Distributional Reinforcement Learning:

Risk Optimization and Moment Matching

Aurélien Garivier

UMPA, ENS de Lyon

joint work with Justin Bec, Juliet Bringas Miranda, Olivier Cappé, Alexandre Marthe, Thomas Le Corre, Claire Vernade

Workshop on Stochastic Processes with Reinforcement
Berlin Mathematics Research Center MATH+ — November 5, 2026

UMPA
ENS DE LYON

Outline

1. Markov decision processes and the law of the return

2. What can dynamic programming optimize?

3. How to represent the law of the return?

Part 1

Markov decision processes and the law of the return

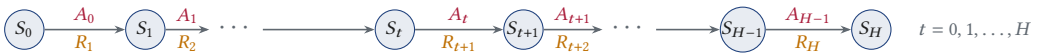
1. Markov decision processes and the law of the return

2. What can dynamic programming optimize?

3. How to represent the law of the return?

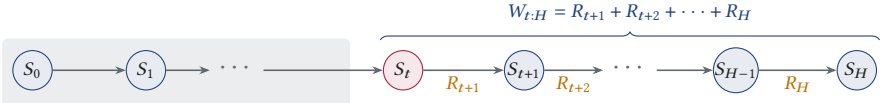
Markov decision processes: the setting

- **Data**: finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, horizon H; for every $t \in \{0, \dots, H-1\}$, a **transition kernel** $p_t(s' \mid s, a)$ and a **reward law** $\rho_t(s, a, s') \in \mathcal{M}_1([-R_{\max}, R_{\max}])$ (a probability law on that interval), of mean $r_t(s, a, s') \coloneqq \int x \, \rho_t(s, a, s')(\mathrm{d}x)$; an initial law $\eta_0 \in \mathcal{M}_1(\mathcal{S})$.
- **Protocol**: $S_0 \sim \eta_0$. For $t = 0, \dots, H-1$: the agent sees the **history** $\mathcal{H}_t = (S_0, A_0, R_1, S_1, \dots, S_t)$ and chooses A_t; nature draws $S_{t+1} \sim p_t(\cdot \mid S_t, A_t)$, then, given the past and S_{t+1}, $R_{t+1} \sim \rho_t(S_t, A_t, S_{t+1})$. Stop at S_H.
- **Policy** $\pi = (\pi_0, \dots, \pi_{H-1})$: $A_t = \pi_t(\mathcal{H}_t)$. **Markov** policy: $A_t = \pi_t(S_t)$; then $(S_t)_{t \le H}$ is a Markov chain with kernels $p_t^\pi(s, s') = p_t(s' \mid s, \pi_t(s))$: a policy = a chain chosen in a family.



The cumulated reward and its law

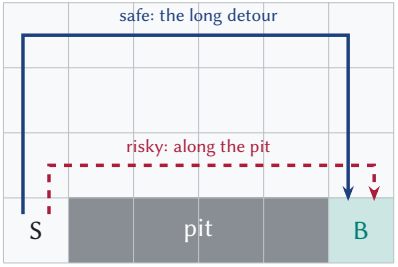
- **Cumulated reward** from time t: $W_{t:H} \coloneqq \sum_{k=t+1}^{H} R_k$; $\quad W_{H:H} = 0$, $|W_{t:H}| \leq (H-t)R_{\max}$.
- **Goal**: find a policy maximizing $\mathbb{E}^{\pi}[W_{0:H}]$.



Definition (Return distributions). For a Markov policy π, $t \in \{0, \dots, H\}$ and $s \in \mathcal{S}$: $v_t^{\pi}(s) \coloneqq \mathcal{L}\big(W_{t:H} \,\big|\, S_t = s\big)$; in particular $v_H^{\pi}(s) = \delta_0$.
Markov property: $\mathcal{L}(W_{t:H} \mid \mathcal{H}_t) = \mathcal{L}(W_{t:H} \mid S_t) = v_t^{\pi}(S_t)$.

Classical RL: the **value** $V_t^{\pi}(s) = \mathbb{E}[W_{t:H} \mid S_t = s] = \int x \, v_t^{\pi}(s)(\mathrm{d}x)$. Distributional RL: the whole law $v_t^{\pi}(s)$. Risk-sensitive RL: a functional of it.

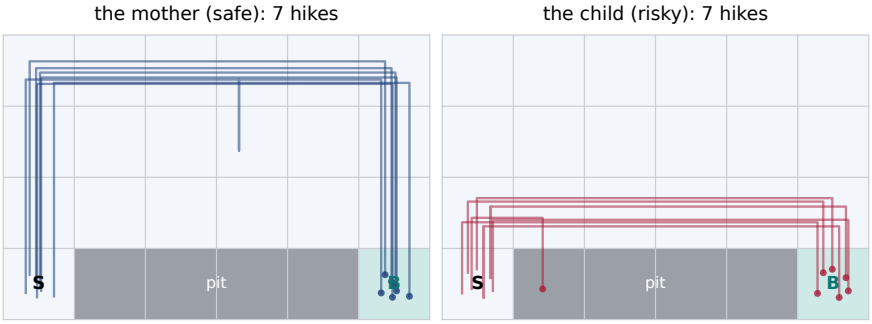
Example 1: the windy hike



- **States**: the 20 cells and a sink; **start** S, **bar** B (absorbing, reward 1 at every step there); the bottom row is a pit (fall = sink, absorbing, no reward).
- **Actions** N, S, E, W. **Wind**: with probability $p = 0.08$ the hiker moves one cell south instead. Walls stop.
- **Reward** $R_{t+1} = \mathbb{1}\{S_{t+1} = \mathrm{B}\}$, i.e. $\rho_t(s, a, s') = \delta_{\mathbb{1}\{s'=\mathrm{B}\}}$; $H = 16$. $W_{0:H}$ = **time spent at the bar** $\in \{0, \dots, 16\}$.
- Two hikers, two Markov policies: **the mother** (safe): climb, east, down; **the child** (risky): one row above the pit.

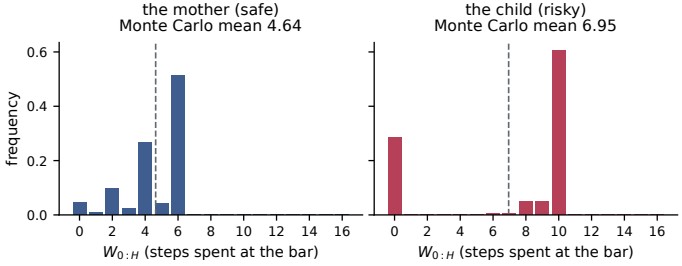
After the cliff walk of Sutton and Barto (2018); $|\mathcal{S}| = 21$, $|\mathcal{A}| = 4$.

The windy hike, simulated



- **The mother** reaches the bar at time 11, for 6 steps there, if the wind spares her on the 8 steps north and east ($0.92^8 \approx 0.51$); an isolated interior push while climbing or moving east costs her two steps; she climbs back, but consecutive pushes can still make her fall: $\mathbb{P}(\text{fall}) \approx 0.24\%$.
- **The child** is at the bar at time 7, for 10 steps, if the wind spares her 6 times ($0.92^6 \approx 0.61$); a push along the pit ($1 - 0.92^4 \approx 0.28$) ends the hike with $W_{0:H} = 0$.

Monte Carlo: the law of the time at the bar



$N = 10^5$ hikes	mother	child
mean of $W_{0:H}$	4.64	**6.95**
standard deviation	1.7	4.4
$\mathbb{P}(W_{0:H} = 0)$	0.05	**0.29**
median	6	10

Empirical frequencies of $W_{0:H}$ over 10^5 independent simulated hikes: an estimate of the law $\nu_0^\pi(\mathrm{S})$.

Example 2: the bike retail store

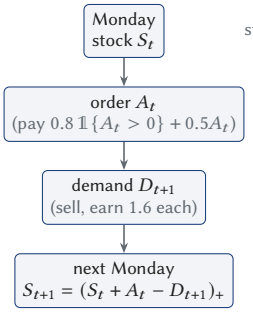
- **State** $S_t \in \{0, \dots, M\}$: bikes in stock on Monday of week t; capacity $M = 15$.
- **Action** $A_t \in \{0, \dots, M - S_t\}$: bikes ordered, delivered at once (a state-dependent action set).
- **Demand** D_{t+1}, i.i.d., $\mathbb{P}(D = k) = 0.1 \cdot 0.9^k$ for $k < M$, $\mathbb{P}(D = M) = 0.9^M$ (mean 7.1); sales $(S_t + A_t) \wedge D_{t+1}$;

$$S_{t+1} = \big(S_t + A_t - D_{t+1}\big)_+.$$

- **Reward**: price 1.6 per bike sold, 0.5 per bike ordered plus 0.8 per order, storage 0.3 per bike on Monday:

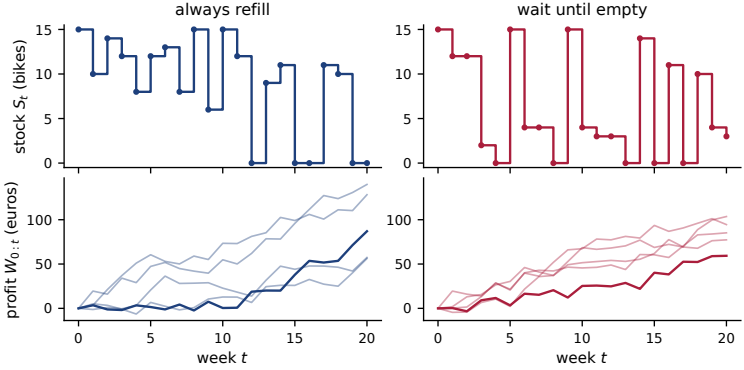
$$R_{t+1} = -0.3\, S_t - 0.8\, \mathbb{1}\{A_t > 0\} - 0.5\, A_t + 1.6\, \big((S_t + A_t) \wedge D_{t+1}\big).$$

- $H = 20$ weeks, full store at the start ($S_0 = 15$). $W_{0:H}$ = **total profit**.



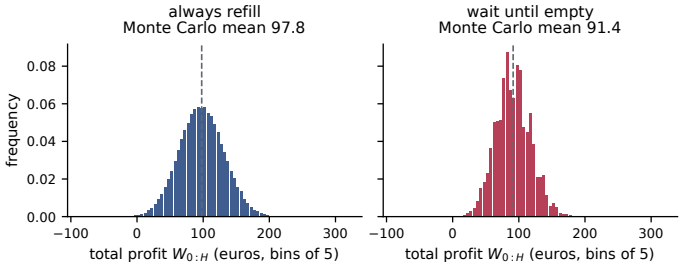
Two managers, two Markov policies:
always refill: $A_t = M - S_t$;
wait until empty: $A_t = M\, \mathbb{1}\{S_t = 0\}$.

The bike store, simulated



- Runs of 20 weeks under each policy: the stock on Mondays along one run (top) and the cumulated profit $W_{0:t}$ along five runs (bottom), the darker curve being the run shown above.
- **Always refill** holds more stock (storage) and orders almost every week (fee); it never misses a sale. **Wait until empty** saves fees and storage, but misses sales when its remaining stock falls short of the demand (an empty store is refilled at once).

Monte Carlo: the law of the total profit



$N = 10^5$ runs	**refill**	**wait**
mean of $W_{0:H}$	**97.8**	91.4
standard deviation	33.6	26.4
10% quantile	54.8	58.6
$\mathrm{CVaR}_{10\%}$	39.7	**47.2**

euros; bins of width 5 on the histograms; $\mathrm{CVaR}_{10\%}$ = mean of the worst tenth of the seasons

Always refill earns more on average, and his worst seasons are worse: more to gain, more to lose.
Wait until empty has better "bad seasons".

Classical policy evaluation: the Bellman equation

Proposition (Bellman equation). Markov policy π, **value** $V_t^\pi(s) \coloneqq \mathbb{E}^\pi[W_{t:H} \mid S_t = s]$ (so $V_H^\pi \equiv 0$). For $t \in \{0, \dots, H-1\}$, $s \in \mathcal{S}$, $a = \pi_t(s)$,

$$V_t^\pi(s) \;=\; \sum_{s'} p_t(s' \mid s, a)\Big[r_t(s, a, s') + V_{t+1}^\pi(s')\Big].$$

Proof. With $a = \pi_t(s)$:

$$V_t^\pi(s) \;=\; \mathbb{E}\big[W_{t:H} \,\big|\, S_t = s\big] \overset{\text{(i)}}{=} \mathbb{E}\big[R_{t+1} + W_{t+1:H} \,\big|\, S_t = s\big] \overset{\text{(ii)}}{=} \sum_{s'} p_t(s' \mid s, a)\, \mathbb{E}\big[R_{t+1} + W_{t+1:H} \,\big|\, S_t = s,\, S_{t+1} = s'\big]$$

$$\overset{\text{(iii)}}{=} \sum_{s'} p_t(s' \mid s, a)\Big[r_t(s, a, s') + \mathbb{E}\big[W_{t+1:H} \,\big|\, S_{t+1} = s'\big]\Big] \;=\; \sum_{s'} p_t(s' \mid s, a)\Big[r_t(s, a, s') + V_{t+1}^\pi(s')\Big].$$

(i) **additivity**: $W_{t:H} = R_{t+1} + W_{t+1:H}$; (ii) **tower property**: condition on $S_{t+1} \sim p_t(\cdot \mid s, a)$; (iii) **linearity**, the **protocol** ($R_{t+1} \sim \rho_t(s, a, s')$ has mean $r_t(s, a, s')$) and the **Markov property**. ∎

Definition (Bellman operator). $V_t^\pi = T_t^\pi V_{t+1}^\pi$, where $T_t^\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is defined by $(T_t^\pi V)(s) \coloneqq \sum_{s'} p_t(s' \mid s, \pi_t(s))\big[r_t(s, \pi_t(s), s') + V(s')\big]$, an affine map.

Backward induction: $V_0^\pi = T_0^\pi T_1^\pi \cdots T_{H-1}^\pi\, 0$, in $O(H|\mathcal{S}|^2)$ operations, for $|\mathcal{S}|^H$ paths.

Planning: the optimal Bellman operator

- Backward induction again, with a maximization at every state: $V_H^\star \equiv 0$ and, for $t = H-1, \dots, 0$,

$$Q_t^\star(s,a) \;\coloneqq\; \sum_{s'} p_t(s' \mid s,a)\Big[r_t(s,a,s') + V_{t+1}^\star(s')\Big], \qquad V_t^\star(s) \;\coloneqq\; \max_{a \in \mathcal{A}} Q_t^\star(s,a) \;\coloneqq\; (T_t^\star V_{t+1}^\star)(s).$$

- The **greedy policy** is the Markov policy $\pi_t^\star(s) \in \arg\max_a Q_t^\star(s,a)$ (any choice of the maximizer).
- **Cost**: $O(H|\mathcal{S}|^2|\mathcal{A}|)$ operations, for $|\mathcal{A}|^{|\mathcal{S}|H}$ Markov policies.

Question. In which sense is the greedy policy optimal? Optimal among which policies?

For a general (history-dependent) policy π, write $V_t^\pi(h_t) \coloneqq \mathbb{E}^\pi[W_{t:H} \mid \mathcal{H}_t = h_t]$ for a history $h_t = (s_0, a_0, r_1, \dots, s_t)$: it depends on the whole history. By the tower property and the protocol, with $a = \pi_t(h_t)$,

$$V_t^\pi(h_t) \;=\; \mathbb{E}^\pi\Big[R_{t+1} + V_{t+1}^\pi(\mathcal{H}_{t+1}) \,\Big|\, \mathcal{H}_t = h_t\Big], \qquad \mathcal{H}_{t+1} = (h_t, a, R_{t+1}, S_{t+1}),\; S_{t+1} \sim p_t(\cdot \mid s_t, a),\; R_{t+1} \sim \rho_t(s_t, a, S_{t+1}).$$

Bellman's principle of optimality

Theorem (Optimality of the dynamic programming policy). [Bellman 1957]

1. For every policy π (history-dependent), every t and every history h_t: $V_t^{\pi}(h_t) \le V_t^{\star}(s_t)$.
2. The greedy policy $\pi^{\star}$ satisfies $V_t^{\pi^{\star}} = V_t^{\star}$ for all t.

Hence, for every s: $V_0^{\star}(s) = V_0^{\pi^{\star}}(s) = \max_{\pi} V_0^{\pi}(s)$, the maximum over *all* policies being attained by the *Markov* policy computed by dynamic programming.

Proof. (1) Backward induction on t. At $t = H$ both sides vanish. Assume the claim at $t+1$ and let $a = \pi_t(h_t)$:

$$V_t^{\pi}(h_t) = \mathbb{E}\big[R_{t+1} + V_{t+1}^{\pi}(\mathcal{H}_{t+1}) \,\big|\, \mathcal{H}_t = h_t\big] \overset{\text{(i)}}{\le} \mathbb{E}\big[R_{t+1} + V_{t+1}^{\star}(S_{t+1}) \,\big|\, \mathcal{H}_t = h_t\big]$$
$$\overset{\text{(ii)}}{=} Q_t^{\star}(s_t, a) \;\le\; \max_{a'} Q_t^{\star}(s_t, a') = V_t^{\star}(s_t).$$

(i) induction hypothesis $V_{t+1}^{\pi}(\mathcal{H}_{t+1}) \le V_{t+1}^{\star}(S_{t+1})$ pointwise + **monotonicity** of $\mathbb{E}$;
(ii) **Markov transition** + **linearity** turns the expectation into $\sum_{s'} p_t(s' \mid s_t, a)\,[r_t(s_t, a, s') + V_{t+1}^{\star}(s')]$;
(2) For $\pi^{\star}$, Markov: (i) is an equality by induction ($V_{t+1}^{\pi^{\star}} = V_{t+1}^{\star}$). ∎

Memory is useless for the expected return, DP gives the optimal policy in $O\big(H|\mathcal{S}|^2|\mathcal{A}|\big)$ operations.

What the proof used, and dynamic programming on the examples

Three properties of the expectation $\nu \mapsto \int x \, \mathrm{d}\nu$:

1. **affine under mixtures**: $\mathbb{E}[\sum_i w_i \nu_i] = \sum_i w_i \mathbb{E}[\nu_i]$ (the next state is random);
2. **additive under convolutions**: $\mathbb{E}[\rho * \nu] = \mathbb{E}[\rho] + \mathbb{E}[\nu]$ (the reward adds to the future);
3. **monotone** for the stochastic order: $\nu \leq_{\mathrm{st}} \nu' \Rightarrow \mathbb{E}[\nu] \leq \mathbb{E}[\nu']$ (replacing a continuation by a better one).

Properties 1–2 make the recursion exact (evaluation); Property 3 makes the greedy choice optimal (planning). Part 2 asks which other functionals Ψ of the law of $W_{0:H}$ enjoy the same privilege.

Windy hike ($H = 16$): $V_0^\star(\mathrm{S}) = 7.18$, against 6.97 for the child and 4.63 for the mother: the optimal hiker for the mean walks one row above the child and reaches the bar at time 9 on its wind-free path.

Bike store ($H = 20$): $V_0^\star(15) = 102.4$ euros, against 97.9 (refill) and 91.2 (wait). The optimal policy is an (s, S) rule: on weeks 0 to 17, refill to 15 when the stock is at most 8, nothing otherwise; the levels drop to $(8, 14)$, then $(5, 11)$, in the last two weeks.

Distributional policy evaluation

For laws on $\mathbb{R}$: **mixture** $\sum_i w_i \nu_i$; **convolution** $\rho * \nu = \mathcal{L}(R+X)$ for independent $R \sim \rho$, $X \sim \nu$.

Theorem (Distributional Bellman equation). Markov policy π; $\nu_H^\pi = \delta_0$ and, for $t \in \{0, \dots, H-1\}$, $s \in \mathcal{S}$, $a = \pi_t(s)$:

$$\nu_t^\pi(s) \;=\; \sum_{s'} p_t(s' \mid s, a)\; \rho_t(s, a, s') * \nu_{t+1}^\pi(s') \;\coloneqq\; (\mathcal{T}_t^\pi \nu_{t+1}^\pi)(s).$$

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ be bounded measurable. Additivity, then conditioning on S_{t+1}:

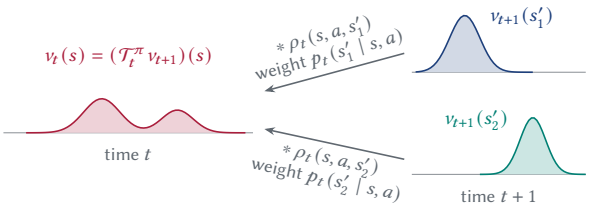
$$\mathbb{E}\big[g(W_{t:H}) \,\big|\, S_t = s\big] = \sum_{s'} p_t(s' \mid s, a)\; \mathbb{E}\big[g(R_{t+1} + W_{t+1:H}) \,\big|\, S_t = s,\, S_{t+1} = s'\big].$$

Given $(S_t, A_t, S_{t+1}) = (s, a, s')$: $R_{t+1} \sim \rho_t(s, a, s')$ (protocol), $W_{t+1:H} \sim \nu_{t+1}^\pi(s')$ (Markov property), and the two are **independent**: the reward is drawn from the transition alone, the future depends on the past through S_{t+1} only (Markov policy). The law of a sum of independent variables is the convolution: the inner expectation is $\int g \, \mathrm{d}(\rho_t(s, a, s') * \nu_{t+1}^\pi(s'))$. As g is arbitrary, the laws coincide. ∎

Backward induction on laws: $\nu_H^\pi = \delta_0$, $\nu_0^\pi = \mathcal{T}_0^\pi \mathcal{T}_1^\pi \cdots \mathcal{T}_{H-1}^\pi \delta_0$. Deterministic rewards ($\rho_t = \delta_{r_t}$, as in both examples): the convolution is a plain **shift**, $\delta_r * \nu = \mathcal{L}(r + X)$, $X \sim \nu$; Part 3 is phrased with shifts.

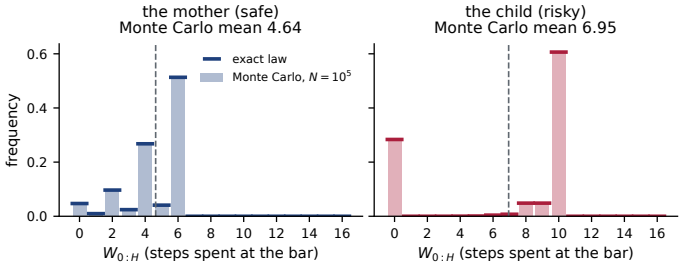
The distributional Bellman operator: a mixture of convolutions

- $\mathcal{T}_t^\pi$: **convolve** each $\nu(s')$ with a known reward law, **mix** with the transition probabilities.
- Affine under mixtures; monotone for the stochastic order; **non-expansive**: $\overline{\mathcal{W}}_1(\mathcal{T}_t^\pi \nu, \mathcal{T}_t^\pi \nu') \le \overline{\mathcal{W}}_1(\nu, \nu')$ for $\overline{\mathcal{W}}_1(\nu, \nu') \coloneqq \max_s \mathcal{W}_1(\nu(s), \nu'(s))$, where $\mathcal{W}_1(\nu, \nu') = \inf \mathbb{E}|X - Y|$ over couplings $X \sim \nu$, $Y \sim \nu'$.
- **The classical equation is its image under the mean**: $\mathbb{E}[\rho * \nu] = \mathbb{E}[\rho] + \mathbb{E}[\nu]$ and $\mathbb{E}[\sum_i w_i \nu_i] = \sum_i w_i \mathbb{E}[\nu_i]$. Part 3 asks which finite families of statistics are closed under these two operations.



Exact computation: finitely many reward values $\Rightarrow$ $\nu_t^\pi(s)$ finitely supported; deterministic integer rewards: laws on $\{-HR_{\max}, \dots, HR_{\max}\}$, $O(H^2|\mathcal{S}|^2 R_{\max})$ operations.

The windy hike: the exact law against Monte Carlo

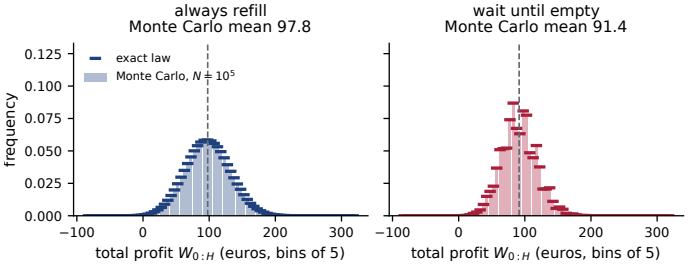


exact law	mother	child
mean $V_0^\pi(\mathrm{S})$	4.632	**6.972**
standard deviation	1.717	4.428
$\mathbb{P}(W_{0:H} = 0)$	0.047	**0.284**
$\mathbb{P}(W_{0:H} = W_{\max})$	0.513	0.606

$W_{\max} = 6 / 10;\ 0.92^8 = 0.513,$
$0.92^6 = 0.606$

Sixteen applications of $\mathcal{T}_t^\pi$ to laws on $\{0, \dots, 16\}$: the exact law (dark marks) against the Monte Carlo frequencies (bars). The differences are the Monte Carlo errors, of order 10^{-3}.

The bike store: the exact law against Monte Carlo

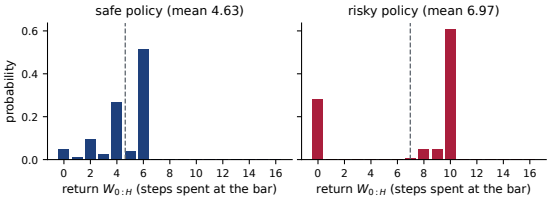


exact law	**refill**	**wait**
mean $V_0^\pi(15)$	**97.86**	91.17
standard deviation	33.61	26.24
10% quantile	54.8	58.5
$\mathrm{CVaR}_{10\%}$	39.9	**47.2**

euros

All amounts are multiples of 0.1: the laws live on a lattice of 5 561 points, and twenty applications of $\mathcal{T}_t^\pi$ (16 states, 16 demand values) give them exactly. Dynamic programming computes the whole law in the same H sweeps as the mean; the overhead is the support size.

What the mean does not say



The child has the larger mean, and more than one chance in four of no bar at all.

The manager who always refills earns 97.9 on average against 91.2, but the worst tenth of his seasons average 39.9 against 47.2. A bank, an insurer, a regulator do not look at the mean only.

- **Quantiles**: value at risk $\mathrm{VaR}_\delta(\nu) = F_\nu^{-1}(\delta) = \inf\{x : F_\nu(x) \geq \delta\}$, $0 < \delta < 1$.
- **Conditional value at risk** $\mathrm{CVaR}_\delta(\nu) = \frac{1}{\delta}\int_0^\delta F_\nu^{-1}(u)\,\mathrm{d}u$: mean of the worst fraction δ.
- **Mean–variance** $\int x\,\mathrm{d}\nu - \tau\mathbb{V}\mathrm{ar}[\nu]$, $\tau > 0$ (Markowitz 1952).
- **β-means** $\Psi_\beta(\nu) = \frac{1}{\beta}\ln\int e^{\beta x}\,\mathrm{d}\nu$ (entropic risk measures), Ψ_0 = the mean.
- **Utilities** $\Psi_f(\nu) = \int f\,\mathrm{d}\nu$, f continuous nondecreasing.

Which functional Ψ of the law should guide a decision, and can dynamic programming optimize it?

Criteria: functionals of the law of the return

Criterion: $\Psi : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R}$, defined on laws with bounded support; the agent prefers π to π' from s when $\Psi(\nu_0^{\pi}(s)) > \Psi(\nu_0^{\pi'}(s))$. $\Psi(\delta_x)$ is the value of a sure amount x: $\Psi_\beta(\delta_x) = x$, $\Psi_f(\delta_x) = f(x)$.

Definition (Standing assumptions).

1. **Monotone**: $\nu \leq_{\text{st}} \nu'$ (stochastic order: $F_\nu \geq F_{\nu'}$) $\Rightarrow \Psi(\nu) \leq \Psi(\nu')$: a sure improvement is never disliked.
2. **Continuous** for $\mathcal{W}_1$ on laws supported in a compact interval.

Definition (Two axioms of decision theory, *not* assumed here). For all finitely supported ν_1, ν_2, ν_3, $w \in [0,1]$ and $c \in \mathbb{R}$:

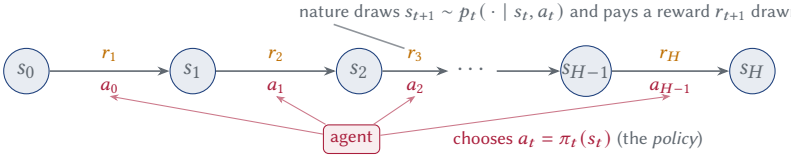
1. **Independence** (von Neumann–Morgenstern):
   $\Psi(\nu_1) \geq \Psi(\nu_2) \implies \Psi\big(w\nu_1 + (1-w)\nu_3\big) \geq \Psi\big(w\nu_2 + (1-w)\nu_3\big)$:
   a comparison survives a mixture with a common third law.
2. **Translation** (the ordinal form of the cash invariance of risk measures):
   $\Psi(\nu_1) \geq \Psi(\nu_2) \implies \Psi(\delta_c * \nu_1) \geq \Psi(\delta_c * \nu_2)$: a comparison survives a common shift.

Which criterion has which property?

	monotone	continuous	independent	translation	where it fails, and why
mean, β-means Ψ_β	✓	✓	✓	✓	affine under mixtures (in the $e^{\beta\cdot}$ scale); $\Psi_\beta(\delta_c * \nu) = c + \Psi_\beta(\nu)$
utilities Ψ_f	✓	✓	✓	×	(f continuous nondecreasing) translation only when f is affine or exponential (Part 2)
CVaR_δ, $0 < \delta < 1$	✓	✓	×	✓	$\mathrm{CVaR}_{1/2}$: $\delta_1 \succ \frac{1}{2}(\delta_0 + \delta_4)$ ($1 > 0$), but half-mixed with δ_4: $1 < 2$
quantiles VaR_δ	✓	×	×	✓	the median of $(\frac{1}{2}+\varepsilon)\delta_0 + (\frac{1}{2}-\varepsilon)\delta_1$ jumps from 0 to 1 as ε changes sign; independence: Part 2
mean–variance	×	✓	×	✓	$\int x\,\mathrm{d}\nu - \tau\mathbb{V}\mathrm{ar}[\nu]$: $\frac{1}{2}(\delta_0 + \delta_M) \geq_{\mathrm{st}} \delta_0$, yet disliked for $M > 2/\tau$; the variance of a mixture has a cross term

- On the hike: the mean prefers the child; minimizing $\mathbb{P}(W_{0:H} = 0)$, maximizing $\mathrm{CVaR}_{10\%}$ or Ψ_β with $\beta < \beta_\times \approx -0.253$ prefer the mother. On the store: the mean prefers the refiller, $\mathrm{CVaR}_{10\%}$ the prudent manager.
- Among these examples, only the β-means have both axioms: this is no accident (Part 2).

Distributional Reinforcement Learning: 1 hope, 1 issue



return $W_{0:H} = r_1 + r_2 + \cdots + r_H$
an additive functional of the chain

Classical RL (Bellman 1957): $V_t(s) = \mathbb{E}[W_{t:H} \mid S_t = s]$

	$t=0$	1	$\cdots$	$H-1$	H
s^1	7.0	6.3	$\cdots$	0.0	0
s^2	10.1	9.3	$\cdots$	0.0	0
s^3	16.0	15.0	$\cdots$	1.0	0
$\vdots$					

backward in time

Distributional RL (Bellemare et al. 2017): $\nu_t(s) = \mathcal{L}(W_{t:H} \mid S_t = s)$

	$t=0$	1	$\cdots$	$H-1$	H
s^1			$\cdots$		δ_0
s^2			$\cdots$		δ_0
s^3			$\cdots$		δ_0
$\vdots$					

backward in time

Question (Today, in finite horizon and without discounting).

1. Which criteria $\Psi(\text{law of } W_{0:H})$ can dynamic programming *optimize*?
2. How to deal effectively with the *distributions*?

The two questions of this talk

Question (Part 2. Dynamic programming for a criterion Ψ).
take $v_H^\star(s) = \delta_0$ and, for $t = H-1, \dots, 0$,

$$v_t^\star(s,a) = \sum_{s'} p_t(s' \mid s,a)\ \rho_t(s,a,s') * v_{t+1}^\star(s')\ ,$$

$$\pi_t^\star(s) \in \arg\max_{a \in \mathcal{A}} \Psi\big(v_t^\star(s,a)\big), \qquad v_t^\star(s) = v_t^\star\big(s, \pi_t^\star(s)\big)\ .$$

Does it return $\max_\pi \Psi(v_0^\pi(s))$ in general? For which Ψ? Evaluating is not optimizing: three moments evaluate the third-moment criterion exactly, yet its greedy maximization fails.

Question (Part 3. Representing the laws). $v_t^\pi(s)$ is an infinite-dimensional object, or a law whose number of atoms grows with H:

- with a finite budget, which statistics to propagate?
- What is the cost of approximation?

Part 2

What can dynamic programming optimize?

1. Markov decision processes and the law of the return
2. What can dynamic programming optimize?
3. How to represent the law of the return?

Greedy distributional dynamic programming for a criterion Ψ

The would-be dynamic programming of Part 1, recalled:

$$v_H^\star(s) = \delta_0; \qquad v_t^\star(s,a) = \sum_{s'} p_t(s' \mid s,a)\ \rho_t(s,a,s') * v_{t+1}^\star(s'),$$

$$\pi_t^\star(s) \in \arg\max_a \Psi\big(v_t^\star(s,a)\big), \qquad v_t^\star(s) = v_t^\star(s, \pi_t^\star(s)).$$

The output $\pi^\star$ is Markov, and $v_t^\star(s) = v_t^{\pi^\star}(s)$ is its return law.

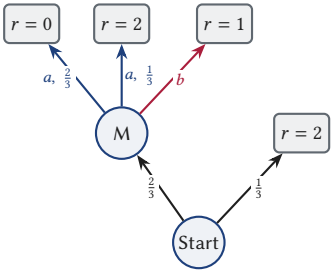
Definition (Bellman-optimizable criterion). [Marthe, G., Vernade 2023] Ψ is *Bellman optimizable* if, in every finite-horizon MDP and for every choice of the maximizers,

$$\Psi\big(v_0^\star(s)\big) \;=\; \max_\pi \Psi\big(v_0^\pi(s)\big) \qquad \text{for every } s \in \mathcal{S},$$

the maximum being over all policies, history-dependent ones included.

- The mean is Bellman optimizable: Part 1. So is every $G \circ \Psi$ with G strictly increasing (same maximizers).
- Which others? Two counterexamples first.

Counterexample 1: the median. Greedy planning misses the optimal policy



$H = 2$; unlabeled rewards are 0; boxes are absorbing, their reward paid on entry.

$\Psi(\nu) = F_\nu^{-1}(\frac{1}{2})$, the (lower) median. One decision only, at M (time 1).

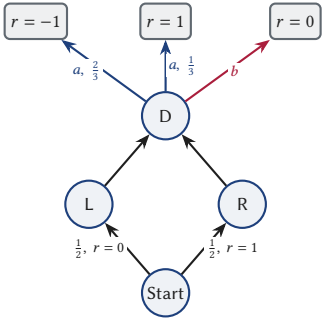
- **Greedy at M**: $\Psi(\nu_1^\star(\mathsf{M}, a)) = \mathrm{med}(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2) = 0$ and $\Psi(\nu_1^\star(\mathsf{M}, b)) = \mathrm{med}(\delta_1) = 1$: it selects b, and outputs $\nu_0^\star(\text{Start}) = \frac{2}{3}\delta_1 + \frac{1}{3}\delta_2$, of median 1.
- **The policy "a at M"**: $\nu_0^\pi(\text{Start}) = \frac{2}{3}(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2) + \frac{1}{3}\delta_2 = \frac{4}{9}\delta_0 + \frac{5}{9}\delta_2$, of median 2.

The optimal policy is Markov and deterministic, and greedy planning does not find it.

Why: $\Psi(\delta_1) \geq \Psi(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2)$, but mixing both with the *same* third law δ_2 (weight $\frac{1}{3}$) reverses the order. The comparison made at M is not the comparison that matters at Start.

Needed: a comparison of laws that survives a mixture with a common third law. This is the **independence** axiom.

Counterexample 2: the third moment. The optimal policy needs memory



$H = 3$; one decision, at D (time 2); boxes absorbing, reward on entry.

$\Psi(\nu) = \int x^3 \, \mathrm{d}\nu$: a utility, monotone and continuous. $W_{0:H} = R_1 + X$ with $R_1 \in \{0, 1\}$ fair, $X \sim \frac{2}{3}\delta_{-1} + \frac{1}{3}\delta_1$ under a, $X = 0$ under b.

- **Greedy at D**: $\Psi(a) = -\frac{2}{3} + \frac{1}{3} = -\frac{1}{3} < 0 = \Psi(b)$: selects b. From Start: $\mathbb{E}[R_1^3] = \frac{1}{2}$.
- **Markov policy "always a"**: $\frac{1}{2}\mathbb{E}[X^3] + \frac{1}{2}\mathbb{E}[(1+X)^3] = \frac{1}{2}(-\frac{1}{3}) + \frac{1}{2} \cdot \frac{8}{3} = \frac{7}{6} > \frac{1}{2}$.
- **History-dependent "a if $R_1 = 1$, b if $R_1 = 0$"**: $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{8}{3} = \frac{4}{3} > \frac{7}{6}$.

No Markov policy is optimal: the best action at D depends on the reward already collected, because the third moment of a sum is not determined by the third moments of its parts.

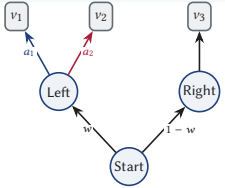
Why: $\Psi(\delta_0) \geq \Psi(\frac{2}{3}\delta_{-1} + \frac{1}{3}\delta_1)$ $(0 \geq -\frac{1}{3})$, but after a common shift by 1: $\Psi(\delta_1) = 1 < \frac{8}{3} = \Psi(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2)$.

Needed: a comparison of laws that survives a common shift. This is the **translation** property.

Two necessary properties

Definition. Ψ has the **independence** property if, for all finitely supported ν_1, ν_2, ν_3 and $w \in [0,1]$, $\Psi(\nu_1) \geq \Psi(\nu_2) \Rightarrow \Psi(w\nu_1 + (1-w)\nu_3) \geq \Psi(w\nu_2 + (1-w)\nu_3)$; the **translation** property if $\Psi(\nu_1) \geq \Psi(\nu_2) \Rightarrow \Psi(\delta_c * \nu_1) \geq \Psi(\delta_c * \nu_2)$ for all $c \in \mathbb{R}$.

Lemma (Necessary conditions). [Marthe, G., Vernade 2023] A Bellman-optimizable criterion has both properties.



Proof. The two counterexamples, in general form: a gray box ν is a *fan*, one action leading to absorbing states with the probabilities and rewards of the atoms of ν. *Independence*: if $\Psi(\nu_1) \geq \Psi(\nu_2)$ but $\Psi(w\nu_1 + (1-w)\nu_3) < \Psi(w\nu_2 + (1-w)\nu_3)$, greedy planning may select a_1 at Left and output $w\nu_1 + (1-w)\nu_3$ at Start, whereas a_2 does strictly better. *Translation*: same with $\delta_c * \nu_1$ against $\delta_c * \nu_2$. ∎

Applied in both directions, independence gives *mixture compatibility* ($\Psi(\nu_1) = \Psi(\nu_2) \Rightarrow$ equality after mixing with ν_3) and translation gives *shift compatibility*: the equality-preserving forms used in Part 3.

Ingredient 1: independence reduces the criterion to a utility

Theorem (Expected-utility theorem). [von Neumann and Morgenstern 1944; Herstein and Milnor 1953] Ψ monotone, continuous, independent, not constant on Dirac masses. Then there is a continuous nondecreasing f such that

$$\Psi(\nu) \geq \Psi(\nu') \quad \Longleftrightarrow \quad \int f \,\mathrm{d}\nu \geq \int f \,\mathrm{d}\nu' \qquad \text{(laws with bounded support)}:$$

Ψ is **ordinally equivalent** to the utility Ψ_f. f is unique up to $f \mapsto \alpha f + \gamma$, $\alpha > 0$.

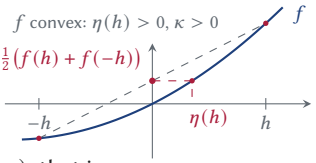
Proof sketch, on an interval $[a, b]$ with $\Psi(\delta_a) < \Psi(\delta_b)$.

1. $\nu_\lambda \coloneqq (1-\lambda)\delta_a + \lambda\delta_b$; $G(\lambda) \coloneqq \Psi(\nu_\lambda)$ is continuous and *strictly* increasing (if $G(\lambda) = G(\lambda')$, independence propagates the equality up to $\lambda = 1$: contradiction).
2. **Calibration**: $\delta_a \leq_{\mathrm{st}} \nu \leq_{\mathrm{st}} \delta_b$, so $\bar{\Psi}(\nu) \coloneqq G^{-1}(\Psi(\nu)) \in [0,1]$ is the unique λ with $\nu \sim \nu_\lambda$; $\bar{\Psi} \sim \Psi$.
3. **$\bar{\Psi}$ preserves mixtures** (independence, twice): $\kappa\nu + (1-\kappa)\nu' \sim \kappa\nu_{\bar{\Psi}(\nu)} + (1-\kappa)\nu_{\bar{\Psi}(\nu')} = \nu_{\kappa\bar{\Psi}(\nu)+(1-\kappa)\bar{\Psi}(\nu')}$.
4. $f(x) \coloneqq \bar{\Psi}(\delta_x)$: by induction on the atoms, $\bar{\Psi}(\nu) = \int f \,\mathrm{d}\nu$ on finitely supported laws; f is nondecreasing (monotonicity) and continuous; $\mathcal{W}_1$-continuity extends to bounded supports.

Ingredient 2: translation forces an ordinary differential equation

$\Psi \sim \Psi_f$ with the translation property; assume $f \in C^2$ and $f' > 0$ (continuous is sufficient: backup B3).

- **The certainty equivalent of a fair coin.** For h small, $\eta(h)$ is the sure amount equivalent to the coin $\frac{1}{2}(\delta_h + \delta_{-h})$: $f(\eta(h)) = \frac{1}{2}\big(f(h) + f(-h)\big)$ (figure). Taylor, $\frac{1}{2}(f(h) + f(-h)) = f(0) + \frac{1}{2}f''(0)h^2 + o(h^2)$ and $f(\eta) = f(0) + f'(0)\eta + o(\eta)$: $\eta(h) = \kappa h^2 + o(h^2)$, $\kappa = f''(0)/(2f'(0))$.



- **Translation.** $\delta_{\eta(h)} \sim \frac{1}{2}(\delta_h + \delta_{-h})$, hence for every x: $\delta_{x+\eta(h)} \sim \frac{1}{2}(\delta_{x+h} + \delta_{x-h})$, that is

$$f\big(x + \eta(h)\big) = \tfrac{1}{2}\big(f(x+h) + f(x-h)\big) \quad \Longrightarrow \quad f(x) + f'(x)\kappa h^2 + o(h^2) = f(x) + \tfrac{1}{2}f''(x)h^2 + o(h^2).$$

- **The ODE and its solutions**: $f'' = \beta f'$ with $\beta \coloneqq 2\kappa = f''(0)/f'(0)$, so $f'(x) = f'(0)e^{\beta x}$:

$\beta = 0$: $f(x) = c_1 x + c_0$, $c_1 > 0$ (the mean); $\qquad \beta \neq 0$: $f(x) = c_1 e^{\beta x} + c_0$, $c_1\beta > 0$ (a β-mean: $\Psi_\beta = \frac{1}{\beta}\ln \Psi_{e^{\beta\cdot}}$).

Theorem (Only the β-means). [Marthe, G., Vernade 2023] Ψ monotone, continuous, not constant on Dirac masses. Then Ψ is Bellman optimizable $\iff$ $\Psi = G \circ \Psi_\beta$ for some $\beta \in \mathbb{R}$ and some strictly increasing G ($\beta = 0$: the mean; $G(u) = \Psi(\delta_u)$).

$\Rightarrow$: the lemma and the two ingredients. $\Leftarrow$: Howard and Matheson (1972), next.

The β-means have a Bellman equation of their own

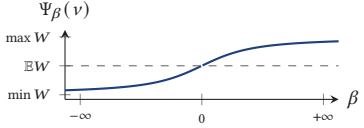
$\beta \neq 0$ ($\beta = 0$ by the limit). $Z^{\pi}_{\beta,t}(s) \coloneqq \mathbb{E}\big[e^{\beta W_{t:H}} \,\big|\, S_t = s\big]$; **$\beta$-value** $V^{\pi}_{\beta,t}(s) \coloneqq \Psi_\beta(\nu^{\pi}_t(s)) = \frac{1}{\beta}\ln Z^{\pi}_{\beta,t}(s)$; for a reward law, $\Psi_\beta(\rho) = \frac{1}{\beta}\ln\int e^{\beta r}\rho(\mathrm{d}r)$.

Theorem (Multiplicative and exponential Bellman equations). $Z_{\beta,H} \equiv 1$; *tilted kernel* $\widetilde{P}^{\pi}_{\beta,t}(s,s') \coloneqq p_t(s' \mid s, \pi_t(s)) \int e^{\beta r}\rho_t(s,\pi_t(s),s')(\mathrm{d}r)$:

$$Z^{\pi}_{\beta,t} = \widetilde{P}^{\pi}_{\beta,t} Z^{\pi}_{\beta,t+1}, \qquad Z^{\pi}_{\beta,0} = \widetilde{P}^{\pi}_{\beta,0}\widetilde{P}^{\pi}_{\beta,1}\cdots\widetilde{P}^{\pi}_{\beta,H-1}\mathbf{1} \quad \text{(Feynman–Kac)}.$$

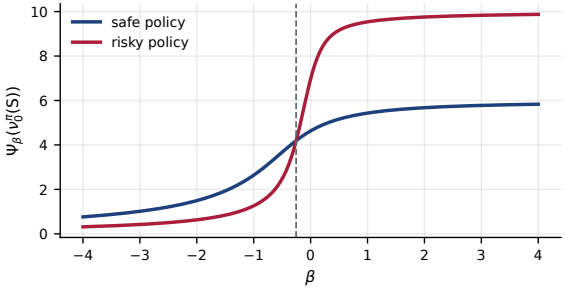
Logarithmic form, $a = \pi_t(s)$: $V^{\pi}_{\beta,t}(s) = \frac{1}{\beta}\ln\sum_{s'} p_t(s' \mid s,a)\, e^{\beta[\Psi_\beta(\rho_t(s,a,s'))+V^{\pi}_{\beta,t+1}(s')]}$.

- The Bellman equation with $\mathbb{E}$ replaced by Ψ_β and r by $\Psi_\beta(\rho)$.
- **Planning** (Howard–Matheson 1972):
  $V^{\star}_{\beta,t}(s) = \max_a \frac{1}{\beta}\ln\sum_{s'} p_t(s' \mid s,a)\, e^{\beta[\Psi_\beta(\rho_t(s,a,s'))+V^{\star}_{\beta,t+1}(s')]}$.
  The proof of Part 1 applies with $\mathbb{E}$ replaced by Ψ_β, monotone and with the tower property $\Psi_\beta\big(\Psi_\beta(W \mid \text{past})\big) = \Psi_\beta(W)$: the greedy policy is optimal among all policies.



$\beta \mapsto \Psi_\beta(\frac{1}{2}(\delta_0+\delta_1))$. Limits: the mean at $\beta = 0$; $(\max,+)$, the best attainable return, at $+\infty$; $(\min,+)$, the guaranteed return, at $-\infty$.

The two hikers, seen through β-means



- $\beta \mapsto \Psi_\beta(v_0^\pi(S))$ for the two hikers, by the exponential Bellman equation.
- Crossing at $\beta_\times \approx -0.253$: **the child** is preferred for $\beta > \beta_\times$, **the mother** for $\beta < \beta_\times$: two ranges of a single parameter.
- Chernoff, $\beta < 0$: $\mathbb{P}(W_{0:H} \leq u) \leq e^{-|\beta|(V_{\beta,0}-u)}$: a risk-averse β-value is a guarantee on the lower tail.

Among monotone continuous criteria, the increasing transforms of β-means, and only they, can be optimized by the greedy backward recursion.

Part 3

How to represent the law of the return?

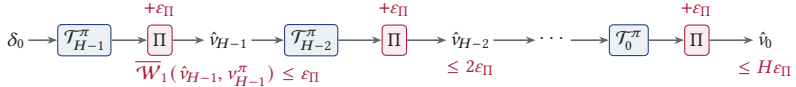
1. Markov decision processes and the law of the return

2. What can dynamic programming optimize?

3. How to represent the law of the return?

Propagating approximate laws: projected dynamic programming

- Finitely many reward values: exact, but the number of atoms may grow geometrically with the horizon; continuous rewards: no finite support at all.
- **Practice** (Bellemare et al. 2017, Dabney et al. 2018): N parameters per law (grid atoms, or quantiles); a **projection** Π after every step.



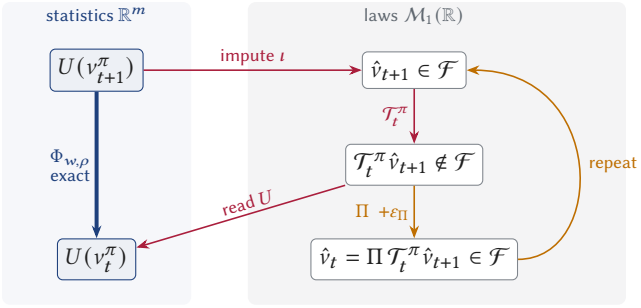
Proposition (Error accumulation). [Marthe, G., Vernade 2023] Supports in an interval of length Δ, $\mathcal{W}_1(\Pi\nu, \nu) \le \varepsilon_\Pi$:

$\overline{\mathcal{W}}_1(\hat{v}_t, v^\pi_t) \le (H-t)\,\varepsilon_\Pi \ \forall t$; $\quad$ N quantiles at levels $\frac{2i-1}{2N}$: $\varepsilon_\Pi \le \frac{\Delta}{2N}$, so at $t=0$: $\le \frac{H\Delta}{2N} \le \frac{H^2 R_{\max}}{N}$.

Proof. $\overline{\mathcal{W}}_1(\hat{v}_t, v^\pi_t) \le \overline{\mathcal{W}}_1(\Pi\mathcal{T}^\pi_t \hat{v}_{t+1}, \mathcal{T}^\pi_t \hat{v}_{t+1}) + \overline{\mathcal{W}}_1(\mathcal{T}^\pi_t \hat{v}_{t+1}, \mathcal{T}^\pi_t v^\pi_{t+1}) \le \varepsilon_\Pi + \overline{\mathcal{W}}_1(\hat{v}_{t+1}, v^\pi_{t+1})$ (non-expansive). ■

Projection error paid at every step: linear accumulation in H. Even the mean of $\hat{v}_0$ need not be preserved (quantile projection; categorical projection when it clips).

Propagate exact statistics, not approximate laws



A. Propagate approximate laws

N parameters per law, a projection Π at every step: error $O(H\Delta/N)$, accumulates.

B. Propagate exact statistics

Never leave $\mathbb{R}^m$: no projection, no imputation. Reconstruct a law *once*, where it is needed.

Which statistics?
How much do they determine?
Which law to reconstruct?

Red detour: the imputation strategy of Rowland et al. (2019). It agrees with the blue arrow for every imputation ι with $U \circ \iota = \mathrm{id}$ *iff* U is Bellman closed: then no imputation is needed. Diagram after Bringas Miranda, G., Cappé (2026).

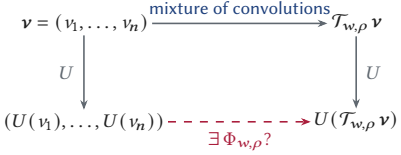
Bellman closedness

Statistic: $U : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R}^m$ (bounded supports).
Mixture of convolutions, weights w, reward laws $\rho_1, \dots, \rho_n$:

$$\mathcal{T}_{w,\rho}(\nu_1, \dots, \nu_n) \;\coloneqq\; \sum_{i=1}^{n} w_i \, \rho_i * \nu_i; \qquad (\mathcal{T}_t^{\pi} \nu)(s) \text{ is of this form.}$$

Definition (Bellman-closed statistic). [Rowland et al. 2019] $\forall n, w, \rho \;\; \exists \Phi_{w,\rho} : (\mathbb{R}^m)^n \to \mathbb{R}^m$ such that, for all $\nu_1, \dots, \nu_n$,

$$U\big(\mathcal{T}_{w,\rho}\boldsymbol{\nu}\big) = \Phi_{w,\rho}\big(U(\nu_1), \dots, U(\nu_n)\big).$$



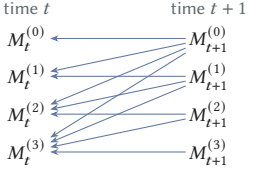
- $\implies U(\nu_t^{\pi}(s)) = \Phi_{p_t(\cdot|s,a),\rho_t(s,a,\cdot)}\big((U(\nu_{t+1}^{\pi}(s')))_{s'}\big)$: **its own Bellman equation**, in every MDP.
- Examples: mean; first m moments; the β-means (Part 2); informally the whole law ($m = \infty$). Only the *information* matters: $g \circ U$ closed for every bijection g.

Moments

Proposition (Moment recursions). $M_t^{(j)}(s) \coloneqq \int x^j \, \nu_t^\pi(s)(\mathrm{d}x)$, $m_j(\rho) \coloneqq \int r^j \rho(\mathrm{d}r)$; $M^{(0)} \equiv 1$, $M^{(1)} = V^\pi$, $M_H^{(j)} \equiv 0$ for $j \geq 1$, and, with $a = \pi_t(s)$,

$$M_t^{(j)}(s) \;=\; \sum_{s'} p_t(s' \mid s, a) \sum_{i=0}^{j} \binom{j}{i} m_{j-i}\big(\rho_t(s,a,s')\big)\, M_{t+1}^{(i)}(s').$$

Proof. Convolution: $\mathbb{E}[(R+X)^j] = \sum_i \binom{j}{i} \mathbb{E}[R^{j-i}]\mathbb{E}[X^i]$ for independent R, X (binomial formula). Mixture: average. ∎



Triangular: first m moments together, exactly, $O(H|\mathcal{S}|^2 m^2)$.

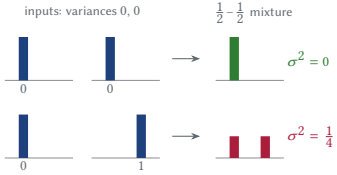
Variance $\sigma_t^2 = M_t^{(2)} - V_t^2$, not autonomous (it needs V); total variance, R and W' independent given s':

$$\sigma_t^2(s) = \underbrace{\textstyle\sum_{s'} p_t(s' \mid s, a)\, \big[\sigma_{t+1}^2(s') + \mathbb{V}\mathrm{ar}\rho_t(s,a,s')\big]}_{\text{inherited, and created by the reward}} + \underbrace{\textstyle\sum_{s'} p_t(s' \mid s, a)\, \big[r_t(s,a,s') + V_{t+1}(s') - V_t(s)\big]^2}_{\text{created by the transition at time } t}$$

A test for closedness, and what fails

Proposition (Closed = compatible with shifts and mixtures). If U is continuous for $\mathcal{W}_1$ on laws with a common compact support, it is Bellman closed iff
(i) $U(\nu) = U(\nu') \Rightarrow U(\delta_r * \nu) = U(\delta_r * \nu')\ \forall r$, and
(ii) $U(\nu_i) = U(\nu_i')\ \forall i \Rightarrow U(\sum_i w_i \nu_i) = U(\sum_i w_i \nu_i')\ \forall w$.

Proof. $\Rightarrow$: $n = 1$, $\rho_1 = \delta_r$; then all $\rho_i = \delta_0$. $\Leftarrow$: $\Phi_{w,\rho}(y_1, \dots, y_n) \coloneqq U(\sum_i w_i \rho_i * \nu_i)$ for any ν_i with $U(\nu_i) = y_i$: well defined by (i)+(ii) for finitely supported ρ_i (a convolution is then a finite mixture of shifts); in general approximate ρ_i in $\mathcal{W}_1$ and use continuity. ■



Variance alone: same inputs, different outputs.

With the mean, it closes.

statistic	shifts	mixtures	why
mean; (mean, variance); first m moments; β-means	✓	✓	binomial formula, linearity; $e^{\beta(r+x)} = e^{\beta r} e^{\beta x}$
variance alone	✓	×	δ_0, δ_1: variance 0; $\frac{1}{2}(\delta_0 + \delta_0)$ vs $\frac{1}{2}(\delta_0 + \delta_1)$: 0 vs $\frac{1}{4}$
median, any quantile	✓	×	δ_0 vs $0.6\delta_0 + 0.4\delta_5$: median 0; half-mixed with δ_1: 0 vs 1
threshold probability $\nu([u, \infty))$	×	✓	linear in ν; $\delta_u, \delta_{u+1} \mapsto 1$, shifted by -1: 0 vs 1
entropy (discrete)	✓	×	mix δ_0 with δ_0, or with δ_1
ess inf, ess sup	✓	✓	$(\min, +)$, $(\max, +)$ recursions: the endpoints of the support are computable (not its gaps); closed but not continuous

Utilities: a functional equation

Utility: $\Psi_f(\nu) = \int f \, \mathrm{d}\nu$, f continuous; the statistic $U = \Psi_{\mathbf{f}} = \left(\int f_1 \, \mathrm{d}\nu, \ldots, \int f_m \, \mathrm{d}\nu \right)$.
Linear in $\nu \Rightarrow$ mixture compatible: closed $\iff$ shift compatible.

Lemma (Shift compatibility of utilities). $\Psi_{\mathbf{f}}$ shift compatible $\iff \forall c \in \mathbb{R} \, \exists A(c) \in \mathbb{R}^{m \times m}$, $\gamma(c) \in \mathbb{R}^m$:

$$\mathbf{f}(x + c) \;=\; A(c)\,\mathbf{f}(x) + \gamma(c) \qquad \forall x \in \mathbb{R}$$

$\iff$ $\mathrm{span}\{1, f_1, \ldots, f_m\}$ translation invariant.

Proof. $\Leftarrow$: integrate: $\Psi_{\mathbf{f}}(\delta_c * \nu) = A(c)\Psi_{\mathbf{f}}(\nu) + \gamma(c)$.
$\Rightarrow$: E = finitely supported signed measures; $\Lambda_c \mu \coloneqq \int \mathbf{f}(x + c) \mu(\mathrm{d}x)$, $\widetilde{\Lambda}_0 \mu \coloneqq (\mu(\mathbb{R}), \Lambda_0 \mu)$. $\widetilde{\Lambda}_0 \mu = 0 \Rightarrow$ $\mu = \kappa(\nu - \nu')$, $\Psi_{\mathbf{f}}(\nu) = \Psi_{\mathbf{f}}(\nu') \Rightarrow$ (compatibility) $\Lambda_c \mu = 0$. So $\ker \widetilde{\Lambda}_0 \subset \ker \Lambda_c$: $\Lambda_c \mu = \mu(\mathbb{R})\gamma(c) + A(c)\Lambda_0 \mu$ (linear algebra). Take $\mu = \delta_x$. ∎

For a general reward law: $\int \mathbf{f} \, \mathrm{d}(\rho * \nu) = \iint \mathbf{f}(x + r) \, \nu(\mathrm{d}x) \rho(\mathrm{d}r) = \left(\int A \, \mathrm{d}\rho \right) \Psi_{\mathbf{f}}(\nu) + \int \gamma \, \mathrm{d}\rho$: the same linear recursion, with averaged coefficients.

Bellman-closed utilities: the exponential polynomials, and nothing else

Theorem (Bellman-closed families). [Rowland et al. 2019; Marthe, G., Vernade 2023] $\Psi_{\mathbf{f}}$ Bellman closed $\iff$ $V = \mathrm{span}\{1, f_1, \ldots, f_m\}$ translation invariant $\Rightarrow$ every f_i is an *exponential polynomial*: a combination of $x^j e^{ux}\cos(vx)$, $x^j e^{ux}\sin(vx)$. Conversely, any basis of a finite-dimensional translation-invariant space of continuous functions is a closed family.

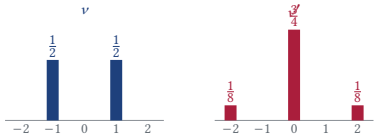
Proof sketch. V finite-dimensional, translation invariant, continuous functions. (1) $f_i(x+c) = \sum_j c_{ij}(c) f_j(x)$ with c_{ij} continuous in c (invert an evaluation matrix). (2) V closed under pointwise limits $\Rightarrow V \subset C^\infty$ (mollify), $f' = \lim_h \frac{f(\cdot+h)-f}{h} \in V$. (3) $D = \frac{\mathrm{d}}{\mathrm{d}x}$ endomorphism of V; Cayley–Hamilton: $\chi_D(D) = 0 \Rightarrow$ linear ODE with constant coefficients. ∎

Corollary (One utility). [Marthe, G., Vernade 2023] f continuous, nonconstant: Ψ_f Bellman closed $\iff$ $f(x) = c_1 x + c_0$ or $f(x) = c_1 e^{\beta x} + c_0$: the mean or an exponential moment (the ODE of Part 2; proof without smoothness: backup slide).

$\{x^j\}_{j\le m}$: moments. $\{x^j e^{\beta x}\}_{j\le m}$: ∂_β^j of the exponential moment. $\{\cos\omega x, \sin\omega x\}$: shift = rotation. No other family of utilities.

The truncated moment problem

- **Strategy B** for a non-closed functional (quantile, CVaR): m exact moments + the exact support endpoints ($(\min, +)$ and $(\max, +)$ recursions; for lattice rewards, the lattice), then **reconstruct** a law.
- **Moment fiber**
  $C(c) \coloneqq \{\nu \in \mathcal{M}_1([a,b]) : \int x^j \,\mathrm{d}\nu = c_j,\ j = 1, \dots, m\}$:
  convex, weakly compact (Akhiezer 1965).
  All moments determine ν (Weierstrass); m moments generally do not.

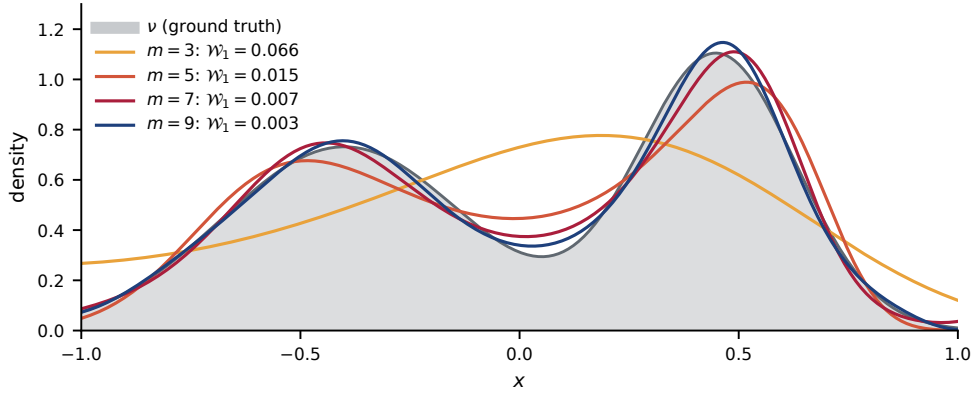


same mean 0, same variance 1,
yet $\mathcal{W}_1(\nu, \nu') = 1$

Question. How far apart can two laws of $C(c)$ be? In which metric?

- **Wasserstein**: $\mathcal{W}_1(\nu, \nu') = \inf_{X\sim\nu,\, Y\sim\nu'} \mathbb{E}|X - Y| = \sup_{\varphi\ 1\text{-Lip}} \int \varphi \,\mathrm{d}(\nu - \nu') = \int_{\mathbb{R}} |F_\nu - F_{\nu'}|$
  (Kantorovich–Rubinstein): controls Lipschitz utilities, distorted means $\int_0^1 w(u) F_\nu^{-1}(u)\,\mathrm{d}u$ (CVaR).
- Tool: polynomials of degree m against Lipschitz functions (approximation theory).

Matching m moments: every member of the fiber gets close



- For each m, one member of $C(c)$ with the first m moments of ν: the density $e^{P(x)}$ with $\deg P = m$ and these moments (**maximum entropy**, later). Not the $\mathcal{W}_1$-closest member, just one of them.
- $\mathcal{W}_1(\nu, \nu_m)$ decreases as m grows, and so does the distance between any two members: the fiber shrinks. How fast? Next.

Jackson's theorem

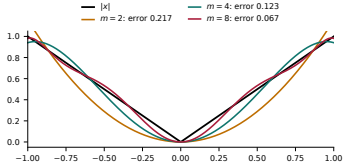
Theorem (Jackson, 1911). [Rivlin 1969, Chap. 1; Cheney 1966, Chap. 4] $\varphi : [-1,1] \to \mathbb{R}$ continuous, $m \geq 1$: there is a polynomial P of degree $\leq m$ with

$$\sup_{|x|\leq 1} |\varphi(x) - P(x)| \;\leq\; 6\,\omega_\varphi(1/m), \qquad \omega_\varphi(\delta) = \sup\{|\varphi(x) - \varphi(y)| : |x-y| \leq \delta\}.$$

1-Lipschitz: within $6/m$. $\varphi \in C^k$, $m \geq 2k$:

$$E_m(\varphi) \;\coloneqq\; \inf_{\deg P \leq m} \|\varphi - P\|_\infty \;\leq\; \left(\frac{\pi}{2}\right)^k \frac{\|\varphi^{(k)}\|_\infty}{(m+1)m\cdots(m-k+2)} \;\leq\; \left(\frac{\pi}{m}\right)^k \|\varphi^{(k)}\|_\infty.$$

- Rate $1/m$ sharp for Lipschitz functions: $\varphi(x) = |x|$.



Chebyshev interpolants of $|x|$ (near-best).

What m moments determine: an agnostic bound

Theorem (Moment matching in Wasserstein distance). [Bringas Miranda, G., Cappé 2026] $\nu, \hat{\nu}$ supported in a common interval of length Δ, same moments of orders $1, \dots, m$:

$$\mathcal{W}_1(\nu, \hat{\nu}) \;\leq\; \frac{6\Delta}{m}.$$

Proof. Affine map onto $[-1, 1]$ (moments still matched, $\mathcal{W}_1$ scaled by $2/\Delta$): prove $12/m$ there. φ 1-Lipschitz, P Jackson's polynomial, $\int P \,\mathrm{d}\nu = \int P \,\mathrm{d}\hat{\nu}$:

$$\int \varphi \;\mathrm{d}(\nu - \hat{\nu}) = \int (\varphi - P)\;\mathrm{d}(\nu - \hat{\nu}) \leq 2 \sup_{|x| \leq 1} |\varphi - P| \leq \frac{12}{m}.$$

■

- **Agnostic**: nothing on the shapes of ν, $\hat{\nu}$.
- **Horizon only through** Δ (exact support length, $\leq 2HR_{\max}$) vs $H\Delta/(2N)$ for projected DP: no accumulation of errors.
- **For the criteria of Part 1**: an L-Lipschitz utility is off by at most $6L\Delta/m$, CVaR_δ by at most $6\Delta/(m\delta)$ (it is $\frac{1}{\delta}$-Lipschitz for $\mathcal{W}_1$); quantiles are not controlled (discontinuous).

The rate $1/m$ cannot be improved

Proposition (A lower bound in $\mathcal{W}_1$). For every $m \geq 1$ and $\Delta > 0$ there are two laws supported in an interval of length Δ, with the same moments of orders $1, \dots, m$, at distance

$$\mathcal{W}_1(\nu, \nu') \;=\; \frac{\Delta}{m+1}.$$

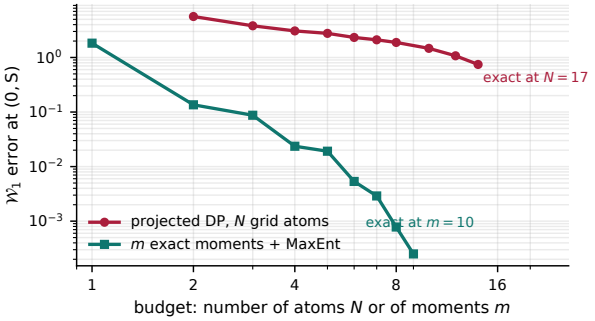
Proof. Let $n = m+1$, $x_k = A + k\Delta/n$ for $k = 0, \dots, n$, and split the binomial weights by parity:

$$\nu = 2^{-(n-1)} \sum_{k \text{ even}} \binom{n}{k} \delta_{x_k}, \qquad \nu' = 2^{-(n-1)} \sum_{k \text{ odd}} \binom{n}{k} \delta_{x_k} \qquad \text{(each of total mass 1).}$$

Moments: $\int P \,\mathrm{d}\nu - \int P \,\mathrm{d}\nu' = 2^{-(n-1)} \sum_{k=0}^{n} (-1)^k \binom{n}{k} P(x_k) = 0$ for $\deg P \leq m$: the n-th finite difference of a polynomial of degree $< n$ vanishes. *Distance*: on $[x_k, x_{k+1})$,
$F_\nu - F_{\nu'} = 2^{-(n-1)} \sum_{i \leq k} (-1)^i \binom{n}{i} = 2^{-(n-1)} (-1)^k \binom{n-1}{k}$, so
$\mathcal{W}_1(\nu, \nu') = \int |F_\nu - F_{\nu'}| \,\mathrm{d}x = \frac{\Delta}{n} 2^{-(n-1)} \sum_{k=0}^{n-1} \binom{n-1}{k} = \frac{\Delta}{n}$. ∎

$m = 1$: $\frac{1}{2}(\delta_A + \delta_{A+\Delta})$ against $\delta_{A+\Delta/2}$, same mean, $\mathcal{W}_1 = \Delta/2$. (Gauss–Legendre quadrature laws against the uniform law give a lower bound of the same order.) Jackson's rate $1/m$ is the truth, up to the constant. To do better, something more than the moments must be known: **regularity** (next).

Same budget, two strategies, on the hike



- The child's law $\nu_0^\pi(\mathrm{S})$.
- **Projected DP**: N atoms on a grid of the a priori support $[0, H]$, projection at each of the 16 steps.
- **Moment matching**: m exact moments, maximum-entropy law on the exact support $\{0, \dots, 10\}$ (next).
- Bound $6\Delta/m$, $\Delta = 10$: 30, 15, 10, 7.5 for $m = 2, 4, 6, 8$; actual 0.14, 0.024, 0.005, 0.0008: far below the worst-case bound.

Budget = size of the representation (N atoms, m moments), not equal computing time; the two methods use different support information.

Which law to pick among those with the prescribed moments, and can smoothness beat $6\Delta/m$?

The maximum-entropy law and its dual

Reference λ on $[a, b]$ (Lebesgue or counting measure); $T = (T_1, \dots, T_m)$ polynomials, $\operatorname{span}\{1, T\}$ = degree $\le m$; known: $c = \int T \, \mathrm{d}\nu$, the moments of the unknown ν, hence the fiber $C(c) = \{\nu' : \int T \, \mathrm{d}\nu' = c\} \ni \nu$.

Maximum-entropy problem: on the fiber, take the law closest to the reference, $\mathrm{KL}(\nu' \parallel \lambda) = \int \ln \frac{\mathrm{d}\nu'}{\mathrm{d}\lambda} \, \mathrm{d}\nu'$ ($+\infty$ unless $\nu' \ll \lambda$; $= -\mathrm{H}(\nu')$ for λ uniform: least relative entropy = most entropy):

$$\hat{\nu}_m \;\coloneqq\; \operatorname*{arg\,min}_{\nu' \in C(c)} \mathrm{KL}\big(\nu' \,\big\|\, \lambda\big) \,.$$

Exponential family: $\nu_\theta = e^{P_\theta} \lambda$, $P_\theta = \langle \theta, T \rangle - \ln Z(\theta)$, $Z(\theta) = \int e^{\langle \theta, T \rangle} \, \mathrm{d}\lambda$: log-density of degree $\le m$.

Proposition (Solution by the dual). [Jaynes 1957; Csiszár 1975] $J(\theta) \coloneqq \ln Z(\theta) - \langle \theta, c \rangle$ is convex on $\mathbb{R}^m$, $\nabla J = \mathbb{E}_{\nu_\theta}[T] - c$, $\nabla^2 J = \mathbb{C}\mathrm{ov}_{\nu_\theta}[T]$. If $\nabla J(\theta^\star) = 0$, then $\nu_{\theta^\star} \in C(c)$ and $\hat{\nu}_m = \nu_{\theta^\star}$: the problem on the fiber is solved by the unconstrained convex problem $\min_{\mathbb{R}^m} J$ (Newton: backup B4).

Proof. *Derivatives:* $\ln Z$ is the cumulant generating function of T: $\partial_j \ln Z = \mathbb{E}_{\nu_\theta}[T_j]$, $\partial_{jk} \ln Z = \mathbb{C}\mathrm{ov}_{\nu_\theta}(T_j, T_k)$, so $u^\top \nabla^2 J \, u = \mathbb{V}\mathrm{ar}_{\nu_\theta} \langle u, T \rangle \ge 0$. *Entropy:* $P_{\theta^\star}$ is affine in T, so $\int P_{\theta^\star} \, \mathrm{d}\nu' = \langle \theta^\star, c \rangle - \ln Z(\theta^\star)$ depends on $\nu' \in C(c)$ only through c: it equals $\int P_{\theta^\star} \, \mathrm{d}\hat{\nu}_m = \mathrm{KL}(\hat{\nu}_m \parallel \lambda)$. For $\nu' = q\lambda \in C(c)$:

$$\mathrm{KL}\big(\nu' \,\big\|\, \hat{\nu}_m\big) = \int \ln q \, \mathrm{d}\nu' - \int P_{\theta^\star} \, \mathrm{d}\nu' = \mathrm{KL}\big(\nu' \,\big\|\, \lambda\big) - \mathrm{KL}(\hat{\nu}_m \parallel \lambda) \;\ge 0, \quad \text{equality iff } \nu' = \hat{\nu}_m.$$

■

A Pythagorean identity

Recall: $\nu_\theta = e^{P_\theta}\lambda$ with P_θ a polynomial of degree $\leq m$; $\hat{\nu}_m = \nu_{\theta^\star}$ is the member with the m prescribed moments.

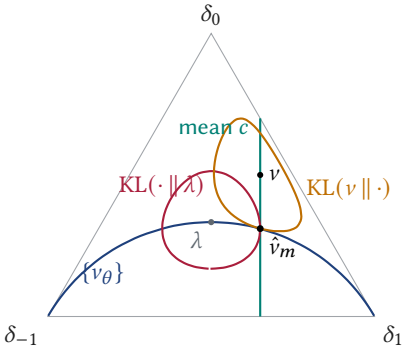
Lemma (The dual objective is a relative entropy). $\nu \ll \lambda$ with moments c, any θ:

$$\mathrm{KL}(\nu \,\|\, \nu_\theta) \;=\; \mathrm{KL}(\nu \,\|\, \lambda) + J(\theta).$$

$\Rightarrow$ $\hat{\nu}_m$ is the member of the family closest to *every* ν with moments c, and

$$\mathrm{KL}(\nu \,\|\, \nu_\theta) \;=\; \mathrm{KL}(\nu \,\|\, \hat{\nu}_m) + \mathrm{KL}(\hat{\nu}_m \,\|\, \nu_\theta)\,.$$

Proof. $\mathrm{KL}(\nu \,\|\, \nu_\theta) = \int \ln \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\,\mathrm{d}\nu - \int \ln p_\theta \,\mathrm{d}\nu =$
$\mathrm{KL}(\nu \,\|\, \lambda) - \langle \theta, c\rangle + \ln Z(\theta)$; then
$J(\theta) - J(\theta^\star) = \mathrm{KL}(\nu \,\|\, \nu_\theta) - \mathrm{KL}(\nu \,\|\, \hat{\nu}_m)$; take $\nu = \hat{\nu}_m$. ∎
Two orthogonalities: closest to λ on the fiber, closest to ν in the family: the error is controlled without knowing ν.



Laws on $\{-1, 0, 1\}$, $m = 1$ (the mean), λ uniform.

Quality of the reconstruction

Theorem (Maximum entropy versus the truth). [Barron and Sheu 1991; Bringas Miranda, G., Cappé 2026] $\nu = e^g\lambda$, g bounded on $\operatorname{supp}\lambda$; $\hat{\nu}_m$ the max-entropy law with the first m moments of ν; $E_m(g) \coloneqq \inf_{\deg P \le m} \sup_{\operatorname{supp}\lambda} |g - P|$. Then

$$\mathrm{KL}(\nu \,\|\, \hat{\nu}_m) \;\le\; \tfrac{1}{2}\, E_m(g)^2, \qquad \mathcal{W}_1(\nu, \hat{\nu}_m) \le \tfrac{\Delta}{2} E_m(g).$$

$g \in C^k$ on an interval of length Δ, $|g^{(k)}| \le M_k$, $m \ge 2k$: $E_m(g) \le (\pi\Delta/2m)^k M_k$: rates m^{-2k} (relative entropy), m^{-k} ($\mathcal{W}_1$).

Proof. P polynomial of degree $\le m$: $\nu_P \coloneqq e^P\lambda / \int e^P \,\mathrm{d}\lambda$ is in the family $\Rightarrow \mathrm{KL}(\nu \,\|\, \hat{\nu}_m) \le \mathrm{KL}(\nu \,\|\, \nu_P)$ (Pythagoras). $r \coloneqq g - P$; $\int e^P \,\mathrm{d}\lambda = \int e^{P-g} \,\mathrm{d}\nu = \mathbb{E}_\nu[e^{-r}]$:

$$\mathrm{KL}(\nu \,\|\, \nu_P) = \int (g - P) \,\mathrm{d}\nu + \ln \int e^P \,\mathrm{d}\lambda = \mathbb{E}_\nu[r] + \ln \mathbb{E}_\nu[e^{-r}] = \ln \mathbb{E}_\nu\left[e^{-(r - \mathbb{E}_\nu r)}\right] \le \tfrac{1}{8}(\sup r - \inf r)^2$$

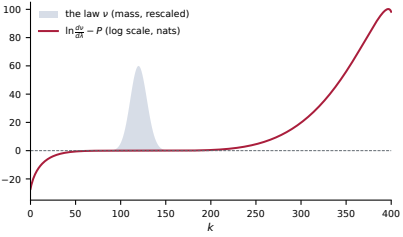
(**Hoeffding's lemma**); oscillation of $g - P = 2\inf_a \|g - P - a\|_\infty$; infimum over P. $\mathcal{W}_1$: Pinsker, $\|\nu - \mu\|_{\mathrm{TV}} \coloneqq \sup_A |\nu(A) - \mu(A)| \le \sqrt{\mathrm{KL}(\nu \,\|\, \mu)\,/2}$, and $\mathcal{W}_1 \le \Delta \|\cdot\|_{\mathrm{TV}}$. Rates: Jackson. ∎

Reading the bound

- **Rates**: m^{-2k} (KL), m^{-k} ($\mathcal{W}_1$) for a C^k log-density, vs $1/m$ agnostic. $\Delta^k M_k$: dimensionless (variation of the log-density over the support).
- **Variance form** (Barron–Sheu), from $e^y - 1 - y \le \frac{1}{2} y^2 e^{y_+}$:

$$\mathrm{KL}(\nu \,\|\, \hat{\nu}_m) \;\le\; \frac{1}{2} \inf_{\deg P \le m,\ \int (g-P)\,\mathrm{d}\nu = 0} \int (g-P)^2 \left(\mathrm{d}\nu + e^P \,\mathrm{d}\lambda \right) :$$

  counts the error *where the mass is* only.
- **Limits**: g bounded = a density bounded away from 0 and ∞ on the reference support (fine on a finite support with a positive reference; not for a density vanishing inside its support); the smooth rates need a smooth log-density.



$\nu = \mathrm{Bin}(400, 0.3)$, $m = 4$: $E_m(g) \approx 4.4$ (useless), yet $\mathrm{KL}(\nu \,\|\, \hat{\nu}_m) = 1.4 \cdot 10^{-8}$; variance form: $2.8 \cdot 10^{-8}$.

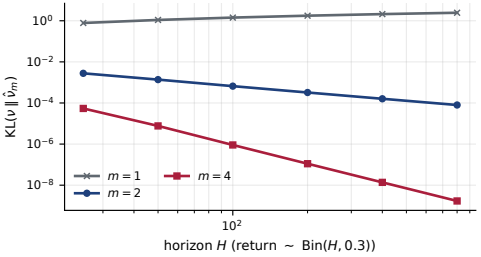
Why it works. P misses g by tens of nats only for $k < 43$ or $k > 215$, where ν and $\hat{\nu}_m$ have mass $\approx 10^{-20}$. On $120 \pm 4\sigma$ the log-density is nearly quadratic, $|g - P| < 10^{-2}$; the mass-weighted error is $\int (g-P)^2 \,\mathrm{d}\nu = 2.7 \cdot 10^{-8}$.

Where the horizon hides

1. **No explicit accumulation over the H steps**: $\frac{1}{2}E_m(g)^2$ is dimensionless: it bounds the final law alone.
2. **In the shape of ν, favourably.** Sum of H i.i.d. lattice rewards (mean μ, variance σ^2), $x = (W_{0:H} - H\mu)/(\sigma\sqrt{H})$, counting reference on the unit lattice (Edgeworth heuristic):

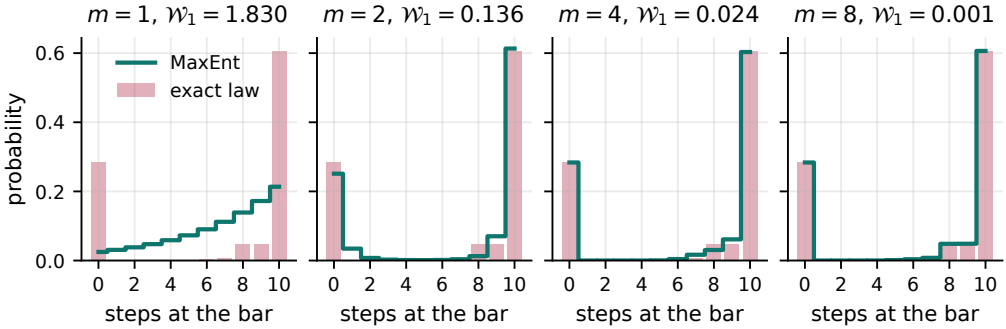
$$\ln \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \approx -\frac{x^2}{2} - \frac{1}{2}\ln(2\pi H\sigma^2) + P_H(x),$$

P_H a polynomial correction with vanishing coefficients, that of x^j of order $H^{-(j-2)/2}$ ($j \geq 3$): a degree-m polynomial follows g where the mass is, suggesting $\mathrm{KL}(\nu \,\|\, \hat{\nu}_m) = O(H^{-(m-1)})$ (m even): observed on the binomial family (figure), a local-limit heuristic, not a theorem.



$\nu = \mathrm{Bin}(H, 0.3)$, counting measure on $\{0, \ldots, H\}$: slopes -1, -3 for $m = 2, 4$; $m = 1$ (truncated geometric law) grows like $\frac{1}{2}\ln H$.

The child's law, reconstructed from m moments



- Reference: counting measure on the exact support $\{0, \dots, 10\}$; Chebyshev moments.
- $m = 1$: geometric profile, blind to the two modes. $m = 2$: both modes already (on a known bounded support, a nearly two-point law is pinned by its mean and variance), the small atoms at 8, 9 not resolved. $m = 4$: they are. $m = 8$: indistinguishable. $m = 10$: the eleven masses exactly (a Vandermonde system).
- **The support matters**: a priori support $\{0, \dots, 16\} \Rightarrow \mathcal{W}_1 = 2.48$ instead of 0.14 ($m = 2$).

Why relative entropy? Moment matching is a projection for KL only

Minimum-divergence reconstruction: f-divergence $D_f(\nu, \nu') = \int f\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right) \mathrm{d}\nu'$ (f strictly convex, C^2, $f(1) = 0$); minimize $D_f(\nu', \lambda)$ over the fiber $C(c)$. Lagrange: an interior minimizer $q\lambda$ satisfies $f'(q) = \theta_0 + \langle \theta, T \rangle$ (θ_0 fixing $\int q \, \mathrm{d}\lambda = 1$), i.e. lies in $\mathcal{E}_f = \{(f')^{-1}(\theta_0 + \langle \theta, T \rangle)\, \lambda\}$; for KL, $(f')^{-1} = \exp$: the exponential family.

Definition (Entropy-adapted divergence). [Bec, Le Corre, G. 2026] D_f is *T-entropy-adapted* if, for every finite $\mathcal{X} \subset \mathbb{R}$ and every reference probability λ charging all of $\mathcal{X}$, the D_f-projection onto $\mathcal{E}_f$ of every $\nu \in C(c)$ is the moment-matching member (when it exists): $\arg\min_{\nu' \in \mathcal{E}_f} D_f(\nu, \nu') = \{\hat{\nu}_m\}$, $\{\hat{\nu}_m\} = \mathcal{E}_f \cap C(c)$.

Theorem (Bec, Le Corre, G. 2026). For every linearly independent family $T = (T_0 \equiv 1, T_1, \ldots, T_m)$, $m \geq 1$, KL is, up to equivalence ($f \sim \alpha f + \gamma(u - 1)$), the only C^2 f-divergence that is T-entropy-adapted.

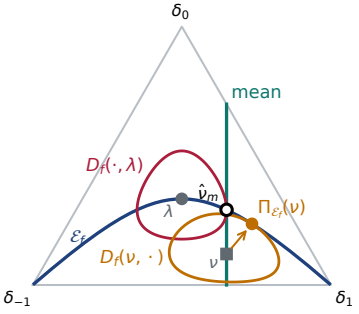


Fig.: reverse KL $D_f(\nu, \nu') = \mathrm{KL}(\nu' \,\|\, \nu)$, family $\nu' \propto \lambda/(1 + gx)$, laws on $\{-1, 0, 1\}$, $m = 1$. The level set of $D_f(\cdot, \lambda)$ touches the fiber at $\hat{\nu}_m$; that of $D_f(\nu, \cdot)$ touches the family away from $\hat{\nu}_m$, at $\Pi_{\mathcal{E}_f}(\nu)$ (mean 0.47).

Other f-divergences also reconstruct from moments; moment matching = projection is specific to KL.

Summary

1. What dynamic programming can optimize. Among monotone continuous criteria, the greedy backward recursion is exact for the strictly increasing transforms of β-means (the mean included) and, for monotone continuous nonconstant criteria, for nothing else: independence $\Rightarrow$ expected utility, translation $\Rightarrow$ an ODE. Otherwise greedy planning may fail, and memory may be needed.

2. What it can propagate exactly. Bellman closed = compatible with shifts and mixtures: for utilities, the exponential polynomials, in particular the first m moments (and the support endpoints). Two laws with the same m moments on an interval of length Δ are within $6\Delta/m$ in $\mathcal{W}_1$ (Jackson), some pairs are exactly $\Delta/(m+1)$ apart: no accumulation of errors, the horizon only through Δ.

3. Which law to reconstruct. The maximum-entropy law is the information projection of every law with the same moments (Pythagoras); for a bounded $g = \ln \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}$, $\mathrm{KL}(\nu \,\|\, \hat{\nu}_m) \leq \frac{1}{2} E_m(g)^2$, hence $\mathcal{W}_1 \leq \frac{\Delta}{2} E_m(g)$ (Hoeffding + Jackson): rates m^{-2k} and m^{-k} for $g \in C^k$. Within the smooth f-divergences, KL is the only one with this projection property.

Thank you!

References (1)

- A. Marthe, A. Garivier, C. Vernade. *Beyond average return in Markov decision processes.* NeurIPS 2023.
  A. Marthe. *Risk-sensitive planning in Markov decision processes: distributional perspective and entropic front.* PhD thesis, ENS de Lyon, 2026.
- R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.
- J. Bringas Miranda, A. Garivier, O. Cappé. *A novel approach to distributional reinforcement learning via moment matching.* Preprint, 2026.
- J. Bec, T. Le Corre, A. Garivier. *On minimum divergence for distribution reconstruction.* arXiv:2610.21872, 2026.
- M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, W. Dabney. *Statistics and samples in distributional reinforcement learning.* ICML 2019.
- M. G. Bellemare, W. Dabney, R. Munos. *A distributional perspective on reinforcement learning.* ICML 2017.
  W. Dabney, M. Rowland, M. G. Bellemare, R. Munos. *Distributional reinforcement learning with quantile regression.* AAAI 2018.
- M. G. Bellemare, W. Dabney, M. Rowland. *Distributional Reinforcement Learning.* MIT Press, 2023.
- R. A. Howard, J. E. Matheson. *Risk-sensitive Markov decision processes.* Management Science, 1972.
  M. J. Sobel. *The variance of discounted Markov decision processes.* J. Appl. Probab., 1982.

References (2)

- J. von Neumann, O. Morgenstern. *Theory of Games and Economic Behavior.* 1944. I. N. Herstein, J. Milnor. *An axiomatic approach to measurable utility.* Econometrica, 1953.
- T. J. Rivlin. *An Introduction to the Approximation of Functions.* 1969. E. W. Cheney. *Introduction to Approximation Theory.* 1966. N. I. Akhiezer. *The Classical Moment Problem.* 1965.
- E. T. Jaynes. *Information theory and statistical mechanics.* Phys. Rev., 1957. I. Csiszár. *I-divergence geometry of probability distributions and minimization problems.* Ann. Probab., 1975.
- A. R. Barron, C.-H. Sheu. *Approximation of density functions by sequences of exponential families.* Ann. Statist., 1991.
- C. Musco, C. Musco, L. Rosenblatt, A. V. Singh. *Sharper bounds for Chebyshev moment matching, with applications.* COLT 2025.
- M. J. Kearns, S. P. Singh. *Finite-sample convergence rates for Q-learning and indirect algorithms.* NeurIPS 1998.
- R. S. Sutton, A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 2018.
- R. Pemantle. *A survey of random processes with reinforcement.* Probability Surveys, 2007. R. Chetrite, H. Touchette. *Variational and optimal control representations of conditioned and driven processes.* J. Stat. Mech., 2015.

One utility, without smoothness: the mean or an exponential moment

Theorem (Bellman-closed utilities). [Marthe, G., Vernade 2023] f continuous, nonconstant. Ψ_f Bellman closed $\iff$ $f(x) = c_1 x + c_0$ or $f(x) = c_1 e^{\beta x} + c_0$ ($c_1 \neq 0$, $\beta \neq 0$).

Proof. $\Leftarrow$: $\int e^{\beta(x+c)}\,\mathrm{d}\nu = e^{\beta c}\int e^{\beta x}\,\mathrm{d}\nu$. $\Rightarrow$: the functional-equation lemma with $m = 1$: $f(x+c) = \alpha(c)f(x) + \gamma(c)\ \forall x, c$.

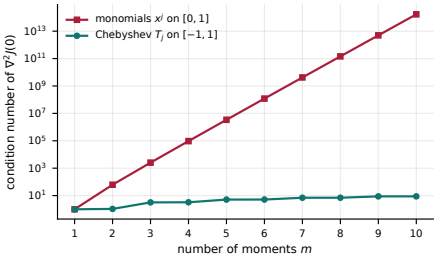
1. *α continuous, positive, multiplicative.* $f(x_1) \neq f(x_2)$: $\alpha(c) = \frac{f(x_1+c)-f(x_2+c)}{f(x_1)-f(x_2)}$ continuous, $\alpha(0) = 1$, $\alpha \neq 0$. $g \coloneqq f - f(0)$: $g(x+c) = \alpha(c)g(x) + g(c)$. Expand $g(x+c+d)$ in two ways: $\big(\alpha(c+d) - \alpha(c)\alpha(d)\big)g(x)$ independent of x, g nonconstant $\Rightarrow \alpha(c+d) = \alpha(c)\alpha(d)$; $\alpha(c) = \alpha(c/2)^2 > 0$; $\ln\alpha$ continuous, additive (Cauchy) $\Rightarrow \alpha(c) = e^{\beta c}$.
2. $\beta = 0$: g continuous, additive $\Rightarrow g(x) = c_1 x$.
3. $\beta \neq 0$: swap $x \leftrightarrow c$ in $g(x+c) = e^{\beta c}g(x) + g(c)$: $g(x)\,(e^{\beta c} - 1) = g(c)\,(e^{\beta x} - 1)$; fix $c \neq 0$: $g(x) = c_1(e^{\beta x} - 1)$.

■

Bridge to Part 2: the translation property, applied in both directions, makes Ψ_f shift compatible, hence (lemma of Part 3 with $m = 1$) $f(x+c) = \alpha(c)f(x) + \gamma(c)$; steps 1–3 then give the characterization without any smoothness.

Backup: computing the maximum-entropy law

- $\lambda = \sum_{k \le n} w_k \delta_{x_k}$ (exact support, or Gauss–Legendre nodes): $J, \nabla J, \nabla^2 J$ finite sums, $O(m^2 n)$.
- **Newton** on the strictly convex J ($m \le 10$ here): $\theta' = \theta - \mathbb{C}\mathrm{ov}_{v_\theta}[T]^{-1}\big(\mathbb{E}_{v_\theta}[T] - c\big)$, backtracking line search. In our runs, 4 to 30 Newton steps bring the moment residual below 10^{-9} (checked and reported by the solver).
- **Certificate**: $\nabla J(\theta)$ = moment mismatch of v_θ; $J(\theta) - J(\theta^\star) = \mathrm{KL}(\hat{v}_m \,\|\, v_\theta)$ (Pythagoras).
- **Conditioning**: monomials $\Rightarrow$ Hilbert-type conditioning, $\kappa \approx e^{3.5m}$. Use *Chebyshev* $T_j = \cos(j \arccos u)$ on the rescaled *exact support*.
- **Existence**: a finite $\theta^\star \iff c$ in the relative interior of the moment polytope $\mathrm{conv}\, T(\mathrm{supp}\, \lambda)$ (boundary: restrict to the face, or a limit); estimated moments may fall outside: shrink c toward the moments of the normalized reference, at the price of an error.



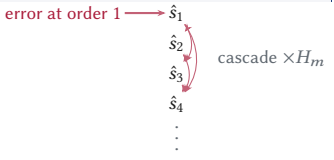
Condition number of $\nabla^2 J(0)$, λ uniform (400 points).

Outlook: the same recursion from samples (discounted setting)

Setting of the paper: states $x \in \mathcal{X}$, a fixed policy π with transition matrix P^π, deterministic reward $0 \le r(x) \le 1$, discounted return $\sum_{t\ge 0} \gamma^t r(X_t)$ with $0 < \gamma < 1$; $s_j^\pi(x) \coloneqq \mathbb{E}_x\big[(\sum_t \gamma^t r(X_t))^j\big]$, $s_0 \equiv 1$; a generative sampler of next states, n calls per state and per phase (total $|\mathcal{X}|nL$, model-based $|\mathcal{X}|n$). Moment recursion = fixed point: $s_j(x) = \sum_{i\le j} \binom{j}{i} r(x)^{j-i}\gamma^i (P^\pi s_i)(x)$. Two sample-based versions (Kearns–Singh 1998): **phased moment TD** (fresh samples into the recursion, phase after phase, one batch for all orders) and **model-based** (estimate P^π once, solve in the empirical MDP).

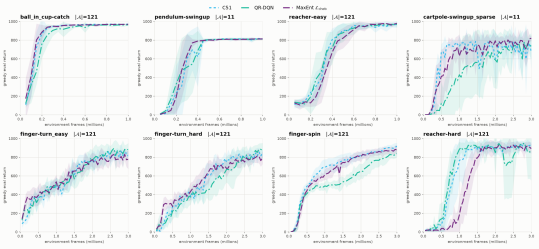
Theorem (Finite-sample estimation of m moments). [Bringas Miranda, G., Cappé 2026] $H_m = \sum_{j\le m} 1/j$; budget $n \;\ge\; \dfrac{c\,H_m^2}{\varepsilon^2(1-\gamma)^2} \ln\dfrac{m|\mathcal{X}|L}{\delta}$ calls per state and per phase ($L = O(\ln(m/\varepsilon)/\ln(1/\gamma))$ phases; $L = 1$ model-based). With probability $\ge 1-\delta$: $\max_{1\le j\le m}(1-\gamma)^j \|\hat{s}_j - s_j^\pi\|_\infty \le \varepsilon$.

- Lower-triangular recursion: an error at order i **cascades** to all $j > i$ (binomial coefficients). Cost of the cascade: the harmonic factor $H_m^2 = O(\ln^2 m)$ only. m moments are barely harder to estimate than the mean.
- Natural scale of the j-th moment: $(1-\gamma)^{-j}$.



Outlook: MaxEnt-DQN, the same idea with a neural network

- Network output, per action: the m natural parameters $\lambda_\theta(x, a)$ of an exponential family on the normalized return support.
- Transition (x, a, r, x'): **target moments** = moment recursion applied to the network's law at $(x', a^\star)$, $a^\star$ greedy for the mean. No categorical or quantile projection for the targets; the network still approximates the law.
- Loss in Chebyshev-moment coordinates, $L = \sum_{k \le m} k^{-2} (\hat{c}_k - c_k(\lambda_\theta))^2$; on $[-1, 1]$, $\mathcal{W}_1(\text{net}, \text{target}) \le 36/m + \sqrt{L}$ (Musco et al. 2025, Thm. 1; vacuous at $m = 8$).
- 8 DeepMind Control tasks (discretized actions): $m = 8$ moments $\approx$ C51 (categorical DQN, 51 atoms), > QR-DQN (quantile DQN, 51 quantiles) on most tasks.

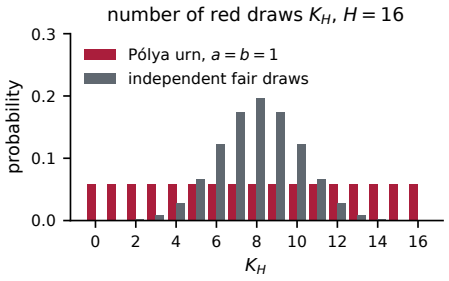


Greedy evaluation return during training, mean over 10 seeds, bands ±1 s.d.

Backup: A reinforced process in the same framework: the Pólya urn

Why: the reinforced processes of this workshop fit the framework (no decision, time-dependent p_t).

- Weights a (red), b (blue); draw a colour with probability proportional to its weight and *add* one unit of it.
- $K_t \in \{0, \dots, t\}$ red draws by time t ($a, b > 0$, $K_0 = 0$): the colour sequence is not Markov, the **count** is: $\mathbb{P}(\text{red at } t+1 \mid \text{past}) = p_t(K_t)$, $p_t(k) = \frac{a+k}{a+b+t}$.
- State $S_t = K_t$, kernel p_t, reward $R_{t+1} = \mathbb{1}\{\text{red}\}$, return K_H; for $v_t(k) \coloneqq \mathcal{L}(K_H - K_t \mid K_t = k)$: $v_H = \delta_0$ and $v_t(k) = (1 - p_t(k))\, v_{t+1}(k) + p_t(k)\, \delta_1 * v_{t+1}(k+1)$.



$a = b = 1$, $H = 16$: K_H uniform; mean 8 as for independent fair draws, variance 24 against 4.

Counts in the state restore the Markov property: the return law propagates exactly. Same mean as independent draws, a much broader law: $K_H/H \longrightarrow \Theta \sim \text{Beta}(a, b)$, a *random* limit (backup).

Backup: the Pólya urn, exact law and moments

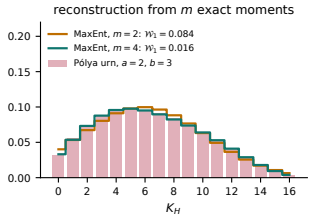
- **Exchangeability**: a colour sequence with k red draws among H has probability $a^{\overline{k}}\, b^{\overline{H-k}} / (a+b)^{\overline{H}}$ whatever the order ($x^{\overline{j}} \coloneqq x(x+1)\cdots(x+j-1)$, $x_{\underline{j}} \coloneqq x(x-1)\cdots(x-j+1)$): the **beta-binomial** law and the factorial moments

$$\mathbb{P}(K_H = k) = \binom{H}{k} \frac{a^{\overline{k}}\, b^{\overline{H-k}}}{(a+b)^{\overline{H}}}, \qquad \mathbb{E}\big[(K_H)_{\underline{j}}\big] = (H)_{\underline{j}}\, \frac{a^{\overline{j}}}{(a+b)^{\overline{j}}} \qquad (0 \le k, j \le H),$$

  $\mathbb{E}K_H = \frac{Ha}{a+b}$, $\mathbb{V}\mathrm{ar}K_H = \frac{Hab\,(a+b+H)}{(a+b)^2(a+b+1)}$ (of order H^2, not H). For $a = b = 1$: uniform on $\{0, \dots, H\}$.

- **De Finetti**: draw $\Theta \sim \mathrm{Beta}(a, b)$ once, then the draws are i.i.d. red with probability Θ; $p_t(k) = \mathbb{E}[\Theta \mid K_t = k]$ is the posterior mean: a reinforced process or a Bayesian predicting a fixed unknown Θ, the same law. $K_H/H \to \Theta$ almost surely: a random limit, no concentration around a constant.

- **Reconstruction**: $\mathcal{W}_1 \le 6H/m$ applies; for $a = 2$, $b = 3$, $H = 16$, the maximum-entropy law from $m = 2$ and 4 moments on $\{0, \dots, 16\}$ is at $\mathcal{W}_1$ distance 0.084 and 0.016 (figure); for $a = b = 1$ the law is uniform, the maximum-entropy law without constraint.



$a = 2$, $b = 3$, $H = 16$: the law of K_H and its maximum-entropy reconstructions.

Backup: exponential tilting, Doob transform and a variational identity

Fixed Markov policy, $\mathbb{P}$ its path law from $S_0 = s$, $W = W_{0:H}$; $Z_t(s) \coloneqq \mathbb{E}[e^{\beta W_{t:H}} \mid S_t = s]$, $Z_H \equiv 1$; $L_t(s,s')(\beta) \coloneqq \int e^{\beta r} \rho_t(s, \pi_t(s), s')(\mathrm{d}r)$, the Laplace transform of the reward law.

- **Feynman–Kac**: $Z_t(s) = \sum_{s'} p_t^{\pi}(s,s')\, L_t(s,s')(\beta)\, Z_{t+1}(s')$; the kernel $p^{\pi} L$ is nonnegative, in general not stochastic.
- **Doob h-transform** with $h = Z$: $p_t^{(\beta)}(s,s') \coloneqq p_t^{\pi}(s,s')\, L_t(s,s')(\beta)\, Z_{t+1}(s')/Z_t(s)$ is a Markov kernel; with the reward laws tilted by $e^{\beta r}/L_t(s,s')(\beta)$, the path law becomes $\mathrm{d}\mathbb{P}_\beta/\,\mathrm{d}\mathbb{P} = e^{\beta W}/Z_0(s)$: the chain tilted toward high ($\beta > 0$) or low ($\beta < 0$) returns, not a conditioning on an event.
- **Gibbs variational identity** ($\beta \neq 0$): $\mathrm{KL}\Big(Q \,\Big\|\, \mathbb{P}_\beta\Big) = \mathrm{KL}(Q \,\|\, \mathbb{P}) - \beta \mathbb{E}_Q W + \ln Z_0(s) \geq 0$, equality iff $Q = \mathbb{P}_\beta$, so

  $$\Psi_\beta\big(v_0^{\pi}(s)\big) = \frac{1}{\beta} \ln Z_0(s) = \sup_{Q \ll \mathbb{P}} \Big\{ \mathbb{E}_Q W - \tfrac{1}{\beta}\, \mathrm{KL}(Q \,\|\, \mathbb{P}) \Big\} \quad (\beta > 0), \qquad = \inf_{Q \ll \mathbb{P}} \Big\{ \mathbb{E}_Q W + \tfrac{1}{|\beta|}\, \mathrm{KL}(Q \,\|\, \mathbb{P}) \Big\} \quad (\beta < 0)\,:$$

  at the optimizer $Q = \mathbb{P}_\beta$: $\Psi_\beta(v_0^{\pi}(s)) = \mathbb{E}_{\mathbb{P}_\beta} W - \frac{1}{\beta}\, \mathrm{KL}\Big(\mathbb{P}_\beta \,\Big\|\, \mathbb{P}\Big)$, an entropy-regularized optimum over *path* laws (the tilted mean $\partial_\beta \ln Z_0$ is not the β-mean).
- Part 3 projects *return* laws, not path laws: the same relative entropy on a different space. With $R_{t+1} = \mathbb{1}\{S_{t+1} = v\}$, W is an occupation time and $\beta > 0$ favours paths visiting v: a reweighting of paths, not an online reinforcement of v.