two facts and fixes what they cannot be attributed to. Seed dispersion does not relocate the stall: five independent seeds cluster on a single operating point on the $\lambda_{\max}$ line, with end-of-budget residuals agreeing on every gate component. The exploration rule does not relocate it either: per-agent independent ϵ-greedy of Section 6.3 clears the gate at $N \in \{1, 2\}$ under the same specification, and a joint-ϵ-greedy lift would reduce the controlled Hamming-1 coverage of Section 6.4 without touching the joint Q-function-table volume. The mechanism reading of Chapter 8 is the explanation each fact requires.

Fact 7.1 (cost Q-function). Under the 6×10^6-episode per-seed budget, the C1 recursion (4.5) on the joint cost Q-function of size $|S|^N|A|^N$ does not reach the gate's Bellman tolerance at $N = 3$, whereas it does at $N = 1$ and $N = 2$. The binding constraint is the per-cell visit budget; the fixed point itself is correctly identified by the C1 recursion.

Fact 7.2 (dual signal). The buffer-mean estimator $\widehat{\Xi}^{\mathcal{B}}_{\alpha}$ on the fixed-capacity buffer of π_b-returns drives the slow update to and against the upper clip $\lambda_{\max}$ at $N = 3$, and stays there. The only signal capable of releasing λ from the boundary is the buffer-side dual gradient $\widehat{\Xi}^{\mathcal{B}}_{\alpha} - C^{\text{eff}}_{\alpha}$; under the protocol and at $N = 3$, this signal does not become reliably negative, so the clipped dual variable is not released within budget.

Fact 7.1 names a quantity that scales as $|S|^N|A|^N$ and is located on the structural side of Q^C. Fact 7.2 names a quantity whose independent-agent baseline term scales as $\sum_i \mathrm{Var}(C_i)$ on the centralised buffer (Eq. (8.1)), with additional contributions from $\mathrm{Cov}(C_i, C_j)$ and the smoothing-shock intercept σ^2_{ξ}, and is located on the dual side. Their analytic drivers are disjoint; their finite-time coupling, through the executed π_b, is the subject of Chapter 8. Neither fact is affected by C1, C2, or C3 considered separately, since each of these addresses a single-agent failure mode that the centralised lift preserves rather than introduces.

7.6 Limitations and scope

The protocol is five seeds per cell at a fixed per-seed budget of 6×10^6 episodes, in a tabular MMDP with deterministic per-step costs, factored transitions, and a shared episode-end shock. The chapter's claim is finite-time, protocol-specific, and qualitative: gate satisfaction at $N \in \{1, 2\}$, gate failure at $N = 3$, with a consistent residual pattern across five seeds at the failed cell. Four boundaries are worth naming.

Scope: stall, not slope. A measured scaling exponent is not on the page. The chapter documents a stall at $N = 3$, not a slope; extracting one would require either a longer per-seed budget on the same cell or additional cells at $N \geq 4$, both of which compound the budget pressure that already censors $N = 3$ at the present setting. The mechanism-level scaling claim — that the stall axes are monotone in N — is

treated in Chapter 8 on the structural quantities $|S|^N|A|^N$ and $\sum_i \mathrm{Var}(C_i)$, not on a measured curve. A properly resolved slope across $N \geq 4$ is outside the present chapter and would more properly belong to the per-agent decomposition study of Chapter 9.

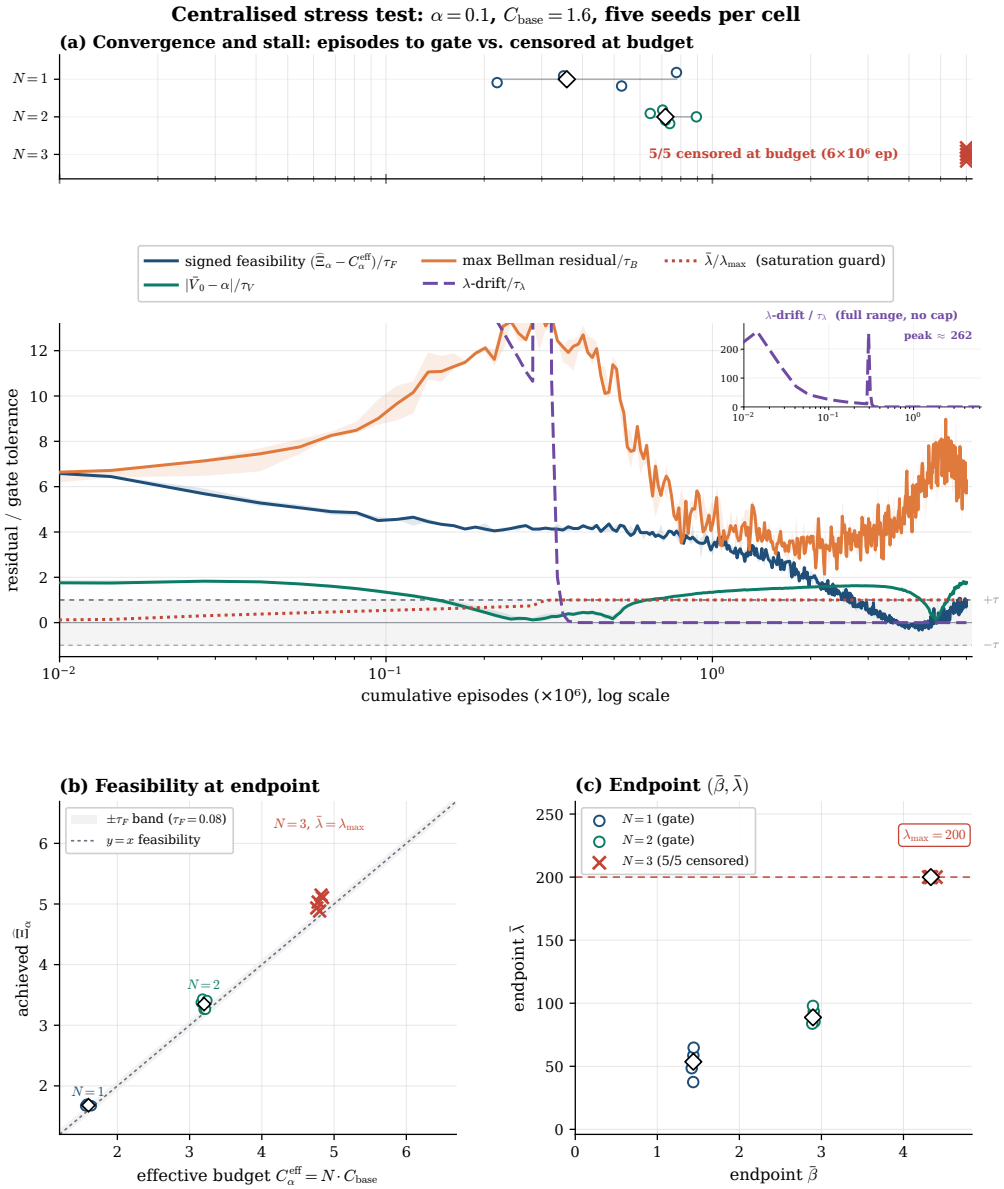

Figure 7.1. Centralised stress test of Algorithm 4.1 at $\alpha = 0.10$, $C_{\text{base}} = 1.60$, per-agent independent ϵ-greedy, five seeds per cell, per-seed budget 6×10^6 episodes. **(a, top)** Episodes-to-gate strip plot: per-seed dots, five-seed median, IQR, and full range for $N = 1, 2$; the $N = 3$ row marks 5/5 censored seeds at the per-seed budget. **(a, middle)** Four time-resolved $N = 3$ gate residuals on log-cumulative episodes, normalised by their gate tolerances of Section 7.1; clip-saturation indicator $\bar{\lambda}/\lambda_{\max}$ overlaid. The inset re-plots $\Delta\bar{\lambda}_W/\tau_\lambda$ at its full unbounded magnitude; the main-panel display caps at $y = 14$ so the G_1, G_2, G_5 traces remain readable, and the inset documents the early-training peak that the cap truncates. **(b)** Buffer feasibility signal $\widehat{\Xi}_\alpha^{\mathcal{B}}$ against the centralised budget C_α^{eff}, with held-out greedy evaluation $\widehat{\Xi}_\alpha^{\text{held}}$ on π_g overlaid; $\pm 5\%$ band shaded. **(c)** Endpoint $(\bar{\beta}, \bar{\lambda})$ locations on the (β, λ) plane with the $\lambda_{\max}$ line marked.

Scope: seed agreement, not tail. Five seeds resolves seed dispersion at $N \in \{1, 2\}$ and confirms seed-level coherence on the failed cell at $N = 3$, but is not enough to estimate the upper tail of episodes-to-gate at $N = 2$ to high precision; within-cell descriptive intervals are correspondingly wide. The five-seed coherence at $N = 3$ is interpreted as agreement on a stall at a tightly clustered operating point, not as a claim about the seed distribution beyond the five seeds run.

Scope: tabular setting. The C1 cost Q-function is enumerated cell by cell. The table-volume reading of Chapter 8 is stated against the representational quantity $|S|^N|A|^N$. Function approximation of Algorithm 4.1, where parameter count and not table size sets per-cell precision, is out of scope here; whether the centralised stall transfers to a parametric approximator under the same per-seed budget is a separate question and is left for the function-approximation extension noted in Chapter 9.

Scope: standardised margin widens with N. The convention $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ keeps the joint constraint per capita constant in N but does not hold the standardised budget-above-mean margin $(C_\alpha^{\text{eff}} - \mathbb{E}[Y])/\sqrt{\text{Var}(Y)}$ constant: under weak per-agent correlation and in the positive-margin regime $C_{\text{base}} > \mathbb{E}[c_i]$, $\mathbb{E}[Y] \propto N$ and $\sqrt{\text{Var}(Y)} \propto \sqrt{N}$. In standard-deviation units the constraint loosens at larger N, so the $N = 3$ stall is not explained by a tightening of the constraint induced by the budget convention. The table-volume mechanism of Chapter 8 is independent of constraint tightness; the dual-noise mechanism depends on both the return variance and the operating point of the constraint, and its finite-time severity at $N = 3$ is therefore observed under a convention that, on the standardised scale, works in its favour. A tail-relative budget scaling that holds the standardised margin fixed across N would be an appropriate robustness experiment for the RCA extension.

8

The Centralised Scaling Bottleneck

The empirical transition between $N = 2$ and $N = 3$ in Chapter 7 is categorical rather than gradual. All seeds clear the operational gate at $N = 2$ within budget; no seed clears it at $N = 3$ within 6×10^6 episodes, while the joint table grows by a further factor of thirty-six between these cells against a fixed per-seed transition budget. The question we take up is not whether the centralised learner slows down but which joint quantities of the representation make the finite-time diagnostic fail.

The two empirical facts isolated in Section 7.5 identify the quantities explained in this chapter. The joint Q^C-table volume has $|S|^N|A|^N$ as its dominant joint state–action factor and grows exponentially in N, setting per-cell Bellman precision through the realised visit count of π_b; the full implemented augmented table additionally carries time and accumulated-cost indices, with the implementation shape developed in Section 8.1. The joint return variance $\mathrm{Var}(Y)$ admits an *independent-agent baseline* obtained by setting the cross-agent covariances to zero in the additive cost decomposition: $\mathrm{Var}(Y) = \sum_i \mathrm{Var}(C_i) + 2\sum_{i<j} \mathrm{Cov}(C_i, C_j) + \sigma_\xi^2$ reduces to $\sum_i \mathrm{Var}(C_i) + \sigma_\xi^2$ when $\mathrm{Cov}(C_i, C_j) = 0$ — the regime of factored transitions under an independent-across-agents reference policy. Under this baseline $\mathrm{Var}(Y)$ is N-linear in the dominant term; the empirical $\mathrm{Var}(Y)$ at $N = 2, 3$ in Figure 8.1b lies above this baseline by an amount attributable to positive policy-induced $\mathrm{Cov}(C_i, C_j)$, an additive correction we report but do not separate out. The quantity sets dual precision through the buffer-mean estimator $\widehat{\Xi}_\alpha$. Neither quantity is touched by C1, C2, or C3; at the asymptotic limit of Borkar and Jain (2014, Theorem 2) the two never meet — C2 decouples the cost Q-function from λ (Proposition 4.1); C3's buffer mean is consistent at the limiting projected dual equilibrium (Section 4.5) under quasi-stationary π_b. They couple in finite time through the executed π_b, whose quasi-stationarity is exactly what the asymptotic theorem rests on.

The numerical anchors are Tables 7.1 and 8.1 together with Figure 8.1: the last two columns of Table 7.1 (the buffer feasibility on the slow update's signal and the maximum Bellman residual on the cost Q-function) are the empirical residuals this chapter's two mechanisms explain, with the mechanism-to-residual mapping itemised in Table 8.1.

The attribution is structural in the sense of Chapter 4: both facts are properties of the centralised representation $(\mathcal{S}^N, \mathcal{A}^N, \mathcal{Y})$ itself, and remain so under any choice of $(a_k, \gamma_k, \eta_k, \lambda_{\max}, W)$.

8.1 Mechanism (1): table-volume dilution of the cost Q-function

The C1 recursion (4.5) is asynchronous stochastic approximation on the joint state–action table. Each episode contributes $T = 12$ updates: twelve cells along the rollout receive a single noisy correction; the rest of the table is untouched. The asynchronous-Q regime of Tsitsiklis (1994) guarantees $Q^k \to Q^\star$ almost surely under $\sum_k a_k = \infty$, $\sum_k a_k^2 < \infty$, and infinitely often visitation of every cell — an asymptotic statement unconditional on N. The finite-time statement rests on the per-cell visit count that π_b produces within budget, and that quantity decays rapidly with N.

Numerical reading. At $N = 2$ the median episodes-to-gate doubles against a forty-fold table growth — the coordination tax expected of a centralised algorithm. At $N = 3$ even an eight-fold lower bound on episodes-to-gate does not track the further thirty-six-fold table growth. The conceptual joint augmented Q-function table scales as $T \cdot |\mathcal{S}|^N \cdot |\mathcal{Y}| \cdot |\mathcal{A}|^N$. The visit-count diagnostic reported below uses the full implementation shape of the logged arrays, $T \cdot |\mathcal{S}|^N \cdot |\mathcal{Y}| \cdot (|\mathcal{A}|^N + 1) \cdot |\mathcal{A}|^N$, in which the additional $(|\mathcal{A}|^N + 1)$ factor is a previous-action slot used only by the tabular implementation — a tabular-index refinement rather than part of the mathematical MMDP state–action space.[1] The implementation shape has 3.3×10^4, 1.3×10^6, and 4.8×10^7 cells at $N = 1, 2, 3$ (Figure 8.1a). The cumulative-cost bin count $|\mathcal{Y}|$ scales linearly in N since per-episode cost is bounded by $N \cdot T \cdot \bar{c} = 6N$ at fixed bin resolution; on this MDP the implementation's $|\mathcal{Y}| \in \{76, 136, 196\}$ at $N = 1, 2, 3$, contributing linearly to the cell count alongside the exponential factors $|\mathcal{S}|^N \cdot |\mathcal{A}|^N$. The per-seed transition budget is fixed at $\sim 7.2 \times 10^7$ ($T = 12$ over 6×10^6 episodes), so uniform-mean per-cell visits collapse from $\sim 2.2 \times 10^3$ to ~ 54 to ~ 1.5. The $N = 1$ figure lies comfortably inside the asynchronous-SA regime; the $N = 3$ figure does not.

The uniform mean misleads on both ends. At $N = 3$, π_b visits $\sim 5.05\%$ of the joint cells over the full budget; the remaining $\sim 95\%$ remain at their initialisation throughout the run and inject their initial value into whatever upstream cell bootstraps through them. Within the visited fraction the median per-cell visit count is 4, and only $\sim 0.11\%$ of cells exceed one hundred visits. Asynchronous SA needs both a contracting per-cell update and a shrinking unvisited fraction as the budget grows; a median of four says the first does not hold, and the concentration of π_b

[1] The implementation indexes each Q-cell by a previous-action context taking values in $\mathcal{A}^N$ together with one no-previous-action sentinel at episode start, realising the augmented-cost convention of Borkar and Jain (2014) at the tabular-index level; the MMDP state–action space and the C1+C2+C3 operators are unchanged.

on policy-relevant trajectories says the second does not, since the unvisited cells are off-support rather than unsampled-yet-on-support.

Bootstrap propagation and the residual pattern. The C1 recursion at backward index t accesses $Q^{C,k}_{t+1}(z', a^\star(z'))$ at the within-sweep value at z'; a contaminated bootstrap at z' contaminates the update at (z, a). The Bellman residual on the C1 operator therefore floors at a level set by the under-visited fraction of the table, independent of per-cell update magnitudes.

This static reading explains the early phase of Figure 7.1a directly: through the first $\sim 200\,000$ episodes the maximum Bellman residual *rises* as bootstrap propagation builds out across an expanding partial table. The residual then descends through 0.5–3×10^6 episodes, briefly approaches the gate band, and re-rises through 4–6×10^6. A static-π_b argument does not, on its own, account for the late re-rise: that signature is a π_b-drift effect on a partially-filled table, deferred to Section 8.3.

An 87× budget estimate. Raising the average per-cell visit count to ~ 130 — an order-of-magnitude floor compatible with the asynchronous-SA regime, well below the $\sim 2\,200$ that closes the $N = 1$ gate — requires, on a coverage-calibrated extrapolation rather than a sample-complexity lower bound,

$$\frac{130 \cdot 4.8 \times 10^7}{T} \;\approx\; 5.2 \times 10^8 \quad \text{episodes per seed,}$$

roughly 87× the present budget. The estimate cuts both ways: the relevant denominator is smaller than $|S|^N|A|^N$ because off-support cells stay at zero visits regardless of budget, but realised visits at sampled cells must exceed the uniform mean for asynchronous-SA precision; conversely, off-support cells need not converge for the deployed greedy policy to be valid, though the gate's worst-cell statistic still registers them. Either reading puts gate clearance orders of magnitude beyond the present budget, and the order is robust: the gap is a regime mismatch between the finite-time visit distribution and the operator the residual estimates, not a 2× or 3× tuning shortfall. None of C^{eff}_α, σ_ξ, the agent coupling structure, or any object internal to C1, C2, C3 enters the floor; mechanism (1) is a property of the joint representation.

Sample-complexity reading. A complementary reading from the asynchronous-Q literature is sharper at the worst cell. Applying the standard sample-complexity scaling of Even-Dar and Mansour (2003) for $\|Q^{C,k} - Q^{C,\star}\|_\infty \le \tau_B$ to the $N = 3$ joint dimensions[2] gives $\approx 3.5 \times 10^8$ already at $d^{-1}_{\min} = 1$, or $\approx 1.9 \times 10^{12}$ on the implementation table shape used for the visit-count diagnostic — respectively $\approx 4.9\times$ and $\approx 2.7 \times 10^4\times$ the per-seed transition budget 7.2×10^7. Both readings

[2]The scaling is $\widetilde{O}\big(|\mathcal{S}^{(N)}|\cdot|\mathcal{A}^{(N)}|\cdot T^3\tau_B^{-2}d^{-1}_{\min}\big)$ with $d_{\min}$ the minimum stationary visitation probability under π_b; this is a literature-standard scaling reading rather than a tight bound for the corrected C1 operator, and $\widetilde{O}$ absorbs logarithmic factors and bounded-cost constants. On the implementation-table reading below, one factor of T is absorbed into the augmented-state cell count, so the effective horizon multiplier is T^2 rather than T^3.

place gate clearance well beyond the $\sim 87\times$ implied by the coverage extrapolation; the order of magnitude is robust.

8.2 Mechanism (2): joint-return variance, and a clip-induced blind spot in the gate

Throughout this section, $\beta_\alpha := \mathrm{VaR}_\alpha(Y^{\pi_b})$ denotes the population VaR threshold at the current behaviour policy; $\widehat{\beta}$ denotes the corresponding sample VaR estimated from $\mathcal{B}$.

The buffer-mean estimator (4.17) was C3's fix at $N = 1$: replace the Q_0^C readout, whose finite-time residual carried the amplifying factor $\beta(1-V_0/\alpha)$ (Proposition 4.2), with a sample mean over a rolling buffer of π_b-returns. The estimator has leading-order variance in the large-buffer quasi-stationary limit, $\mathrm{Var}((Y-\beta_\alpha)^+)/(\alpha^2 W)$, as derived below; $W = 8\,000$ was sized for the single-agent $\mathrm{Var}(Y)$. The centralised lift inherits C3 unchanged — W, η_k, and $\lambda_{\max} = 200$ do not scale with N. What scales is $\mathrm{Var}(Y)$.

Under the centralised lift of Section 2.2.5, the buffer-stored realised return is $Y = \sum_t C_t^{(N)}(S_t, A_t) + \xi$ with $C_t^{(N)} = \sum_i c_{0,t}$ additive across agents and $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ shared (Section 6.2). With ξ independent of the per-agent costs,

$$\mathrm{Var}(Y) \;=\; \sum_{i=1}^{N} \mathrm{Var}(C_i) \;+\; 2\sum_{i<j} \mathrm{Cov}(C_i, C_j) \;+\; \sigma_\xi^2, \tag{8.1}$$

$C_i = \sum_t c_{0,t}(S_t^i, A_t^i)$ and intercept $\sigma_\xi^2 = 0.01$. Under factored transitions and an independent-agent reference policy that factorises across agents, the cross-agent covariances vanish and the independent-agent baseline $\sum_i \mathrm{Var}(C_i) + \sigma_\xi^2$ is N-linear in the dominant term. Empirically (Figure 8.1b) the post-warm-up median $\mathrm{Var}(Y)$ rises monotonically across $N = 1, 2, 3$ and lies above the independent-agent reference at $N = 2, 3$, indicating residual $\mathrm{Cov}(C_i, C_j) > 0$ induced by the centralised π_b on top of the additive baseline.

At the large-buffer quasi-stationary limit, $\widehat{\Xi}_\alpha$ is the sample α-CVaR of i.i.d. draws of Y. The Rockafellar–Uryasev variational form is stationary in β at the population quantile $\beta_\alpha = \mathrm{VaR}_\alpha(Y)$, so by the envelope theorem the plug-in $\widehat{\beta}$ contributes to the asymptotic variance only at second order; the leading $O(W^{-1})$ term is the oracle one (Brown, 2007; Trindade et al., 2007),

$$\mathrm{Var}\big(\widehat{\Xi}_\alpha\big) \;=\; \frac{\mathrm{Var}\big((Y-\beta_\alpha)^+\big)}{\alpha^2\, W} \;+\; o\big(W^{-1}\big)\,. \tag{8.2}$$

The N-dependence in (8.2) is localised on a single factor. The denominator $\alpha^2 W$ is held fixed across the N-sweep — α is fixed by the constraint specification and W is fixed by the centralised stress-test protocol of Section 6.6, neither carrying any explicit N-dependence. The dependence enters only through the upper-tail variance of Y: $\mathrm{Var}((Y-\beta_\alpha)^+)$ scales as $\mathrm{Var}(Y)$ up to an α-dependent constant in

the CLT regime, hence N-linear in the dominant term under the independent-agent baseline of (8.1); under positive policy-induced covariance the leading constant is larger. With W fixed, this is the gradient noise the dual update consumes.

Saturation as a finite-time consequence. The dual update

$$\lambda_{k+1} \;=\; \mathrm{clip}\big[\,\lambda_k + \eta_k\big(\widehat{\Xi}_\alpha - C_\alpha^{\mathrm{eff}}\big),\; 0,\; \lambda_{\max}\big] \tag{8.3}$$

is a clipped random walk with mean drift $\mu(\lambda) := \Xi_\alpha^{\pi_b(\lambda)} - C_\alpha^{\mathrm{eff}}$ and per-update diffusion of scale $\eta_k \sigma_{\widehat{\Xi}}$, where $\sigma_{\widehat{\Xi}} = \sqrt{\mathrm{Var}((Y-\beta_\alpha)^+)/(\alpha^2 W)}$ in the independent-agent baseline of (8.2). The conclusion is that $\sigma_{\widehat{\Xi}}$ grows in N while μ on the per-capita budget convention does not, so the saturation threshold in signal-to-noise units lowers with N even though the constraint loosens on the standardised scale.

A finite-time stochastic-approximation reading (cf. Borkar (2008, §6.1) for the underlying SA noise–drift balance) supplies the qualitative picture in two stages. *Reach:* a sustained positive drift $\mu > 0$ accumulates on the slow λ-cadence and drives the iterate to the upper clip $\lambda_{\max}$ in finite time. *Detach:* once at the boundary, release requires a stretch of windows in which the empirical drift turns negative; when the per-update signal-to-noise ratio $\mathrm{SNR} := \mu/\sigma_{\widehat{\Xi}}$ is approximately 1 — i.e. the average per-update drift in λ is comparable in magnitude to the within-window dispersion of the buffer gradient — such windows are routinely overrun by within-window dispersion, and the iterate fails to leave the boundary on the finite slow timescale. This is a heuristic noise–drift diagnostic on the slow update rather than a hitting-probability bound: distance to the boundary, the step-size schedule, the sign and amplitude of the drift, the noise tail, and the time horizon would all enter a quantitative hitting statement and are not pinned by the present protocol.

The N-dependence enters $\sigma_{\widehat{\Xi}}$ through $\mathrm{Var}((Y-\beta_\alpha)^+) \propto \mathrm{Var}(Y)$, N-linear in the dominant term under the independent-agent baseline of (8.1), so $\sigma_{\widehat{\Xi}} \propto \sqrt{N/W}$ at fixed W. The numerator μ at the closest-approach phase is the residual infeasibility recorded above ($\mu \lesssim 0.06$ at $N = 3$); on the per-capita convention $C_\alpha^{\mathrm{eff}} = N\,C_{\mathrm{base}}$ it carries no N-amplification, and the standardised-margin loosening of Section 7.6 works in its favour. The saturation threshold $\mu/\sigma_{\widehat{\Xi}} \sim 1$ is therefore denominator-driven across the sweep. At $N = 3$ with $W = 8\,000$, observed $\sigma_{\widehat{\Xi}} \approx 0.03$ on the path-window dispersion of the dual signal in the closest-approach phase, and the residual gap $\mu \lesssim 0.06$ in that regime before final ascent, give $\mathrm{SNR} \approx 2$ (per-seed median): on the slow λ-cadence, the per-update positive drift is of the same order as the within-window dispersion of the buffer gradient, and the iterate saturates at $\lambda_{\max}$ under a still-positive average gradient (Figure 7.1a, middle).

Once $\bar{\lambda} = \lambda_{\max}$, the clip reflects every positive proposal: release requires a downward window — a stretch in which the buffer mean under-estimates $\Xi_\alpha^{\pi_b}$ enough for the gradient to turn strongly negative. The slack-driven recoveries documented in Section 7.4 are this mechanism in action: each downward kick is overrun by the next high-variance window, since at the operating (W, N) windows of either sign occur on the slow λ-cadence. The held-out $\Xi_\alpha^{\pi_g}$ near the feasibility band is a separate signal on the deployment signal. Under the protocol it is consistent with two sources, not

mutually exclusive, for the buffer-versus-held-out gap (Section 7.4), neither of which the gate's $\bar{\lambda}$-drift criterion can detect at the clip (*first signature* below).

First signature: the clip blinds the λ-drift gate. Section 7.4 records that λ-drift / τ_λ peaks at ≈ 262 through the cascaded warm-up and then collapses to near zero for the remainder of the run. The collapse is mechanism (2)'s fingerprint, not its absence. With λ-drift $:= |\lambda_{k+1} - \lambda_k|/\tau_\lambda$, the clip in (8.3) pins $\lambda_{k+1} = \lambda_k = \lambda_{\max}$ whenever the proposed increment is positive and the current iterate already lies at the upper clip. At the boundary the drift is identically zero *by construction*, irrespective of the gradient that drives the unclipped proposal of Section 7.4, which remains positive throughout the saturation window. The five-component KKT-based operational gate of Section 6.6 treats λ-drift as a stationarity proxy for the dual update; at the upper clip the proxy is measure-blind, returning drift ≈ 0 regardless of the true gradient.

Second signature: the dual–eval feasibility gap. Section 7.4 also records the dual–eval asymmetry on the feasibility residual: $(\widehat{\Xi}_\alpha - C_\alpha^{\text{eff}})/\tau_F \approx 3.19$ on the buffer of π_b-returns that drives the slow update, against ≈ 0.81 on a held-out evaluation of the deployed greedy policy. The held-out $\Xi_\alpha^{\pi_g}$ lying inside the τ_F-band is a separate measurement on the deployment signal; it records that the greedy policy reaches the feasibility neighbourhood, while the signal the slow update consumes does not. The relative gap of ≈ 0.12 between the two signals is consistent with two sources — product-form behaviour–greedy gap at $N = 3$ (Section 7.4) and the π_b-drift bias bounded by (8.4) — which the protocol does not separate. The gap localises mechanism (2) on the buffer estimator: what the slow update consumes is the buffer mean over π_b-returns, carrying the π_b-drift bias of Section 8.3 on top of its $1/W$ variance, while the held-out evaluation operates on π_g directly and is free of both. λ-drift cannot detect this gap by construction: the clip has removed the variable.

8.3 The two mechanisms couple through π_b

Lemma 8.1 (Asymptotic algorithm-level separation of the two mechanisms)**.** *Under the three-timescale separation $a_k \gg \gamma_k \gg \eta_k$ of Borkar (2008) and quasi-stationarity of π_b on the slow λ-timescale, the two finite-time quantities driving the mechanisms of Sections 8.1–8.2 are functionals of the trajectory through distinct projections, each entering the algorithm through a different component:*

1. *the per-cell visit count $\nu_k(z, a)$ on the joint cost Q-function, indexing mechanism (1), is a functional of the marginal (z, a)-occupancy of π_b and acts on the Q-table updates;*
2. *the buffer-tail variance $\sigma^2_{\widehat{\Xi}}$ of (8.2), indexing mechanism (2), is a functional of the empirical return distribution under π_b and acts on the slow dual update.*

At the asymptotic fixed point the visit-count functional enters through the Q-table update while the buffer-tail variance enters through the slow dual update: the two

are separated by the algorithmic channel through which they act, although both are induced by the same trajectory distribution.

Proof. $\nu_k(z,a) = \sum_{k' \le k} \mathbb{1}\{Z_t^{k'} = z, A_t^{k'} = a \text{ for some } t\}$ depends on the realised joint trajectory through the (z,a)-projection only; $\sigma^2_{\widehat{\Xi}}$ depends on the empirical return distribution through the $Y^k = \sum_t C_t^{(N)}(S_t^k, A_t^k) + \xi^k$ projection only. Both projections are deterministic functions of the same trajectory and are not statistically independent; the separation is at the algorithmic level. At the asymptotic limit (Borkar, 2008, §§6.1–6.3), π_b is quasi-stationary on the slow scale, the fast-timescale recursion reads the trajectory only through the visit-count projection, and the slow-timescale recursion reads it only through the return-distribution projection. The two mechanisms therefore touch the algorithm at distinct levels — the Q-function table and the slow dual update. □

Lemma 8.1 is asymptotic: in finite time the two mechanisms couple through π_b itself, which the rest of this section examines.

In finite time, they meet at π_b. The behaviour policy is ϵ-greedy on the Lagrangian Q-function $J = Q^R - \lambda Q^C$ (Algorithm 4.1). When Q^C is far from its fixed point on the joint table — the situation mechanism (1) creates at $N = 3$ — the $\arg\max J$ direction shifts as under-visited cells fill in and as their bootstrapped neighbours update, and π_b drifts with it. The buffer $\mathcal{B}$ then samples a non-stationary distribution; the Y_i stored at episode i are draws from $\pi_b^{(i)}$, with no single $\pi_b^\star$ governing the window. The $1/W$-variance regime of (8.2) assumed i.i.d. draws from a fixed law, so under π_b-drift the buffer mixes a moving target and the dual gradient $\widehat{\Xi}_\alpha - C_\alpha^{\text{eff}}$ is no longer an unbiased estimate of $\Xi_\alpha^{\pi_b} - C_\alpha^{\text{eff}}$ for any fixed π_b.

A total-variation bound on the per-window bias. At fixed β, assuming a uniform bound $\|Y\|_\infty < \infty$ on the buffer return, the per-window bias of the buffer mean of the R–U surrogate $\beta + \alpha^{-1}\mathbb{E}[(Y-\beta)^+]$ admits the standard total-variation bound

$$\begin{aligned} &\Big|\, \mathbb{E}\Big[\widehat{\Xi}_\alpha(\beta) \,\Big|\, \mathcal{B}\Big] - \Xi_\alpha(\beta; Y^{\pi_b^{(k)}}) \,\Big| \\ &\qquad \le \frac{T\,\|Y\|_\infty}{\alpha} \sup_{i,j \in [k-W+1,\,k]} \mathrm{TV}_{\text{step}}\big(\pi_b^{(i)}, \pi_b^{(j)}\big), \end{aligned} \tag{8.4}$$

where $\mathrm{TV}_{\text{step}}(\pi_1,\pi_2) := \sup_z \frac{1}{2}\sum_a |\pi_1(a \mid z) - \pi_2(a \mid z)|$ is the worst-state per-step total variation. This follows from $(Y-\beta)^+ \in [0, \|Y\|_\infty]$, the variational definition of total variation, and the simulation lemma $\mathrm{TV}(P_T^{\pi_1}, P_T^{\pi_2}) \le T \cdot \mathrm{TV}_{\text{step}}(\pi_1,\pi_2)$ on the path measure of the finite-horizon MDP: for any two stationary policies π_1, π_2, $|\mathbb{E}^{\pi_1}[(Y-\beta)^+] - \mathbb{E}^{\pi_2}[(Y-\beta)^+]| \le \|Y\|_\infty \,\mathrm{TV}(P_T^{\pi_1}, P_T^{\pi_2}) \le T\|Y\|_\infty \,\mathrm{TV}_{\text{step}}(\pi_1,\pi_2)$; the window mean is a uniform mixture of episodes generated under $\{\pi_b^{(i)}\}_{i=k-W+1}^{k}$, so its bias against the current $\Xi_\alpha^{\pi_b^{(k)}}$ inherits the worst pairwise step-level TV in the window through the constant $T\|Y\|_\infty/\alpha$. The uniform-bound assumption $\|Y\|_\infty < \infty$ holds exactly under the bounded smoothing alternatives of Section 6.2 ($\xi \sim U[-\delta,\delta]$, truncated Gaussian); under the Gaussian shock of the experiments ξ has exponentially light tails, so the bound applies as a high-probability diagnostic after truncating ξ

at a level whose tail mass is exponentially small, with $\|Y\|_\infty$ replaced by $T\bar{c}$ plus the truncation level. Extending (8.4) to the plug-in CVaR with $\beta = \beta_\alpha(Y^{\pi_b})$ requires the usual quantile regularity on $\beta \mapsto \mathbb{P}(Y > \beta)$, which is inherited locally from the shock smoothness of Section 6.2; we use the bound as a finite-time diagnostic of how π_b-drift converts the $1/W$-variance regime of (8.2) into a biased buffer mean. The standard quasi-stationarity argument (Borkar, 2008, §§6.2–6.3) closes only when the supremum in (8.4) vanishes on the slow η_k-timescale; mechanism (1)'s sustained reshuffling of $\arg\max J$ on a partially-filled table is precisely the mechanism by which it does not.

The non-monotone Bellman residual recorded in Section 7.4 maps back into this picture cleanly. The early peak through 200 000–500 000 episodes is mechanism (1) building out across an expanding partial table; the descent through $1–3\times10^6$ episodes is the same mechanism contracting on cells π_b visits at the operating point reached by then; the re-rise through $4–6 \times 10^6$ episodes is what $\arg\max J$ reshuffling does once the dual saturates at $\lambda_{\max}$ and ongoing C2 updates on Q^R shift the Lagrangian's maximiser across cells whose C1 estimate was already fragile.

The two mechanisms couple in a self-reinforcing cycle at $N = 3$, which we read as a four-step loop:

1. an under-converged Q^C on the joint table reshuffles $\arg\max J$, so the executed π_b shifts cell by cell;
2. the shifted π_b writes a non-stationary buffer, so the buffer-mean gradient $\widehat{\Xi}_\alpha - C_\alpha^{\text{eff}}$ drifts off the population zero;
3. the drifting gradient sends $\bar{\lambda}$ to $\lambda_{\max}$;
4. the saturated λ freezes the selector at exactly the cells whose Q^C estimate was most fragile, returning the system to step 1.

Two of the three conditions of Borkar and Jain (2014, Theorem 2) are thereby strained: the asynchronous-Q regime (i) by mechanism (1) and quasi-stationarity of π_b on the slow λ-timescale (ii) by the loop above; only quantile regularity (iii), supplied at every N by the shared shock ξ, survives. Each mechanism is capable of blocking gate clearance in the present diagnostic — mechanism (1) by holding the Bellman residual above tolerance, mechanism (2) by saturating the dual — and the π_b signal forbids treating them sequentially: the same Q^C that has not converged is what controls π_b and the buffer's sampling distribution.

8.4 Hyperparameter adjustment at fixed representation

Before interpreting the bottleneck as structural, we tested each hyperparameter in turn — step-size exponents inside the band of Borkar (2008), ϵ across an order of magnitude, buffer sizes $W \in \{4\,000, 8\,000, 16\,000\}$, dual clips $\lambda_{\max} \in \{50, 200, 500\}$. None of the resulting configurations cleared the gate at $N{=}3$ under the per-seed budget, and the failure pattern was the same on each: the cluster on the $\lambda_{\max}$ line

in Figure 7.1c moved up or down on the boundary but did not detach from it, and the Bellman residual on the joint cost Q-function stayed above tolerance on the cells π_b visits. The pattern is consistent with the structural reading below: each knob can relieve one symptom of the diagnostic, but none of them changes the joint table volume or the joint-return variance at fixed centralised representation, the two structural quantities of Sections 8.1–8.2 that bound the relief available.

Q-function-side knobs (a_k, ϵ, seeds). These touch the visit *distribution* but not the total visit *count*, fixed by the per-seed budget at $\sim 7.2\times10^7$ transitions. Step-size adjustments rescale per-update corrections at cells π_b visits; they do not manufacture samples for cells π_b does not. The under-visited fraction sets the Bellman-residual floor of mechanism (1), which is on this reading insensitive to a_k. Raising ϵ flattens the visit distribution rather than its total mass, trading precision on policy-relevant cells for coverage of off-support ones; at large ϵ, the executed π_b deviates far enough from $\arg\max J$ that the policy-evaluation reading of Q^C at π_b stops being informative about the deployed greedy policy. Across-seed averaging tightens confidence intervals on within-cell statistics but does not shift the per-seed trajectory of any KKT residual.

Dual-side knobs ($\lambda_{\max}$, W). Raising $\lambda_{\max}$ delays saturation without changing the SNR of the dual gradient: in the noise–drift regime of Section 8.2, where the per-update positive drift is of the same order as the within-window dispersion of the buffer gradient, a random walk on the slow λ-cadence can saturate at whatever upper boundary is set on the finite slow timescale. Raising W reduces $\mathrm{Var}(\widehat{\Xi}_\alpha)$ at the $O(1/W)$ rate and is the one knob that targets mechanism (2) on its variance axis; its cost is exactly the bias of (8.4), since W is also the window over which π_b must be approximately stationary. A finite optimal $W^\star(N)$ trades the two; on the present table volume at $N = 3$ it does not, empirically, release the dual from the upper boundary, though this is an empirical reading rather than a proof.

8.5 What the per-agent decomposition replaces

The two structural quantities of Table 8.1 are properties of the joint representation, not of the algorithm that runs over it. Neither is touched by C1, C2, or C3, and hyperparameter adjustment at fixed representation (Section 8.4) does not relocate either of them. The forward direction motivated by this attribution is to replace these two centralised joint quantities by per-agent counterparts along Tasche (1999)'s identity for risk contributions, in which the joint CVaR is written as a sum of per-agent contributions on a shared joint-tail event. Chapter 9 develops the construction and its consequences for both mechanisms in detail; Table 8.1 records the targeted replacements alongside the structural quantities they address.

Table 8.1. The two structural quantities of the $N = 3$ centralised stall. Both quantities are properties of the joint representation, neither is touched by C1, C2, or C3, and neither admits a local-tuning fix at fixed N. The per-agent decomposition of Gautier et al. (2023b) replaces the exponentially scaling joint cost Q-function table by per-agent local Q-functions and the single joint dual signal by per-agent risk-contribution signals; joint-tail dependence is relocated rather than eliminated, surviving through the shared VaR_α event and the common joint-tail index set that conditions every per-agent contribution estimator (Chapter 9, *Relocation, not elimination*). Detailed development in Chapter 9.

	Mechanism (1) Table-volume dilution	**Mechanism (2)** Joint-return variance
Structural quantity	$\lvert S\rvert^N \cdot \lvert A\rvert^N$	$\sum_i \mathrm{Var}(C_i) + 2\sum_{i<j} \mathrm{Cov}(C_i, C_j) + \sigma_\xi^2$ (independent-agent baseline: $\sum_i \mathrm{Var}(C_i) + \sigma_\xi^2$)
Sets	per-cell Bellman precision of Q^C on the joint table	variance of $\widehat{\Xi}_\alpha$ on a fixed-W buffer
Scaling in N	exponential	N-linear under the independent-agent baseline; larger under positive policy-induced covariance; σ_ξ^2 intercept
Ch. 7 signature	Bellman residual non-monotone, no gate-band window	$\bar{\lambda} = \lambda_{\max}$ on every seed; dual–eval feasibility gap; λ-drift uninformative at clip
Driver in Algorithm 4.1	C1 recursion under asynchronous SA (Tsitsiklis, 1994)	C3 buffer-mean estimator on π_b-returns
Tuning that does not help	a_k, ϵ, seeds	$\lambda_{\max}$, seeds; W trades variance for drift bias
What RCA replaces it with	N per-agent augmented tables of size $T \cdot \lvert\mathcal{S}_0\rvert \cdot \lvert\mathcal{Y}_i\rvert \cdot \lvert\mathcal{A}_0\rvert$, linear in N rather than exponential	N parallel per-agent risk-contribution signals; per-agent estimator variance tied to local tail variation rather than the full joint-return tail

8.6 Limitations of this attribution

Our reading is empirical and tabular, anchored on five seeds at $N \in \{1, 2, 3\}$ on a single MMDP. Two boundaries specific to this chapter, beyond the empirical scope already named in Section 7.6, should be added.

Constraint tightness compounds with mechanism (2). Section 7.6 records that, under $C_\alpha^{\mathrm{eff}} = N \cdot C_{\mathrm{base}}$, the standardised margin $(C_\alpha^{\mathrm{eff}} - \mathbb{E}[Y])/\sqrt{\mathrm{Var}(Y)}$ widens by $\sqrt{N}$ in the positive-margin regime $C_{\mathrm{base}} > \mathbb{E}[c_i]$. The $N = 3$ stall is therefore not explained by an induced tightening of the constraint; if anything, mechanism (2)'s finite-time severity is recorded under a budget convention that, on the standardised scale, works in the algorithm's favour. The chapter does not separate the variance-driven contribution to mechanism (2) from any operating-point dependence of $|\Xi_\alpha^{\pi_b} - C_\alpha^{\mathrm{eff}}|$ on N. A tail-relative budget convention that holds the standardised margin fixed across N would isolate the two; this robustness sweep is left to Chapter 9.

Weak coupling assumed throughout. The dominant scaling argument of mechanism (2) rests on $\sum_i \mathrm{Var}(C_i)$ as the independent-agent baseline of the covariance-aware decomposition (8.1). The Taxi MMDP of Chapter 6 satisfies the factored-transition, additive-cost, no-shared-resource conditions of weak coupling by construction; the empirical excess of the medians over the independent-agent reference in Figure 8.1b indicates residual $\mathrm{Cov}(C_i, C_j) > 0$ induced by the centralised π_b itself — a finite-time policy-coupling effect; the dynamics themselves remain decoupled by

construction. Settings outside weak coupling, with shared resources or correlated per-agent randomness, can introduce larger off-diagonal covariances in $\mathrm{Var}(Y)$, in which case mechanism (2) follows a different scaling. The present attribution applies inside weakly coupled MMDPs and inherits the scope of De Nijs et al. (2021)'s taxonomy.

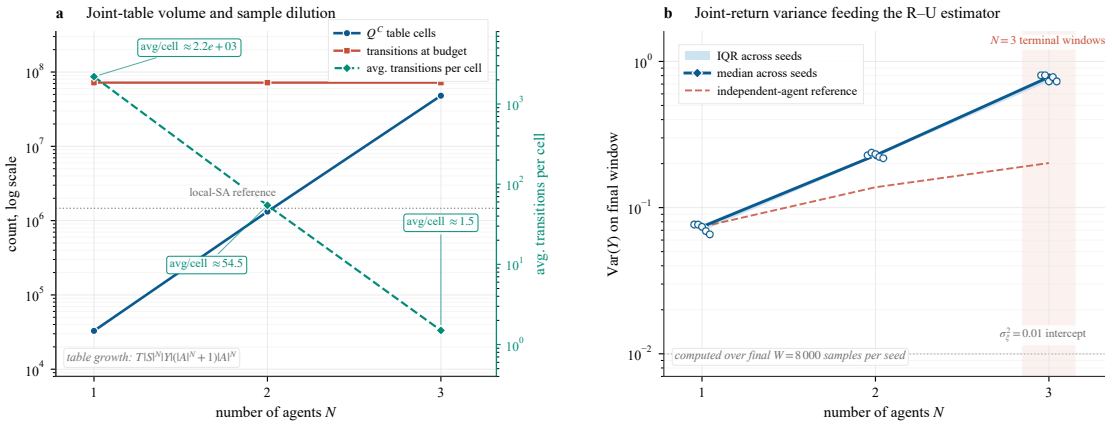

Figure 8.1. The two structural quantities of the centralised stall on a shared $N \in \{1, 2, 3\}$ axis. **(a)** Mechanism (1): joint Q^C-table cell count $T \cdot |S|^N \cdot |Y| \cdot (|A|^N + 1) \cdot |A|^N$ (left axis, log) against the fixed per-seed transition budget of $\sim 7.2 \times 10^7$, with average transitions-per-cell on the right axis (log). Dotted horizontal: an order-of-magnitude floor on per-cell visits compatible with the asynchronous-Q regime of Tsitsiklis (1994). **(b)** Mechanism (2): $\mathrm{Var}(Y)$ on the post-warm-up $W = 8\,000$-sample window, five seeds per cell. Per-seed dots open, five-seed median diamond solid, IQR shaded; dashed reference is the independent-agent baseline $\sum_i \mathrm{Var}(C_i) + \sigma_\xi^2$ of (8.1) with $\sigma_\xi^2 = 0.01$ as intercept.

9

Toward Risk-Contribution Allocation

Chapter 8 closed on a structural diagnosis. The centralised stall at $N = 3$ rests on two quantities of the joint representation — the joint Q^C-table volume $|S|^N \cdot |A|^N$, and the joint return variance $\mathrm{Var}(Y)$ driving the buffer-mean dual signal, N-linear under the independent-agent baseline $\sum_i \mathrm{Var}(C_i) + \sigma_\xi^2$ and larger under positive policy-induced covariances. Neither is touched by C1, C2, or C3, nor by any hyperparameter the construction exposes. The forward direction is therefore not a sharper algorithm on the joint table but a representation that exposes per-agent quantities to the same single-agent machinery Chapter 4 validated. Tasche (1999)'s identity is the algebraic ground; Gautier et al. (2023b)'s offline construction is the closest precedent.

We sketch the construction, isolate the analytical and empirical questions it opens, and stop short of the implementation that would close them.

9.1 Tasche's identity as the bridge

The identity of Tasche (1999),

$$\begin{aligned} \mathrm{CVaR}_\alpha\Big(\textstyle\sum_{i=1}^N C_i\Big) \;&=\; \sum_{i=1}^N \mathrm{RC}_\alpha(C_i), \\ \mathrm{RC}_\alpha(C_i) \;&:=\; \mathbb{E}\Big[\, C_i \;\Big|\; \textstyle\sum_j C_j > \mathrm{VaR}_\alpha\big(\sum_j C_j\big)\Big], \end{aligned} \tag{9.1}$$

holds under integrability of the C_i together with continuity of the CDF of $Y = \sum_j C_j$ at its VaR_α threshold; without the latter, the standard boundary correction of Rockafellar and Uryasev (2000) applies and the per-agent allocation is not pinned at a single vector. The episode-end smoothing shock $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ of Section 6.2, introduced for the centralised diagnostic, supplies continuity for $Y + \xi$. The identity is algebraic, not dynamical: it does not require independence of the C_i across i, nor weak coupling of the underlying MMDP.

The role this plays in the forward design is structural. The single dual update of Algorithm 4.1 against $\widehat{\Xi}_\alpha - C_\alpha^{\mathrm{eff}}$ becomes N parallel updates against $\widehat{\mathrm{RC}}_\alpha^{(i)} - C_\alpha^{(i)}$ with

$\sum_i C_\alpha^{(i)} = C_\alpha^{\text{eff}}$. The conditioning event in each per-agent estimator is the shared joint-tail event $\{Y > \text{VaR}_\alpha(Y)\}$; the conditional expectations themselves are per-agent. The one joint object that survives the per-agent split is therefore a scalar quantile estimator $\widehat{\beta}$ on a joint-return buffer, refreshed once per dual update, which is independent of the joint table volume of mechanism (1). The structural reduction is from a $|S|^N \cdot |A|^N$ joint object and a scalar dual signal driven by $\text{Var}(Y)$ to N per-agent augmented Q-functions on local state–action–cost spaces $T \cdot |S_i| \cdot |Y_i| \cdot |A_i|$, N buffers of size W, and one shared scalar $\widehat{\beta}$ on the joint return. The per-agent augmented dimension $|Y_i|$ is local: it scales with the per-agent cost upper bound $T \cdot \bar{c}$, not with the joint $N \cdot T \cdot \bar{c}$, so the exponential dependence in N is removed from both the state factor $|S_i|$ and the action factor $|A_i|$ while the augmented-cost factor stays at the single-agent scale.

Relocation, not elimination. Cross-agent dependence does not disappear: it is relocated. The shared conditioning event $\{Y > \widehat{\beta}\}$ ties the N contribution estimators to the same random subset of buffer indices, so $\widehat{\beta}$-noise propagates into every $\widehat{\text{RC}}_\alpha^{(i)}$ in tandem rather than independently; $\widehat{\beta}$ itself is a sample quantile of $Y = \sum_i C_i$, whose finite-window variance inherits the upper-tail shape of the joint return, including any positive $\text{Cov}(C_i, C_j)$ induced by the executed product-form π_b. The architecture through which this dependence is consumed changes. Formally: in the centralised system, Algorithm 4.1 drives a single scalar dual variable $\lambda \in \mathbb{R}_{\geq 0}$ from one buffer-mean estimator $\widehat{\Xi}_\alpha$ computed on the joint-return buffer $\mathcal{B} \subset \mathbb{R}$, so the slow update consumes a single scalar projection of all N agents' contributions. In the per-agent system, the dual is a vector $(\lambda^{(1)}, \ldots, \lambda^{(N)}) \in \mathbb{R}^N_{\geq 0}$, each $\lambda^{(i)}$ driven by its own estimator $\widehat{\text{RC}}_\alpha^{(i)}$ on the per-agent buffer $\mathcal{B}_i \subset \mathbb{R}$ and conditioned on the shared scalar $\widehat{\beta} = \text{Quantile}_{1-\alpha}(\mathcal{B}_{\text{joint}})$. The N slow-side recursions no longer share a Bellman recursion with each per-agent Q-function table; the only joint object both sides consume is $\widehat{\beta}$. The structural reduction is from one exponentially scaling joint table $T \cdot |S|^N \cdot |Y| \cdot |A|^N$ and one scalar dual signal driven by $\text{Var}(Y)$, to N per-agent augmented tables of total dimension $\sum_i T \cdot |S_i| \cdot |Y_i| \cdot |A_i|$ linear in N and N parallel per-agent dual signals, with a single shared scalar $\widehat{\beta}$ as the only surviving joint object.

9.2 Algorithmic sketch

Status of this construction. The construction below is an *algorithmic sketch*: a forward direction motivated by the diagnosis of Chapter 8, not a thesis contribution. It is specified only at the level required to frame the analytical questions of Section 9.3 and the empirical agenda of Section 9.4; implementation, ablation, and empirical validation are deferred to follow-up work. We commit to no per-agent V_i schedule, no $\widehat{\beta}$ refresh cadence, and no budget-allocation rule beyond the symmetric default; each is an open item of Section 9.3, and a formal pseudocode specifying any of them would prejudge the corresponding question. The four-contribution scope

of Section 1.4 (C1, C2, C3, C4) is unchanged by this chapter: Chapter 9 is the "forward direction" arrow that Figure 1.1 marks as dashed.

Each agent $i \in [N]$ runs the Q-function-side machinery of Algorithm 4.1 on its local state–action space: per-agent reward and cost Q-functions Q_i^R, Q_i^C on $|S_i|\cdot|A_i|$ tables, updated by the Bellman-consistent recursion (4.5), with shared per-agent selector $a_i^\star = \arg\max_{a_i}[Q_i^R - \lambda^{(i)} Q_i^C]$ on the within-sweep next-step values. The executed joint policy π_b is the product of per-agent ϵ-greedy actions on $J_i = Q_i^R - \lambda^{(i)} Q_i^C$, identical in form to the per-agent variant of Section 6.3. Greedy invariance (Proposition 4.1) carries over per-agent at any fixed $\lambda^{(i)} \geq 0$.

The dual-estimator role of C3 is what changes. A joint-return buffer $\mathcal{B}_{\text{joint}}$ of capacity W holds the trajectory sums $Y^k = \sum_i Y_i^k$ on the most recent W episodes; from it,

$$\widehat{\beta} \;=\; \text{Quantile}_{1-\alpha}(\mathcal{B}_{\text{joint}}) \tag{9.2}$$

is evaluated at each dual-cadence step. Per-agent buffers $\mathcal{B}_i$ hold Y_i^k on the same index set, sharing indices with $\mathcal{B}_{\text{joint}}$ but not contents. The per-agent risk-contribution estimator is the empirical mean of Y_i^k over the subset of indices for which $Y^k > \widehat{\beta}$,

$$\widehat{\text{RC}}_\alpha^{(i)} \;=\; \frac{1}{|\mathcal{T}|} \sum_{k \in \mathcal{T}} Y_i^k, \qquad \mathcal{T} \;=\; \{k \in [1, W] : Y^k > \widehat{\beta}\}, \tag{9.3}$$

with $|\mathcal{T}| \approx \alpha W$ at the saddle. The per-agent dual updates are

$$\lambda^{(i)} \;\leftarrow\; \big[\, \lambda^{(i)} + \eta_k \big(\widehat{\text{RC}}_\alpha^{(i)} - C_\alpha^{(i)}\big) \,\big]_+, \qquad i = 1, \ldots, N, \tag{9.4}$$

running in parallel on the slow timescale, with per-agent budgets $\sum_i C_\alpha^{(i)} = C_\alpha^{\text{eff}}$.

Operational outline. Per episode, the N per-agent learners apply the C1+C2 backwards Bellman update of Algorithm 4.1 to (Q_i^R, Q_i^C) on $\mathcal{M}_i$ with selector $a_i^\star = \arg\max_{a_i}[Q_i^R - \lambda^{(i)} Q_i^C]$, and append the joint and per-agent returns (Y^k, Y_i^k) at a shared episode index to $\mathcal{B}_{\text{joint}}$ and $\{\mathcal{B}_i\}$. On the slow timescale, $\mathcal{B}_{\text{joint}}$ supplies $\widehat{\beta}$ (9.2), the per-agent buffers supply $\{\widehat{\text{RC}}_\alpha^{(i)}\}$ (9.3), and the N projected dual ascents (9.4) run in parallel. The per-agent V_i recursion under the shared $\widehat{\beta}$, the schedule on which $\widehat{\beta}$ refreshes relative to the slow updates, and the question of whether an adaptive $\{C_\alpha^{(i)}\}$ requires a fourth timescale are the open analytical items of Section 9.3; a formal pseudocode would prejudge each. The C1+C2 side reuses Algorithm 4.1 per agent without modification; the C3 side is fully specified by (9.2)–(9.4) under any schedule consistent with Section 9.3.

Three structural properties of the construction match the bottleneck reading of Chapter 8. (i) The per-agent Q-function-side updates run on local state–action spaces of total dimension $\sum_i T\cdot|\mathcal{S}_i|\cdot|\mathcal{Y}_i|\cdot|\mathcal{A}_i|$, linear in N rather than the exponential $T\cdot|\mathcal{S}|^N\cdot|\mathcal{Y}|\cdot|\mathcal{A}|^N$ of mechanism (1). (ii) The single scalar buffer-mean signal driving one λ in Algorithm 4.1 is replaced by the vector $\{\widehat{\text{RC}}_\alpha^{(i)}\}_{i=1}^N$ driving $\{\lambda^{(i)}\}_{i=1}^N$, each anchored on a local buffer $\mathcal{B}_i$ tied to per-agent cost variation rather than to the

full joint return. (iii) The single object surviving the per-agent split is the scalar quantile $\widehat{\beta}$ on $\mathcal{B}_{\text{joint}}$, refreshed once per dual cadence; this is the shared dependence flagged as retained rather than removed (*Relocation, not elimination*, Section 9.1). Figure 9.1 contrasts the resulting per-agent architecture with the centralised one of Chapter 8.

Population-level consistency. The construction relies on $\widehat{\mathrm{RC}}_\alpha^{(i)}$ being consistent for $\mathrm{RC}_\alpha(C_i)$ at the large-buffer quasi-stationary limit. Under quasi-stationarity of π_b on the buffer's mixing scale (Borkar, 2008, §§6.2–6.3) — treated here as effective stationarity of the return distribution on the window of length W, so that buffered returns may be regarded as approximately i.i.d. draws from a single law on this timescale — continuity of F_Y at $\mathrm{VaR}_\alpha(Y)$ (supplied by the smoothing shock ξ on Y), and integrability of the per-agent returns, sample-quantile consistency (Serfling, 1980, Ch. 2) gives $\widehat{\beta} \to \mathrm{VaR}_\alpha(Y)$ almost surely as $W \to \infty$ on the buffer-mixing scale. A standard plug-in argument under continuity of $\beta \mapsto \mathbb{E}[C_i \mid Y > \beta]$ at $\mathrm{VaR}_\alpha(Y)$, inherited from the same shock-induced density continuity, then gives $\widehat{\mathrm{RC}}_\alpha^{(i)} \to \mathrm{RC}_\alpha(C_i)$ almost surely for each i, and summing over i recovers $\sum_i \widehat{\mathrm{RC}}_\alpha^{(i)} \to \mathrm{CVaR}_\alpha(Y)$ by (9.1): the N per-agent dual gradients are population-level consistent with the same joint constraint $\sum_i \mathrm{RC}_\alpha(C_i) \le C_\alpha^{\text{eff}}$ that Algorithm 4.1 enforces. The statement is sample-mean consistency on a fixed-distribution buffer; it does not establish that the limiting projected dual equilibrium of the per-agent parallel updates (9.4) coincides with the centralised limiting projected dual equilibrium under the moving π_b that the parallel duals themselves induce, nor does it provide a non-asymptotic statement under genuinely non-stationary π_b requiring an explicit mixing-rate condition on the policy drift. Both are open three-timescale questions of Section 9.3.

Distance from offline RCA. The online construction replaces Gautier et al. (2023b)'s iterative-sequential planning — worst-trade-off agent, lowered risk-contribution target, offline replan, Monte Carlo joint risk — by parallel-asynchronous learning: all N dual variables move on the common dual cadence and per-agent learners run continuously, with no agent held fixed while another adjusts. It inherits the empirical-convergence gap of Gautier et al. (2023b) and adds the three-timescale machinery that Section 9.3 catalogues. That machinery is the price of the fully online endpoint of a graded family. At the conservative end, the iterative-sequential outer loop is retained and only the per-agent inner problem is made online. This already secures the structural relocation described in Section 9.1 while avoiding the full three-timescale coupling required by that endpoint. The conservative member is therefore the natural point of departure; the fully online scheme is the principled limiting construction.

Per-agent budget allocation. The identity (9.1) holds at any allocation $\{C_\alpha^{(i)}\}$ summing to C_α^{eff} at the saddle; it does not select one. The symmetric Taxi MMDP of Chapter 6 admits the uniform allocation $C_\alpha^{(i)} = C_\alpha^{\text{eff}}/N$ as the symmetric default, which is the choice used in Section 9.4. In asymmetric domains this requires recon-

sideration; an adaptive allocation tracking realised contributions introduces a fourth timescale and is left to Section 9.3.

9.3 Open analytical questions

The asymptotic guarantee of Borkar and Jain (2014, Theorem 2) on Algorithm 4.1 rests on three-timescale separation between (π_b, V_t, Q_t) on the fast scale, β on the intermediate scale, and λ on the slow scale, plus quasi-stationarity of π_b on the slow scale. The per-agent split breaks two distinct elements of the argument; a third question, on budget allocation, follows below.

The joint VaR object. In Algorithm 4.1, β is updated by a per-episode SA recursion driven by $\alpha - V_0(z_0)$, with V_0 the centralised tail-probability estimate produced by the same backwards Bellman update that produces Q^C. In online RCA, $\widehat{\beta}$ of (9.2) is a sample quantile on a rolling joint-return buffer. The two are different SA objects — the centralised β tracks the recursive Bellman value V_0, the RCA $\widehat{\beta}$ tracks the empirical CDF of Y on a fixed-window buffer — targeting the same limiting quantity under quasi-stationarity of π_b. Finite-time noise on $\widehat{\beta}$ scales as $O(1/\sqrt{W\alpha(1-\alpha)})$ (Serfling, 1980, Ch. 2), independent of per-agent table volumes. Whether $\widehat{\beta}$ admits the same three-timescale embedding as Algorithm 4.1's β is the first analytical question.

Coupled per-agent dual updates. The N updates (9.4) are not asymptotically independent. Each $\widehat{\mathrm{RC}}_\alpha^{(i)}$ depends on $\widehat{\beta}$ through the conditioning event and on $\pi_b = \prod_j \pi_b^{(j)}$ through the joint cost trajectory; a movement of $\lambda^{(j)}$ shifts $\pi_b^{(j)}$, shifts the joint cost distribution, and shifts both $\widehat{\beta}$ and $\widehat{\mathrm{RC}}_\alpha^{(i)}$ for every $i \neq j$. The quasi-stationarity argument that closes the centralised proof must be rephrased: π_b is quasi-stationary on the slow scale not because a single λ moves slowly but because the entire vector $(\lambda^{(1)}, \ldots, \lambda^{(N)})$ moves on a common slow scale. The basic two-timescale framework of Borkar (2008, §6.1) already admits vector iterates on each scale, and the saddle-point structure of (9.4) is the natural extension of the scalar case. The non-standard ingredient is the conditional structure of $\widehat{\mathrm{RC}}_\alpha^{(i)}$ on the shared event $\{Y > \widehat{\beta}\}$, which makes the slow-update noise process non-i.i.d. in a way that the natural-timescale averaging machinery of Borkar (2008, §§6.2–6.3) handles only after a controlled-noise hypothesis is verified for the joint $(\widehat{\beta}, \widehat{\mathrm{RC}}^{(i)})$ process. The cleanest path through these conditions — either by absorbing $\widehat{\beta}$ into the controlled state of the noise process, or by separating the quantile recursion onto its own intermediate timescale — is open.

Budget allocation as a fourth timescale. A static $\{C_\alpha^{(i)}\}$ leaves N scalar saddle conditions summing to the joint feasibility by Tasche's identity; the slow-scale analysis runs as N parallel scalar SA recursions on a common quasi-stationary π_b. An adaptive allocation updated on a slow timescale to track realised $\widehat{\mathrm{RC}}_\alpha^{(i)}$ proportions

introduces a fourth timescale slower than η_k (i.e. with step sizes ν_k satisfying $\nu_k \ll \eta_k$ asymptotically) and needs a corresponding tier in the quasi-stationarity hierarchy. Whether the gain in matching realised contributions outweighs the analytical cost is open. The static uniform allocation is the correct default for the symmetric Taxi MMDP; asymmetric domains require the question to be revisited.

9.4 Empirical agenda

A follow-up study would address three questions, in order. First, *closing the N=3 cell.* Online RCA should clear the five-component operational gate at N=3 on the Taxi MMDP within the same 6×10^6 per-seed episode budget that stalled the centralised learner. The prediction follows from Section 8.5: per-cell visit counts on the per-agent augmented table rise from ~ 1.5 on the joint table to $\sim 8.8 \times 10^3$ per agent, well inside any asynchronous-SA threshold, and dual-signal variance is split across N parallel per-agent signals. Failure would isolate a third mechanism this diagnosis did not name.

Second, *slope at $N \in \{2, \ldots, 6\}$.* A scaling exponent of episodes-to-gate against N on the per-agent representation is the first slope the centralised side could not measure; the exponential dependence of mechanism (1) is replaced by a linear one, so a meaningfully resolved slope should be reachable within compute budgets compatible with the present protocol.

Third, *tail-relative budget normalisation.* The per-capita convention $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ of Section 8.6 widens the standardised margin by $\sqrt{N}$, conflating mechanism (2)'s severity with constraint tightness. The tail-relative convention $C_\alpha^{\text{eff}} = \mathbb{E}[Y] + \kappa\sqrt{\text{Var}(Y)}$ holds the standardised margin fixed across N and disentangles the two; the per-agent representation makes the re-allocation tractable through (9.1) at the saddle. Domain extensions beyond Taxi — MMDPs with shared dynamic resources (de Nijs et al., 2021) and with cross-agent cost correlations — complete the agenda, the latter preserving Tasche's identity but enlarging the off-diagonal $\text{Cov}(C_i, C_j)$ terms in mechanism (2).

9.5 Scope and extensions

Two assumptions structure the construction as stated. Tabular representation underlies the closed-form residual of Proposition 4.2 and the cell-count reading of mechanism (1); weak coupling underlies the per-agent factorisation of the cost Q-function side. Both are restrictive, and the question of how the construction transports is open on each axis.

Function approximation: linear features. Linear function approximation on per-agent features carries through cleanest. Equation (4.5) has a projected-Bellman-operator analogue (Tsitsiklis and Van Roy, 1997), the projection-induced bias is bounded by the per-agent feature-coverage error, and the C3 buffer-mean estimator

is unchanged: mechanism (1)'s table-volume reading becomes a per-agent feature-coverage reading; mechanism (2)'s joint-return-variance reading is a property of the return distribution and transfers without modification.

Function approximation: deep parametric Q-functions. A different question. Mechanism (1)'s table-volume reading does not transfer verbatim, since parameter count is decoupled from cell count and per-cell visit reasoning is replaced by per-parameter capacity reasoning; mechanism (2) again transfers unchanged. Whether the centralised stall transfers to a centralised parametric Q-function on the Taxi MMDP is an empirical question the present results do not settle; the per-agent appeal retains a parametric reading nonetheless — per-agent networks have substantially smaller parameter counts than a joint network on the same MMDP.

Off-policy evaluation in the dual estimator. A third thread, separate from the function-class question. Targeting π_g directly through an importance-sampled CVaR estimator (Chow and Ghavamzadeh, 2014) is delicate: variance on tail events is the standing difficulty (Tamar et al., 2015), and its interaction with the per-agent decomposition is open.

Strong coupling. Tasche's identity (9.1) does not require weak coupling — it is an identity on random variables, not on dynamics — so the dual-side decomposition transfers to strongly coupled settings unchanged. The Q-function-side decomposition does not: without factorised dynamics, a per-agent Bellman recursion on $|S_i||A_i|$ tables is not well-defined. Strongly coupled multi-agent CVaR-constrained planning is therefore a separate problem. Centralised training with decentralised execution would attach the C1+C2+C3 machinery to the training-time joint cost Q-function; coordination-graph factorisations of the joint Q^C over a sparse interaction structure would extend the single-agent decoupling clique-by-clique. Each constitutes a separate research programme rather than an extension of online RCA.

9.6 Closing

The thesis set out to scale Borkar and Jain's learner to a multi-agent setting. The single-agent fix proved more demanding than first appeared — the gap between the printed recursion and a Bellman-consistent one was small enough to be confused with calibration error and structurally distinct enough to matter — and the multi-agent extension more resistant still. We close with the diagnosis: an exact identity at z_0 that localises the finite-time coupling Borkar and Jain (2014, Theorem 2) treats as a limiting equivalence (Proposition 4.2); three role separations on a single object that preserve the limiting projected dual equilibrium under the standard quasi-stationarity argument of Borkar (2008) (C1, C2, C3); and two structurally disjoint quantities of the joint representation that the corrected learner does not reach within budget at N=3 (Chapter 8). We do not establish finite-time rates under the modified operator, an N-scaling exponent for the centralised stall, or a

convergence theorem for the parallel-asynchronous per-agent dual scheme sketched above.

What we leave for the next stage is the per-agent construction the diagnosis points to (Sections 9.1–9.2): online RCA on the Taxi MMDP under the same protocol, then the N-slope at $N \in \{2, \ldots, 6\}$ on the per-agent representation, both as set out in Section 9.4 and contingent on the analytical questions of Section 9.3. The thesis closes here.

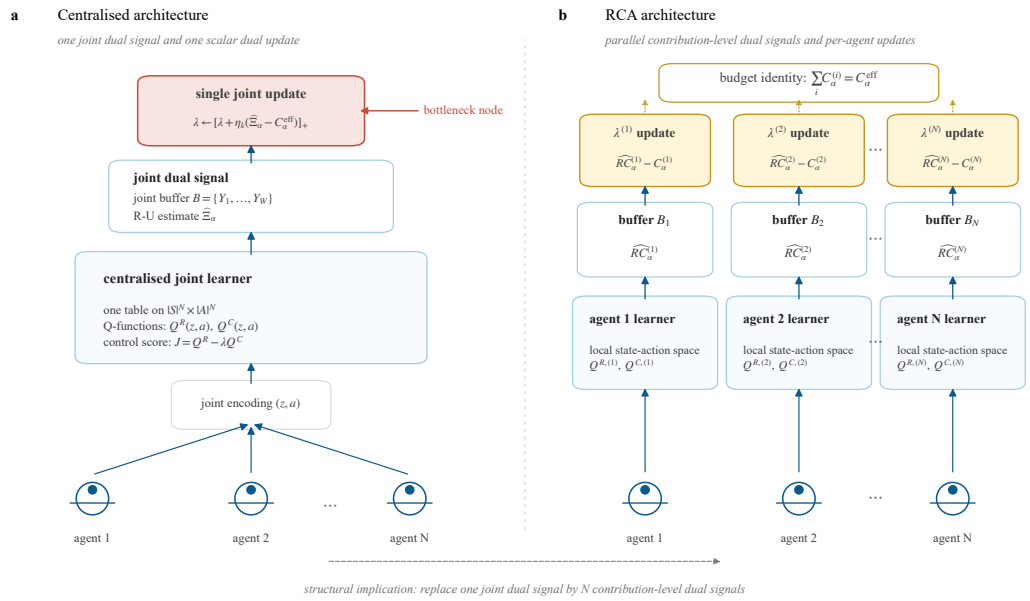

Figure 9.1. Centralised representation versus Risk-Contribution Allocation. **(a) Centralised, this thesis.** N agents feed a single joint encoding (z, a) into one centralised learner over $|S|^N \cdot |A|^N$, with Q_t^R and Q_t^C stored on the joint table and the score $J_t = Q_t^R - \lambda Q_t^C$ driving action selection. A single rolling joint-return buffer $\mathcal{B} = (Y_1, \ldots, Y_W)$ supplies $\widehat{\Xi}_\alpha$ to one scalar dual update $\lambda \in \mathbb{R}_{\geq 0}$. Mechanisms (1) and (2) of Chapter 8 act on the joint table volume and on this single joint dual signal respectively. **(b) Risk-Contribution Allocation, this chapter.** N per-agent learners run on local state–action spaces with per-agent Q_i^R, Q_i^C tables; per-agent buffers $\mathcal{B}_i$ supply $\widehat{\mathrm{RC}}_\alpha^{(i)}$ via (9.3) to N parallel per-agent dual updates (9.4), with $(\lambda^{(1)}, \ldots, \lambda^{(N)}) \in \mathbb{R}^N_{\geq 0}$, per-agent budgets summing to the joint constraint by (9.1). The only joint object that survives the split is the shared scalar $\widehat{\beta}$ that selects the joint-tail event consumed by every $\widehat{\mathrm{RC}}_\alpha^{(i)}$. The reuse of C1, C2, C3 inside each per-agent learner is described in the body of this chapter.

References

Agrawal, P., P. Varakantham and W. Yeoh (2016). 'Scalable Greedy Algorithms for Task/Resource Constrained Multi-Agent Stochastic Planning'. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI Press, pp. 10–16. URL: https://www.ijcai.org/Proceedings/16/Papers/009.pdf.

Altman, E. (1999). *Constrained Markov Decision Processes*. Boca Raton, FL: Chapman and Hall/CRC. DOI: 10.1201/9781315140223.

Bastani, O., Y. J. Ma, E. Shen and W. Xu (2022). 'Regret Bounds for Risk-Sensitive Reinforcement Learning'. In: *Advances in Neural Information Processing Systems 35*, pp. 36259–36269. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/eb4898d622e9a48b5f9713ea1fcff2bf-Abstract-Conference.html.

Bellemare, M. G., W. Dabney and R. Munos (2017). 'A Distributional Perspective on Reinforcement Learning'. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 449–458. URL: https://proceedings.mlr.press/v70/bellemare17a.html.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Vol. 48. Texts and Readings in Mathematics. Gurgaon, India: Hindustan Book Agency. DOI: 10.1007/978-93-86279-38-5.

Borkar, V. S. and R. Jain (2014). 'Risk-Constrained Markov Decision Processes'. In: *IEEE Transactions on Automatic Control* 59.9, pp. 2574–2579. DOI: 10.1109/TAC.2014.2309262.

Brown, D. B. (2007). 'Large Deviations Bounds for Estimating Conditional Value-at-Risk'. In: *Operations Research Letters* 35.6, pp. 722–730. DOI: 10.1016/j.orl.2007.01.001.

Chow, Y. and M. Ghavamzadeh (2014). 'Algorithms for CVaR Optimization in MDPs'. In: *Advances in Neural Information Processing Systems 27*, pp. 3509–3517. URL: https://proceedings.neurips.cc/paper_files/paper/2014/hash/35f1050a4381d2d216bf56ad46b0277d-Abstract.html.

Dabney, W., G. Ostrovski, D. Silver and R. Munos (2018). 'Implicit Quantile Networks for Distributional Reinforcement Learning'. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1096–1105. URL: https://proceedings.mlr.press/v80/dabney18a.html.

De Nijs, F., E. Walraven, M. M. de Weerdt and M. T. J. Spaan (2017). 'Bounding the Probability of Resource Constraint Violations in Multi-Agent MDPs'. In:

Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, pp. 3562–3568. DOI: 10.1609/aaai.v31i1.11037.

De Nijs, F., E. Walraven, M. M. de Weerdt and M. T. J. Spaan (2021). 'Constrained Multiagent Markov Decision Processes: A Taxonomy of Problems and Algorithms'. In: *Journal of Artificial Intelligence Research* 70, pp. 955–1001. DOI: 10.1613/jair.1.12233.

Ding, D., K. Zhang, T. Başar and M. R. Jovanović (2020). 'Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes'. In: *Advances in Neural Information Processing Systems 33*, pp. 8378–8390. URL: https://proceedings.neurips.cc/paper/2020/hash/5f7695debd8cde8db5abcb9f161b49ea-Abstract.html.

Efroni, Y., S. Mannor and M. Pirotta (2020). 'Exploration-Exploitation in Constrained MDPs'. arXiv: 2003.02189 [cs.LG].

Even-Dar, E. and Y. Mansour (2003). 'Learning Rates for Q-Learning'. In: *Journal of Machine Learning Research* 5, pp. 1–25. URL: https://jmlr.org/papers/v5/evendar03a.html.

Gautier, A., B. Lacerda, N. Hawes and M. J. Wooldridge (2023a). 'Multi-Unit Auctions for Allocating Chance-Constrained Resources'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10, pp. 11560–11568. DOI: 10.1609/aaai.v37i10.26366.

Gautier, A., M. Rigter, B. Lacerda, N. Hawes and M. J. Wooldridge (2023b). 'Risk-Constrained Planning for Multi-Agent Systems with Shared Resources'. In: *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 113–121. URL: https://dl.acm.org/doi/10.5555/3545946.3598626.

Gautier, A., A. Stephens, B. Lacerda, N. Hawes and M. J. Wooldridge (2022). 'Negotiated Path Planning for Non-Cooperative Multi-Robot Systems'. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 472–480. URL: https://dl.acm.org/doi/10.5555/3535850.3535904.

Haskell, W. B. and R. Jain (2015). 'A Convex Analytic Approach to Risk-Aware Markov Decision Processes'. In: *SIAM Journal on Control and Optimization* 53.3, pp. 1569–1598. DOI: 10.1137/140969221.

Lim, S. H. and I. Malik (2022). 'Distributional Reinforcement Learning for Risk-Sensitive Policies'. In: *Advances in Neural Information Processing Systems 35*, pp. 30977–30989. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c88a2bd0e793550d0e885aa6e31ca277-Abstract-Conference.html.

Meuleau, N., M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. Dean and C. Boutilier (1998). 'Solving Very Large Weakly Coupled Markov Decision Processes'. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference*. Cambridge, MA: AAAI Press / MIT Press, pp. 165–172. URL: https://aaai.org/papers/00165-aaai98-023-solving-very-large-weakly-coupled-markov-decision-processes/.

Muni, A., V. Taboga, E. Derman, P.-L. Bacon and E. Delage (2026). 'Reward Redistribution for CVaR MDPs using a Bellman Operator on L^∞'. arXiv: 2602.03778 [cs.LG].

Prashanth, L. A. and M. C. Fu (2022). 'Risk-Sensitive Reinforcement Learning via Policy Gradient Search'. In: *Foundations and Trends in Machine Learning* 15.5, pp. 537–693. DOI: 10.1561/2200000091.

Rockafellar, R. T. and S. Uryasev (2000). 'Optimization of Conditional Value-at-Risk'. In: *The Journal of Risk* 2.3, pp. 21–41. DOI: 10.21314/JOR.2000.038.

Russell, S. J. and A. L. Zimdars (2003). 'Q-Decomposition for Reinforcement Learning Agents'. In: *Proceedings of the Twentieth International Conference on Machine Learning*. AAAI Press, pp. 656–663. URL: https://dl.acm.org/doi/10.5555/3041838.3041921.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons. DOI: 10.1002/9780470316481.

Shen, S., C. Ma, C. Li, W. Liu, Y. Fu, S. Mei, X. Liu and C. Wang (2023). 'RiskQ: Risk-Sensitive Multi-Agent Reinforcement Learning Value Factorization'. In: *Advances in Neural Information Processing Systems 36*, pp. 34791–34825. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/6d3040941a2d57ead4043556a70dd728-Abstract-Conference.html.

Stooke, A., J. Achiam and P. Abbeel (2020). 'Responsive Safety in Reinforcement Learning by PID Lagrangian Methods'. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9133–9143. URL: https://proceedings.mlr.press/v119/stooke20a.html.

Tamar, A., Y. Glassner and S. Mannor (2015). 'Optimizing the CVaR via Sampling'. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 2993–2999. DOI: 10.1609/aaai.v29i1.9561.

Tasche, D. (1999). *Risk Contributions and Performance Measurement*. Working paper. Technische Universität München. URL: https://www.researchgate.net/publication/2378752.

Tessler, C., D. J. Mankowitz and S. Mannor (2019). 'Reward Constrained Policy Optimization'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkfrvsA9FX.

Trindade, A. A., S. Uryasev, A. Shapiro and G. Zrazhevsky (2007). 'Financial Prediction with Constrained Tail Risk'. In: *Journal of Banking & Finance* 31.11, pp. 3524–3538. DOI: 10.1016/j.jbankfin.2007.04.014.

Tsitsiklis, J. N. (1994). 'Asynchronous Stochastic Approximation and Q-Learning'. In: *Machine Learning* 16.3, pp. 185–202. DOI: 10.1023/A:1022689125041.

Tsitsiklis, J. N. and B. Van Roy (1997). 'An Analysis of Temporal-Difference Learning with Function Approximation'. In: *IEEE Transactions on Automatic Control* 42.5, pp. 674–690. DOI: 10.1109/9.580874.

Wang, K., N. Kallus and W. Sun (2023). 'Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR'. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 35864–35907. URL: https://proceedings.mlr.press/v202/wang23m.html.

Wu, J. and E. H. Durfee (2010). 'Resource-Driven Mission-Phasing Techniques for Constrained Agents in Stochastic Environments'. In: *Journal of Artificial Intelligence Research* 38, pp. 415–473. DOI: 10.1613/jair.3004.

A

Taxi MMDP schedule

This appendix gives the per-agent reward, cost, and transition schedule for the Taxi MMDP introduced in Chapter 6, Section 6.2. The schedule is per agent and is lifted to the joint MMDP via the factor decomposition of (6.2); the joint budget is $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ with $C_{\text{base}} = 1.60$. Numerical values reproduce `grid_env_taxi.py`; at $N = 1$ they match the single-system reference `grid_env_separate_cost.py` bit-for-bit under any common seed sequence.

Geometry, indexing, and notation

Three lanes, indexed $\ell \in \{0, 1, 2\}$, with the convention

$\ell = 0, 2$: shoulder (safe, low reward); $\ell = 1$: centre (high reward, hazardous in central columns).

The per-agent row state is the lane ℓ. The column is a shared finite-horizon time index with $T = 12$: decision columns are $x \in \{0, \ldots, T-1\}$, all agents share $x_i \equiv t$, and the terminal column index is T. The column is recorded separately from the per-agent lane state and is not an additional row-state factor. Initial lane state shared across agents: $\ell_{i,0} = 1$ at column 0. Three local actions $a_i \in \mathcal{A}_0 = \{\texttt{up}, \texttt{stay}, \texttt{down}\}$, mapped to lane increments $\delta(\texttt{up}) = -1,\ \delta(\texttt{stay}) = 0,\ \delta(\texttt{down}) = +1$. For any lane ℓ and action a, the post-action intended lane is

$$\widetilde{\ell} = \text{clip}\big(\ell + \delta(a),\ 0,\ 2\big). \tag{A.1}$$

The hazard band is the central column range

$$\mathcal{H} = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}. \tag{A.2}$$

Lane transition kernel $P_{0,t}(\ell_i' \mid \ell_i, a_i)$

Let $\widetilde{\ell}_i$ be the intended lane (Eq. (A.1)) and $x_i' = x_i + 1$ the next column. The kernel is hazard-band-conditioned:

- **Outside the hazard band** ($x_i' \notin \mathcal{H}$): the transition is deterministic, $\Pr(\ell_i' = \widetilde{\ell}_i) = 1$.

- **Inside the hazard band** ($x_i' \in \mathcal{H}$): the kernel is asymmetric:

Intended $\widetilde{\ell}_i$	$\Pr(\ell_i' = 0)$	$\Pr(\ell_i' = 1)$	$\Pr(\ell_i' = 2)$
0 (shoulder)	0.85	0.15	0
1 (centre)	0.075	0.85	0.075
2 (shoulder)	0	0.15	0.85

The centre lane is sticky and the shoulder lanes are biased towards the centre; outside the hazard band the lane evolves deterministically. Boundary clipping in Eq. (A.1) saturates the lane index at the boundary; the state remains non-terminal.

Per-step cost $c_{0,t}(s_i, a_i) = c_t^{\text{haz}}(s_i) + c^{\text{sw}}(a_i)$

Hazard cost $c^{\text{haz}}(s_i)$**.** Charged on the current (ℓ_i, x_i). The centre lane within the hazard band pays a column-indexed ramp; shoulders pay zero:

$$c^{\text{haz}}(\ell_i, x_i) = \begin{cases} h(x_i), & \ell_i = 1 \text{ and } x_i \in \mathcal{H}, \\ 0, & \text{otherwise}, \end{cases} \tag{A.3}$$

with the hazard ramp

x	2	3	4	5	6	7	8	9	10
$h(x)$	0.20	0.20	0.30	0.30	0.40	0.40	0.30	0.30	0.20

The cumulative hazard for an agent that holds the centre lane through the entire band is $\sum_{x \in \mathcal{H}} h(x) = 2.60$. Since the episode-end smoothing shock is centred, $\mathbb{E}[\xi] = 0$ (Section 6.2 of Chapter 6), the expected smoothed cumulative cost holding centre throughout is 2.60, which sits well above $C_{\text{base}} = 1.60$ and is what makes the centre-lane trade-off non-trivial under the CVaR constraint.

Lane-switch cost $c^{\text{sw}}(a_i)$**.** Charged on the action: $c^{\text{sw}}(\texttt{stay}) = 0$, $c^{\text{sw}}(\texttt{up}) = c^{\text{sw}}(\texttt{down}) = 0.10$.

Per-step bound. $\bar{c} = \max_x h(x) + 0.10 = 0.40 + 0.10 = 0.50$ per agent, hence $\bar{C}^{(N)} \leq N \cdot 0.50$ and the joint cumulative-cost upper bound is $T \cdot N \cdot 0.50$.

Stage reward $r_{0,t}(s_i, a_i)$

The stage reward is keyed to the post-action intended lane $\widetilde{\ell}_i$ and to whether the next column is the terminal column. Writing $x_i' = x_i + 1$:

$$r_{0,t}(s_i, a_i) = \begin{cases} 0.75, & \widetilde{\ell}_i = 1 \text{ (centre, every step)}, \\ 0.20, & \widetilde{\ell}_i \in \{0, 2\} \text{ and } x_i' < T, \\ 0.00, & \widetilde{\ell}_i \in \{0, 2\} \text{ and } x_i' = T. \end{cases} \tag{A.4}$$

Equivalently:

Intended $\widetilde{\ell}_i$	**Non-terminal step** $(x'_i < T)$	**Terminal step** $(x'_i = T)$
1 (centre)	0.75	0.75
$\{0, 2\}$ (shoulder)	0.20	0.00

Centre-lane intent is rewarded uniformly at every step, including the terminal one, so ending in the centre is the unique terminal reward maximiser. The reward is a function of the intended (clipped, pre-stochastic) lane alone, with the realised next lane playing no role.

Boundary, absorption, termination

Lane boundary. The lane index saturates at $\ell \in \{0, 2\}$ via the clipping in Eq. (A.1). The state remains non-terminal at the boundary.

No per-agent absorption. There is no per-agent absorbing goal state. Per-step quantities (reward, hazard cost, lane-switch cost) accrue at every stage including the terminal stage; agents that have reached the boundary continue to incur lane-switch costs if a non-stay action is selected.

Episode termination. The episode terminates after T joint decisions (column indices $x = 0, 1, \ldots, T-1$); the truncation is recorded at the terminal column index. Stage rewards at the terminal index follow Eq. (A.4); no additional one-shot terminal bonus is paid.

Joint MMDP and CVaR constraint

The joint MMDP is the centralised lift of $\mathcal{M}_0$ over N agents, with joint lane-state space $\mathcal{S}^{(N)} = \mathcal{S}_0^N$, $|\mathcal{S}^{(N)}| = 3^N$, joint action space $\mathcal{A}^{(N)} = \mathcal{A}_0^N$, $|\mathcal{A}^{(N)}| = 3^N$, and factorised transition / reward / cost as in (6.2). The joint cumulative cost $Y_T^{(N)} = \sum_{t=0}^{T-1} C_t^{(N)}(s_t, a_t)$ is constrained at the per-capita invariant joint budget $C_\alpha^{\text{eff}} = N \cdot C_{\text{base}}$ with $C_{\text{base}} = 1.60$; the episode-end smoothing shock $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ with $\sigma_\xi = 0.10$ enters the buffer and the terminal indicator only, not the Q^C stage source (Section 6.2 of Chapter 6).

Risk-neutral policy and the constrained trade-off

The risk-neutral optimum for an isolated agent is to hold the centre lane throughout: per-step reward 0.75 on the centre strictly dominates the 0.20 shoulder reward. This policy accrues a deterministic cumulative hazard of 2.60; with the centred episode-end shock $\mathbb{E}[\xi] = 0$ the expected smoothed cumulative cost is 2.60, which exceeds $C_{\text{base}} = 1.60$. The CVaR-constrained learner of Algorithm 4.1 therefore trades reward for tail-risk reduction by occasionally moving to the shoulder lanes during the

hazard band, paying lane-switch cost 0.10 to avoid the larger column-indexed hazard. The $(1-\alpha)$-tail of the joint cumulative cost is held below C_α^{eff} at the saddle point with $\bar{\lambda}(\widehat{\Xi}_\alpha - C_\alpha^{\text{eff}}) = 0$ and $\bar{\lambda} \geq 0$.

Reproducibility

The schedule above reproduces `grid_env_taxi.py` verbatim. The per-agent dynamics, reward, and cost are bit-equivalent to the single-system reference `grid_env_separate_cost.py`; at $N = 1$ the joint MMDP yields an identical (state, action, transition, reward, cost) trajectory under any common seed sequence, and Algorithm 4.1 therefore produces an identical learner state at $N = 1$ to what it produces on the single-system environment. All numerical constants in this appendix are frozen across $N \in \{1, 2, 3\}$ and across all seeds in the centralised stress-test protocol of Chapter 6.