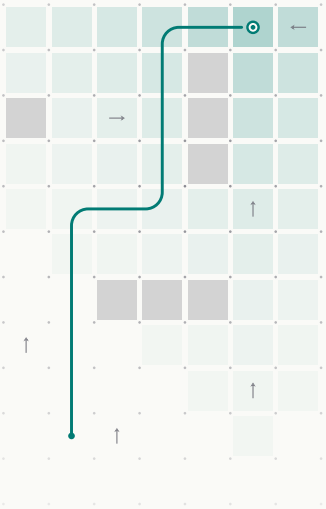
RL COFFEE · 6 NOVEMBER 2026

The Duality of Perception and Action in Brains, Minds, and Machines

A tour through my thesis

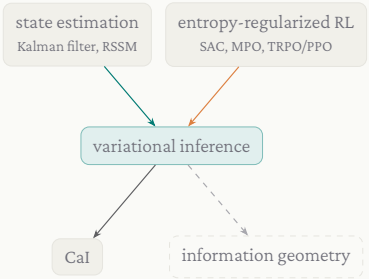


Nikola Milosevic

Max Planck Institute for Human Cognitive and Brain Sciences & Leipzig University

What the thesis is about

Control as Inference (CaI)



Bayes vs. ELBO on a bandit

Bayes' rule

$$\pi^*(a) = p(a \mid \mathcal{O}) = \frac{p(\mathcal{O} \mid a)\,\pi_0(a)}{p(\mathcal{O})}$$

same π^* if π is free, an approximation in a family

ELBO: free-energy minimization

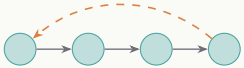
$$\min_{\pi} \; \mathrm{KL}\left(\pi \,\|\, \pi_0\right) - \mathbb{E}_{\pi}\left[\log p(\mathcal{O} \mid a)\right]$$

$p(\mathcal{O} \mid a) \propto e^{\eta\, r(a)}$: ELBO = KL-regularized return, posterior = softmax policy

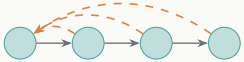
CaI is not always an MDP

Levine (2018); Neu, Jonsson & Gómez (2017).

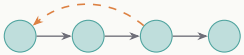
Three horizons in one visitation measure



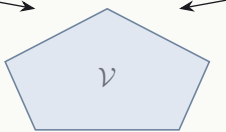
episode
reset after T



discounting
reset w.p. $1-\gamma$



survival
reset w.p. $1-\gamma\sigma(s,a)$



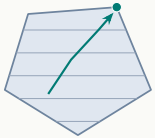
$\min_{\nu\in\mathcal{V}} \Gamma(\nu)$
reset sets the horizon, Γ the objective

$$\nu^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}^\pi(s_t = s,\, a_t = a)$$

Inspired by stochastic resetting, Evans, Majumdar & Schehr (2020), and PageRank as policy evaluation, Avrachenkov, Gregoris & Litvak (2026).

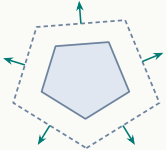
Three relaxations of the MDP

Standard
Reward maximization



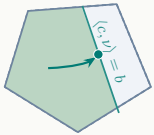
linear objective:
optimum at a vertex

Chapter 1
Survival



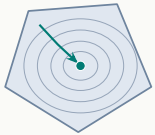
survival occupancy:
mass leaks by hazard

Chapter 2
Constrained



restricted set: barrier
keeps iterates feasible

Chapter 3
Nonlinear



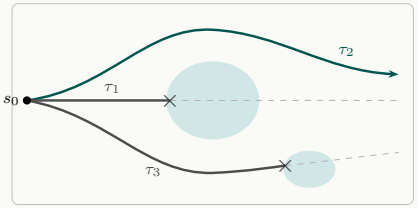
concave utility:
mirror descent to an
interior optimum

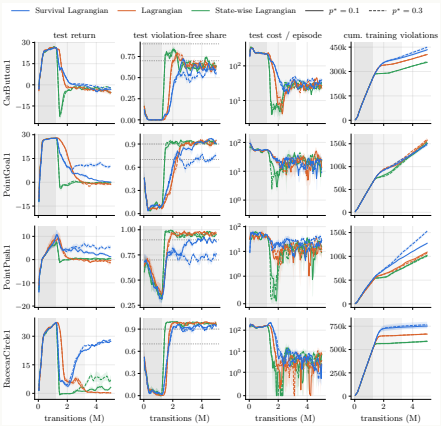
Horizon: hazards end the episode (Ch. 1)

Prior: constraints shape the geometry (Ch. 2)

Likelihood: free energies, not rewards (Ch. 3)

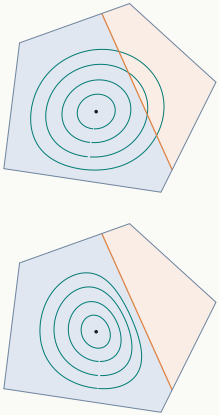
Ch. 1: constraints as a survival process

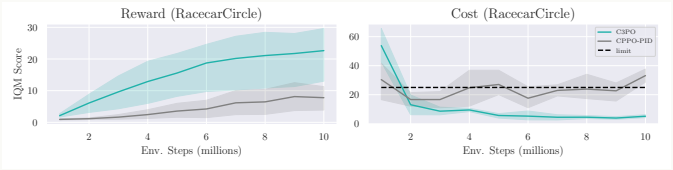




Milosevic & Franz et al. (2026), Stochastic Decision Horizons for Constrained RL.

Ch. 2: constraints shape the geometry



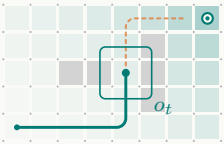


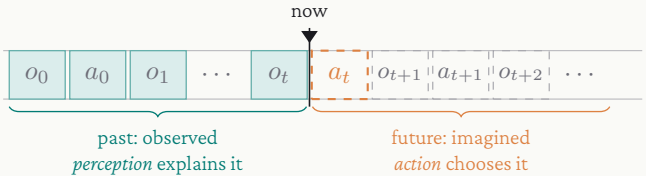
Milosevic, Müller & Scherf (2025): C-TRPO (ICML) and C3PO (EXAIT workshop); Milosevic & Scherf (2025).

Ch. 3: Model-Based CaI with General Utilities

Active Inference and Kalman Duality

An agent lives on a history process





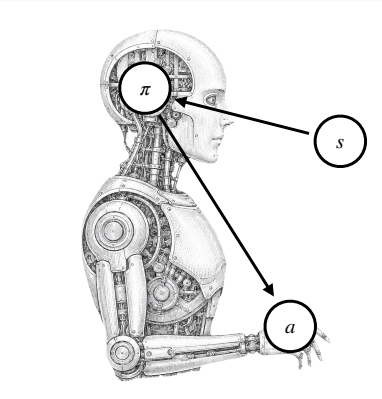
PERCEPTION

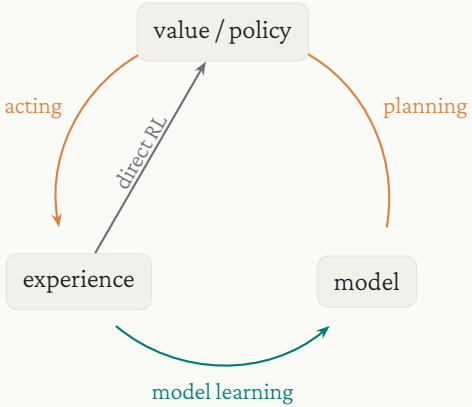
fit a world model to the observed past

ACTION

plan in it over the imagined future

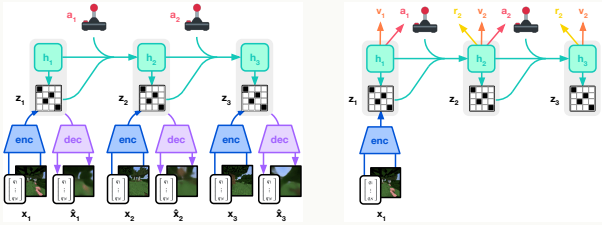
An agent that learns a model and plans in it





Sutton (1991); Sutton & Barto (2018).

World models in practice: DreamerV3



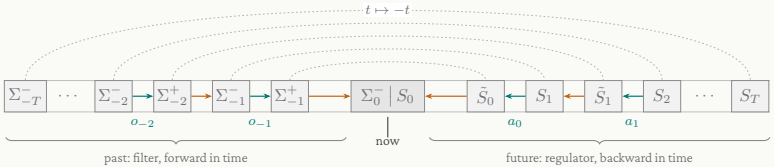
(a) World Model Learning **(b)** Actor Critic Learning

perception: an evidence bound for the model · action: an entropy-regularized actor

two losses, not derived from one objective

Hafner et al. (2023), Fig. 3.

Kalman's duality is time reversal



Past: Kalman filter, belief $\mathcal{N}(\hat{s}, \Sigma)$ on the hidden state
Σ_t^-, Σ_t^+: covariance before / after o_t

Future: regulator, cost-to-go $J_t(s) = s^\top S_t s + c_t$
S_t, $\tilde{S}_t$: value Hessian before / after improving at a_t

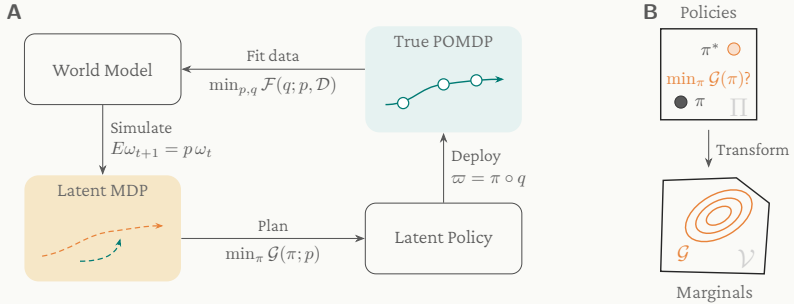
mirror cells: $\Sigma_{-t}^- \leftrightarrow S_t$, $\Sigma_{-t-1}^+ \leftrightarrow \tilde{S}_t$

Kalman (1960); Todorov (2008): beyond Gaussians only if the controller picks transitions.

THE PUZZLE OF CH. 3: WHAT IS THE DUALITY OF PERCEPTION AND ACTION?

> Kalman filter and LQR are dual.
> Is perception the mirror image of action,
> or only in linear–Gaussian problems?

One free energy for perception and action



$$\mathcal{J} = \mathcal{F} + \mathcal{G}$$

Milosevic, Hinrichs & Scherf (2026), Active Inference as a Convex MDP.

Where does the bound put its density?

variational posterior of the future: the imagined Markov process

$$\tilde{q}(\tau) = \hat{\mu}(s_0) \prod_t \pi_t(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t), \qquad \tilde{q}(s_t) = \nu_t^{\pi}(s_t)$$

Mismatched No More

one density on the whole path $\tilde{q}(\tau)$



dynamics factors cancel

linear: a KL-regularized MDP

Active inference

a density per step $\tilde{q}(s_t)$



no dynamics factor in the target

nonlinear: a convex MDP

The nonlinearity is a mutual information

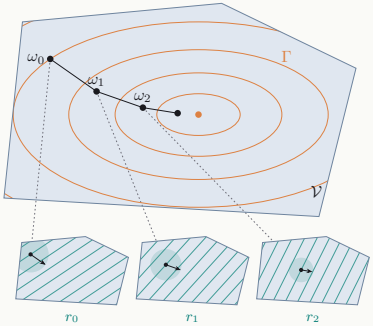
EXPECTED FREE ENERGY (risk + ambiguity, Da Costa et al. 2020)

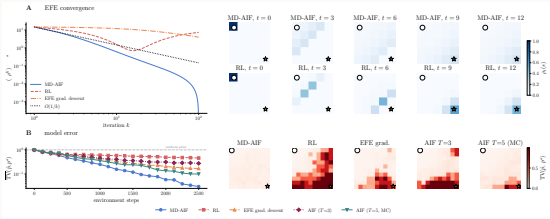
$$\mathcal{G}(\pi) = \sum_t \underbrace{\mathbb{E}_{(s,a)\sim\nu_t^\pi}\big[\log \nu_t^\pi(s,a) - \log \tilde{p}(s,a)\big]}_{\text{risk}} + \underbrace{\mathbb{E}_{s\sim\nu_t^\pi}\,\mathbb{E}_{o\sim p(o|s)}\big[-\log p(o \mid s)\big]}_{\text{ambiguity}}$$

$$\mathcal{G}(\pi) = \sum_t \Big(\underbrace{\mathbb{E}_{(s,a)\sim\nu_t^\pi}\big[-\log \tilde{p}(s,a)\big] + \mathbb{E}_{s\sim\nu_t^\pi}\,\mathbb{E}_{o\sim p(o|s)}\big[-\log p(o \mid s)\big]}_{\text{linear in } \nu_t^\pi} + \underbrace{\mathbb{E}_{(s,a)\sim\nu_t^\pi}\big[\log \nu_t^\pi(s,a)\big]}_{\text{nonlinear}} \Big)$$

Da Costa et al. (2020), Eq. (13), with state–action preferences, in thesis notation; parameter learning omitted.

Mirror Descent on $\mathcal{V}$: each step is a soft MDP

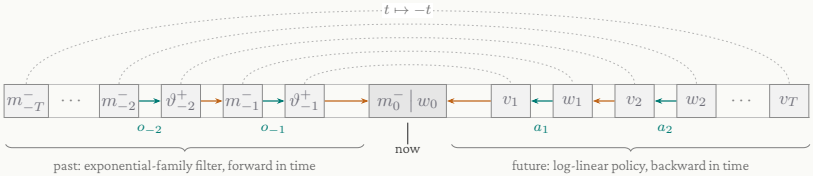




convex MDP = a sequence of soft MDPs: reward $r_k = -\nabla\Gamma(\omega_k)$, KL to π_k

Milosevic, Hinrichs & Scherf (2026); Hazan et al. (2019); Zahavy et al. (2021); Moreno et al. (2024); Lu, Freund & Nesterov (2018).

Kalman duality in POMDPs?



Past: belief $q(s) \propto e^{\vartheta^\top \psi(s)}$ on the hidden state
m^-: predicted moments, before o_t
ϑ^+: natural parameters, after o_t

Future: policy $\pi(a \mid s) \propto e^{\theta^\top \phi(s,a)}$
w_t: $Q_t(s,a) = w_t^\top \phi(s,a)$, the soft Q-function
v_t: $V_t(s) = v_t^\top \psi(s)$, the soft value

the steps mirror, the objects do not: equal recursions only for Gaussians

Perception and control, algorithmically dual

	Perception	Control
exponential step	measurement (Bayes' rule)	policy improvement
mixture step	prediction	backup through the model
direction	forward in time	backward in time

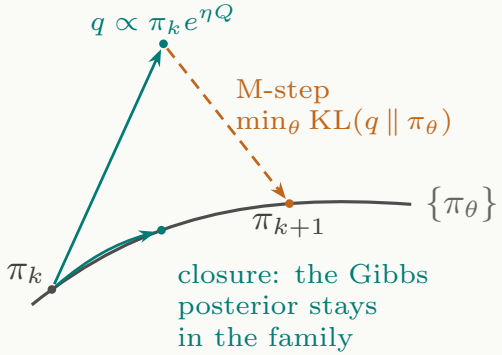
the same mirror step on two information manifolds

Milosevic et al. (2026), The Dually Flat Geometry of Planning as Inference.

MPO: The E-step is a posterior

E-STEP prior π_k, likelihood $e^{\eta Q}$

M-STEP trivial if the family is closed



Abdolmaleki et al. (2018); Minka (2001).

A TRPO-style surrogate advantage bound

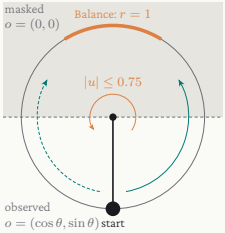
$$\Gamma(\nu^{\pi}) \leq \Gamma(\nu_k) - \mathbb{A}_k(\pi) + \left(L + \frac{2\gamma\|A_k\|_{\infty}}{1-\gamma}\right) D_{\mathrm{KL}}^{\max}(\pi\|\pi_k)$$

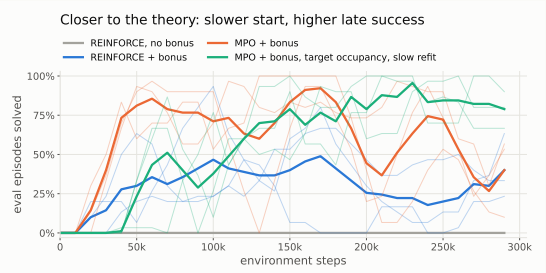
curvature of the objective + curvature of the policy-occupancy map

a planner with a model pays only L

Kakade & Langford (2002); Schulman et al. (2015); Achiam et al. (2017).

Proof of principle: PROBE



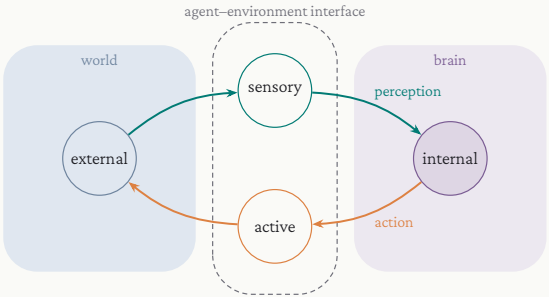


DREAMER kNN state entropy on the latent, actor in imagination

CLOSER TO THEORY MPO actor, kNN reward on the target occupancy, slow model refit (preliminary)

Ashlag et al. (2026), PROBE; Hafner et al. (2023); Abdolmaleki et al. (2018).

What would a brain compute?



DOPAMINE

prediction errors of a utility, not of a reward

CORTEX

perception manifold vs. control manifold:
which map ties them?

Friston (2010); Parr, Da Costa & Friston (2020); Schultz (1997); Stauffer et al. (2014); Doya (2021); Kataoka & Doya (2026).

Open questions

1. Is the duality structural or only algorithmic? Does it have real consequences?
2. When do nonlinear objectives pay off: exploration, robustness, safety?
3. Is the policy-dependence a real obstacle for model-based CaI?

THANK YOU

Questions?

nmilosevic@cbs.mpg.de

Backup: the LQG control problem

Linear–Gaussian world model, $s_t \in \mathbb{R}^n$, $a_t \in \mathbb{R}^p$, stage cost $\ell(s,a) = s^\top M s + a^\top R a$:

$$\begin{aligned}
\min_{\pi} \quad & J(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{T-1} \ell(s_t, a_t) + s_T^\top M_T s_T\Big] \\
\text{subject to} \quad & s_{t+1} = A s_t + B a_t + w_t, && w_t \sim \mathcal{N}(0, W), \\
& o_t = C s_t + v_t, && v_t \sim \mathcal{N}(0, N), \\
& a_t \sim \pi_t(\cdot \mid o_{0:t}), && t = 0, \dots, T-1, \\
& s_0 \sim \hat{\mu} = \mathcal{N}(0, \Sigma_0),
\end{aligned}$$

$M \succeq 0, R \succ 0, W \succ 0, N \succ 0$, independent noises. LQR: $a_t \sim \pi_t(\cdot \mid s_t)$, no emission.

Thesis Ch. 3, the LQG example.

Backup: the Riccati recursion splits like the Kalman filter

$$(\Sigma^+)^{-1} = (\Sigma^-)^{-1} + C^\top N^{-1} C \qquad \tilde{S}_t^{-1} = S_{t+1}^{-1} + B R^{-1} B^\top$$

measurement — minimize over a

$$\Sigma_{t+1}^- = A \Sigma_t^+ A^\top + W \qquad S_t = M + A^\top \tilde{S}_t A$$

prediction — backup

$$A \leftrightarrow A^\top, C \leftrightarrow B^\top, W \leftrightarrow M, N \leftrightarrow R, \Sigma_{-t}^- \leftrightarrow S_t, \Sigma_{-t-1}^+ \leftrightarrow \tilde{S}_t$$

Backup: the closed-family control problem

Features $\psi(s)$ on states and $\phi(s,a)$ on state–action pairs:

$$\min_{\pi} \quad \mathbb{E}_{\pi}\Big[\sum_{t=0}^{T-1} \ell(s_t,a_t)\Big] + \eta^{-1}\,\mathrm{KL}\left(\pi \,\|\, \pi_{\mathrm{ref}}\right) \qquad \ell = (w^{\ell})^{\top}\phi \quad \text{linear in } \phi$$

subject to	$s_{t+1} \sim p(\cdot \mid s_t, a_t),$	$\mathbb{E}\big[\psi(s_{t+1}) \mid s_t, a_t\big] = \mathsf{P}\,\phi(s_t,a_t)$ kernel closure
	$o_t \sim p(\cdot \mid s_t),$	$p(o \mid s) \propto e^{\vartheta_{\ell}(o)^{\top}\psi(s)}$ conjugate
	$a_t \sim \pi_{\theta_t}(\cdot \mid s_t),$	$\pi_{\theta} \propto e^{\theta^{\top}\phi},\ \log Z_{\theta} \in \operatorname{span}\psi$ partition closure
	$s_0 \sim \hat{\mu},$	$\hat{\mu} \propto e^{\vartheta_0^{\top}\psi}$

LQG: ψ, ϕ quadratic, $\ell = s^{\top}Ms + a^{\top}Ra$; the KL vanishes as $\eta \to \infty$.

Thesis Ch. 3, closed families (closure assumption).

Backup: the split beyond Gaussians

Perception: belief $q_t(s) \propto e^{\vartheta_t^\top \psi(s)}$

$$\vartheta_t^+ = \vartheta_t^- + \vartheta_\ell(o_t)$$
measurement

$$m_{t+1}^- = \mathsf{P}\,\xi_t$$
prediction

Control: policy $\pi_{\theta_t}(a \mid s) \propto e^{\theta_t^\top \phi(s,a)}$

$$\theta_t^{k+1} = \theta_t^k + \eta\, w_t$$
improvement

$$w_t = w_t^r + \mathsf{P}^\top v_{t+1}$$
backup

$m = \mathbb{E}_q[\psi]$, $\xi_t = \mathbb{E}[\phi(s_t, a_t)]$, conjugate likelihood $p(o \mid s) \propto e^{\vartheta_\ell(o)^\top \psi(s)}$
kernel closure $P\psi = \mathsf{P}\phi$, partition closure $V_t = \eta^{-1} \log \frac{Z_{\theta_t + \eta w_t}}{Z_{\theta_t}} = v_t^\top \psi$
$\mathsf{P} \leftrightarrow \mathsf{P}^\top$ replaces $A \leftrightarrow A^\top$; without closure, both sides project (moment matching)

Backup: the general duality is algorithmic

	Perception	Control
exponential step	belief: $\vartheta_t^+ = \vartheta_t^- + \vartheta_\ell(o_t)$ prior $\times$ likelihood	policy: $\theta_t^{k+1} = \theta_t^k + \eta\, w_t$ $\pi_k \times e^{\eta Q}$
mixture step	moments forward: $m_{t+1}^- = \mathsf{P}\,\xi_t$	values backward: $w_t = w_t^r + \mathsf{P}^\top v_{t+1}$

mixture steps are adjoint: $\langle v_{t+1}, \mathsf{P}\,\xi_t \rangle = \langle \mathsf{P}^\top v_{t+1}, \xi_t \rangle$ (a measure paired with a function)

exponential steps add natural parameters, but of different objects: a belief and a policy

Gaussians only: the Legendre map inverts a matrix, so both recursions become the same equations (Kalman).

Kalman (1960); Todorov (2008).

Backup: LQG is a special case

	closed family	LQG, $x = (s; a)$
features	$\psi(s),\ \phi(s,a)$	quadratic in s, in x
reward	$r = (w^r)^\top \phi$	$-(s^\top M s + a^\top R a)$
kernel closure	$P\psi = \mathsf{P}\phi$	linear–Gaussian p maps quadratics to quadratics
partition closure	$\log Z_\theta = v^\top \psi$	completing the square, $\Theta_{aa} \prec 0$
backup	$w_t = w_t^r + \mathsf{P}^\top v_{t+1}$	$H_t = \left(\begin{smallmatrix} M & 0 \\ 0 & R \end{smallmatrix}\right) + [A\ B]^\top S_{t+1} [A\ B]$
improvement	$\theta_t^{k+1} = \theta_t^k + \eta w_t$	$\Sigma^{-1}\ \mathrel{+}= 2\eta H_{t,aa}$, gain $K_t = H_{t,aa}^{-1} H_{t,as}$

linear cost: $s^\top M s + a^\top R a = \mathrm{tr}\left(\mathrm{diag}(M,R)\, x x^\top\right) = \langle \mathrm{diag}(M,R), \phi(s,a) \rangle$

Thesis Ch. 3, closed families; Wu et al. (2025).

Backup: Mismatched No More in our KL form

Trajectory law $\tilde{q}$ under the learned model p, target p_r under the environment P:

$$\mathbb{E}_{\tilde{q}}\Big[\log \frac{\tilde{q}(\tau)}{p_r(\tau)}\Big] = \sum_t \Big(\underbrace{\mathbb{E}_{\nu_t}\,\mathrm{KL}\left(p(\cdot \mid s,a) \,\|\, P(\cdot \mid s,a)\right)}_{\text{model}} + \underbrace{\mathbb{E}_{\nu_t}\,\mathrm{KL}(\pi_t \| \pi_{\text{ref}})}_{\text{policy}} - \langle r, \nu_t \rangle\Big)$$

Dynamics factors cancel up to the model term: a KL-regularized MDP. The paper states its bound with discounting, one policy in both laws and $(1-\gamma)\log r$.

Eysenbach et al. (2022), Eq. (2)–(3), Thm. 3.1.

Backup: the mutual information, step by step

EXPECTED FREE ENERGY, REORDERED

$$\mathcal{G}(\pi) = \sum_t \Big(\underbrace{\mathbb{E}_{\nu_t}\big[-\log \tilde{p}(s,a)\big] + \mathbb{E}_{\nu_t}\,\mathcal{H}\big[p(o \mid s)\big]}_{\text{linear}} \; - \; \mathcal{H}(\nu_t) \Big)$$

SPLIT THE ENTROPY: POLICY, THEN STATE

$$\mathcal{H}(\nu_t) = \underbrace{\mathbb{E}_{s\sim\nu_t}\mathcal{H}\big[\pi_t(\cdot \mid s)\big]}_{\text{soft-RL entropy}} + \underbrace{\langle \mathcal{H}_P, \nu_{t-1}\rangle}_{\text{linear in } \nu_{t-1}} + \underbrace{I_{\nu_{t-1}}(s_t;\, s_{t-1}, a_{t-1})}_{\text{mutual information}}$$

chain rule for $s_t \sim p(\cdot \mid s_{t-1}, a_{t-1})$, $\mathcal{H}_P(s,a) = \mathcal{H}\big[p(\cdot \mid s,a)\big]$

$$\mathcal{G}(\pi) = \text{linear in } \nu - \text{soft-RL entropy} - \textstyle\sum_t I_{\nu_{t-1}}(s_t; s_{t-1}, a_{t-1})$$

Backup: one mirror step is one soft-DP pass

MIRROR STEP linearize Γ at ω^k, add the KL to the current policy

$$\min_{\pi} \sum_{t=0}^{T-1} \left(\left\langle \nabla_t \Gamma(\omega^k),\, \omega_t^{\pi} \right\rangle + \eta^{-1}\, \mathbb{E}_{s \sim \omega_t^{\pi}}\, \mathrm{KL}\left(\pi_t(\cdot \mid s) \,\|\, \pi_t^k(\cdot \mid s) \right) \right)$$

SOFT MDP local reward $r^k := -\nabla\Gamma(\omega^k)$, reference policy π^k, temperature η^{-1}, on the imagined process $\tilde{q}$

BACKWARD PASS $V_T := 0,\ Q_t(s,a) = r_t^k(s,a) + \mathbb{E}_{s' \sim p(\cdot \mid s,a)} V_{t+1}(s')$

$$V_t(s) = \max_{\pi_t(\cdot \mid s)} \left\{ \mathbb{E}_{a \sim \pi_t} \left[Q_t(s,a) \right] - \eta^{-1}\, \mathrm{KL}\left(\pi_t \,\|\, \pi_t^k \right) \right\} \overset{\text{Gibbs}}{=} \eta^{-1} \log \mathbb{E}_{a \sim \pi_t^k(\cdot \mid s)} e^{\eta Q_t(s,a)}$$

$$\pi_t^{k+1}(a \mid s) = \pi_t^k(a \mid s)\, e^{\eta (Q_t(s,a) - V_t(s))} \qquad \text{the Gibbs posterior of the current policy}$$

linear Γ: r^k fixed, one pass · EFE: $\nabla\Gamma$ moves with ω^k, repeat the pass

Thesis Ch. 3, planning by mirror descent.

Backup: certainty equivalence is block-wise free-energy minimization

LQG: $a_t = -K_t^{\mathrm{LQR}} \hat{s}_{t|t}$

- retrospective block $\rightarrow$ Kalman posterior $q = \mathcal{N}(\hat{s}, \Sigma)$
- prospective block $\rightarrow$ regulator $\pi(a \mid s) = \delta(a + Ks)$
- deployed $\varpi = \pi \circ q$: mean $-K\hat{s}$ (the CE controller), extra cost $\mathrm{tr}(H_{aa} K \Sigma K^\top)$
- separation makes the block-wise minimization exact

Backup: episodes are stationary on the clock chain

States (s, t), reset after step $T-1$: a resetting kernel with $\rho(s,t,a) = \mathbf{1}\{t = T-1\}$.

- $\omega \in \mathcal{Q} \iff \omega/T \in \ker K$: the step-wise polytope is a visitation polytope
- soft backward pass = soft average-reward Bellman equation in the gauge $V_T = 0$, gain $= \langle \hat{\mu}, V_0 \rangle / T$
- horizon only as the policy-independent normalization $1/T$

Backup: the fixed point of the clean loop

Well-specified tabular model, exact recognition, population refit, coverage; $\Gamma_\lambda = \langle \ell, \omega \rangle + \lambda \sum \omega \log \omega$:

- every full-support policy refits p_{true}
- unique interior fixed point $(\pi^\star, p_{\text{true}})$
- $D_\Psi(\omega^\star \| \omega^k) \leq \left(1 - \frac{1}{1+L_T}\right)^k D_\Psi(\omega^\star \| \omega^0)$, independent of λ

With finite data the refit is performative; Banach needs small sensitivity (Mandal et al. 2023).