

Robust Machine Learning, Reinforcement Learning, and Autonomy: A Unifying Theory via Performance and Risk Tradeoff

John S. Baras

Institute for Systems Research
ECE, CS, ME, AE, Bioeng., DOIT, AMSSC
University of Maryland College Park, USA

January 18, 2024
UC FORTH ICS
Heraklion, Crete, Greece

Acknowledgments

- **Joint work with:** Erfan Noorani, Christos Mavridis, Matthew James, Nital Patel, Maben Rabi, Usman Fiaz, Faizan Tariq, Niles Suriyarachchi
- **Sponsors:** ONR, DARPA, NSF, ARL, ARO, Lockheed Martin, Northrop Grumman, ABB, STEER, Nokia Bell-Labs

Uncertainty: a major challenge in systems, control, communications

ROBUSTNESS – Critical inference, decision-making, control

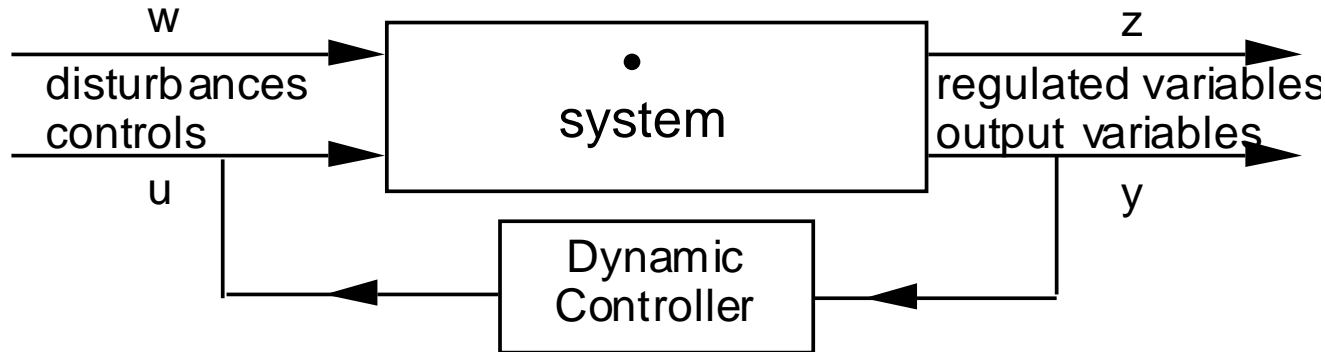
- **Control and Uncertainty**
 - Uncertainty: System Model, Performance Model, Sensor Model, Computation Model, Action Model
Environment
Noises, Disturbances, etc.
Goals and Sub-goals
 - Feedback, Adaptive, Robust, Intelligent, Learning
- **Control system:** A mapping from inputs, outputs to actions
Influenced by objectives
- **Representations:** factoring of this mapping

Some Approaches to Deal with Uncertainty

- **Differential games and uncertainty**
 - Robust control (Baras, Basar, Bernard, Fleming, Helton, James, Isidori, Bensoussan, Ball, ...)
 - Intelligent control
 - AI, planning, performance feedback
 - Learning, connectionist systems
 - Logic, knowledge-based systems
hybrid control systems
- **Reinforcement learning:** Approximate DP, Temporal Difference (TD) methods, Adaptive Critics, Q-learning, Recurrent Network Implementations
 - Barto, Sutton, Tsitsiklis, Bertsekas, Werbos, Watkins, ...

Review of some Oldies but Goodies

Generic Output Feedback Robust Control



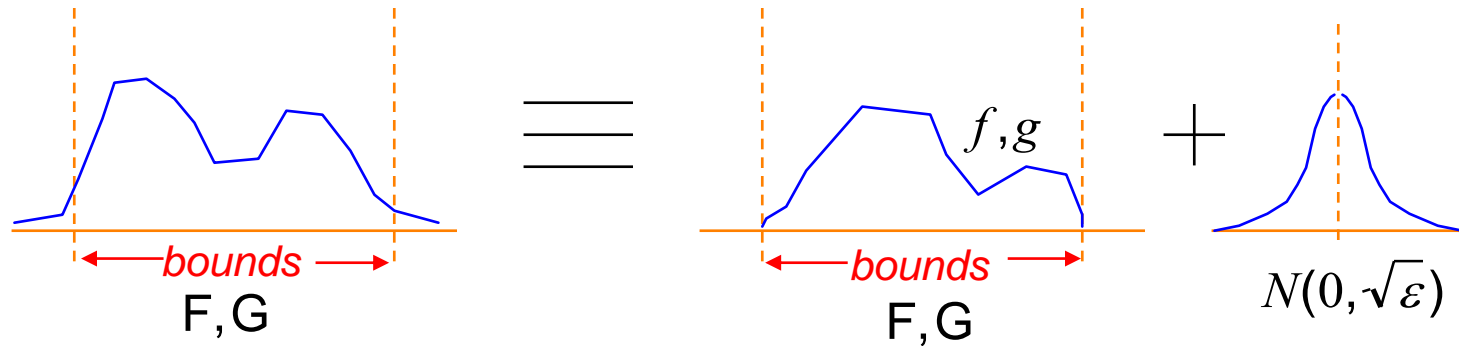
Uncertainty:

Model of system

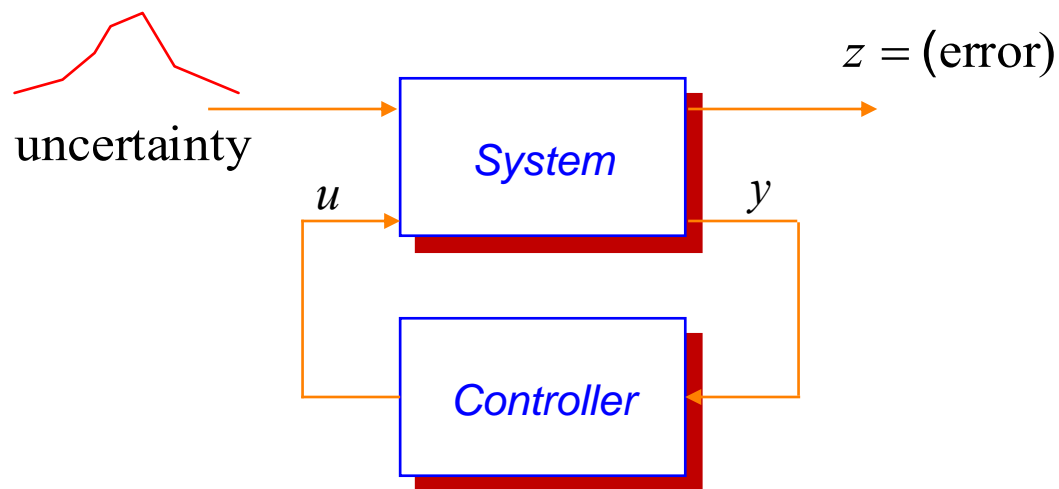
Exogenous disturbances w

Controlling Uncertainties:

Make regulated variables z
behave as desired
despite uncertainties



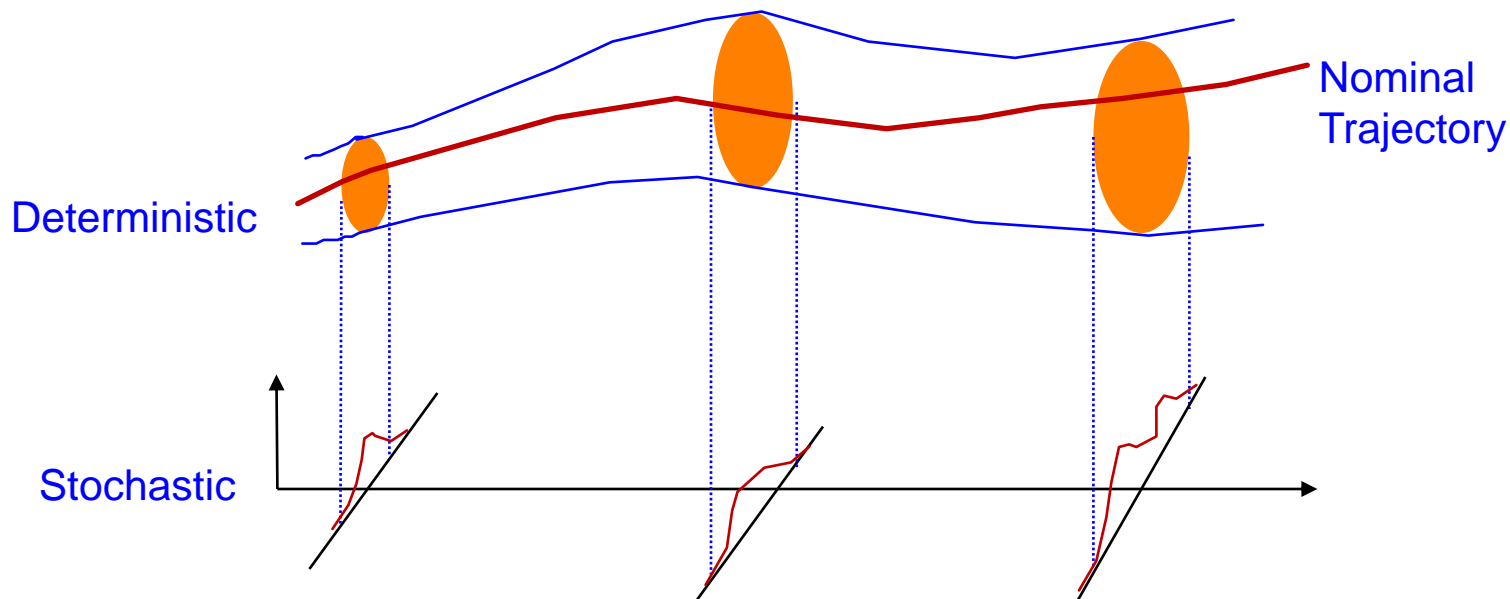
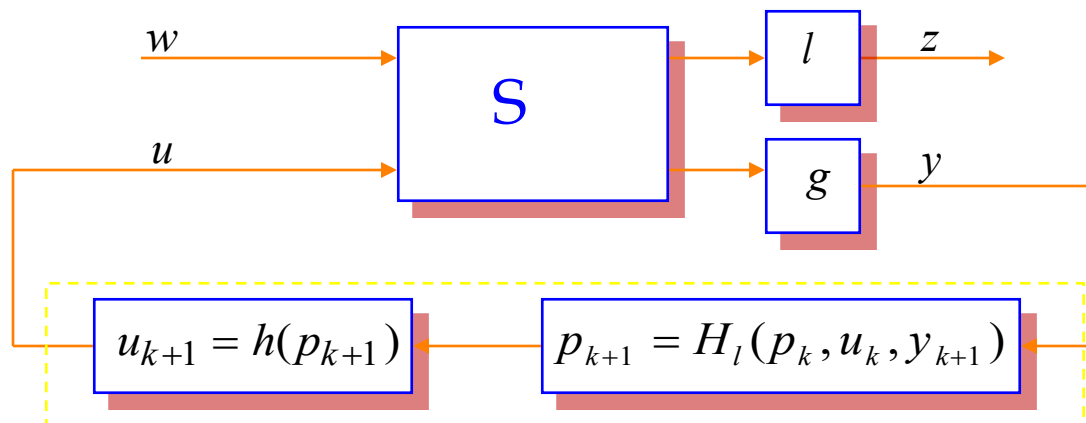
Observed uncertainty = bounded noise + outliers



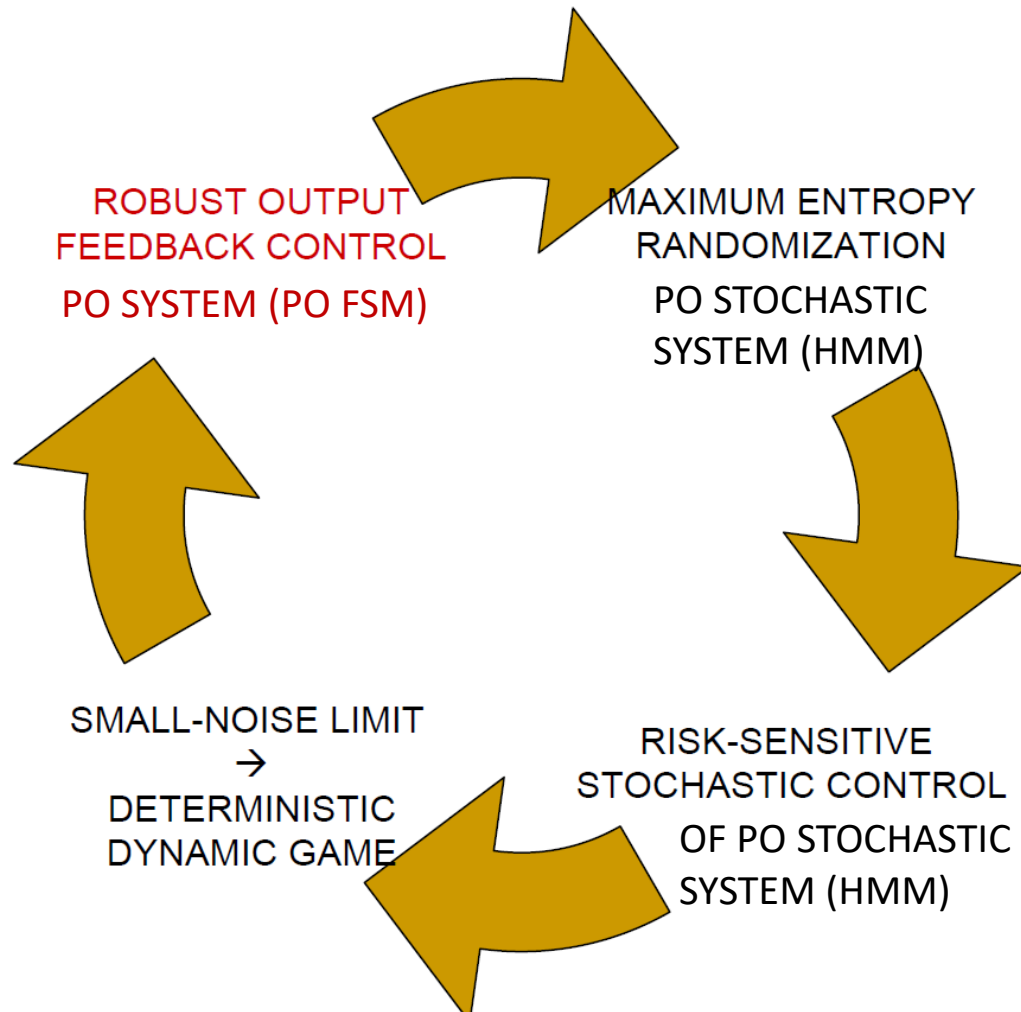
Generic Output Feedback Robust Control: Our General Solution Idea

- Estimation - the information state
 - *Express problem in terms of information state*
- Information state feedback control:
 - *Solve state feedback problem for information state*
- Coupling - information state feedback
 - *Plug information state into optimal feedback*

Our Method and Key Result

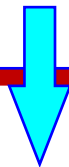


HOW? The Circuit Taken in Our Approach

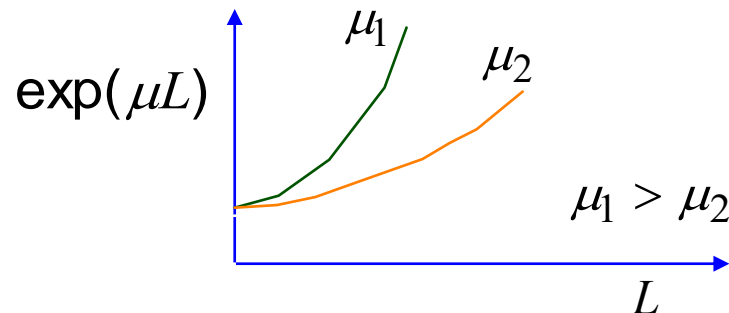


□ Objective: Minimize

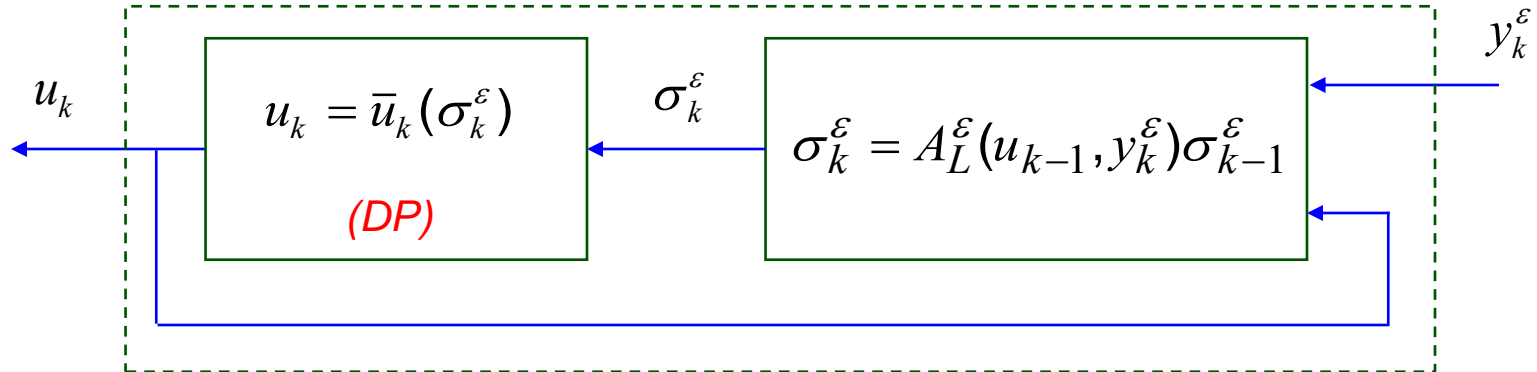
$$J^\varepsilon(u) = \mathbb{E} \left[\exp \left(\frac{\mu}{\varepsilon} \sum_{k=1}^K |z_k|^2 \right) \right], \mu > 0$$



$$\exp(\mu L) = \underbrace{1 + \mu L}_{\substack{\mu \text{ small} \\ (\text{min. variance})}} + \frac{\mu^2 L^2}{2} + \frac{\mu^3 L^3}{6} + \dots$$

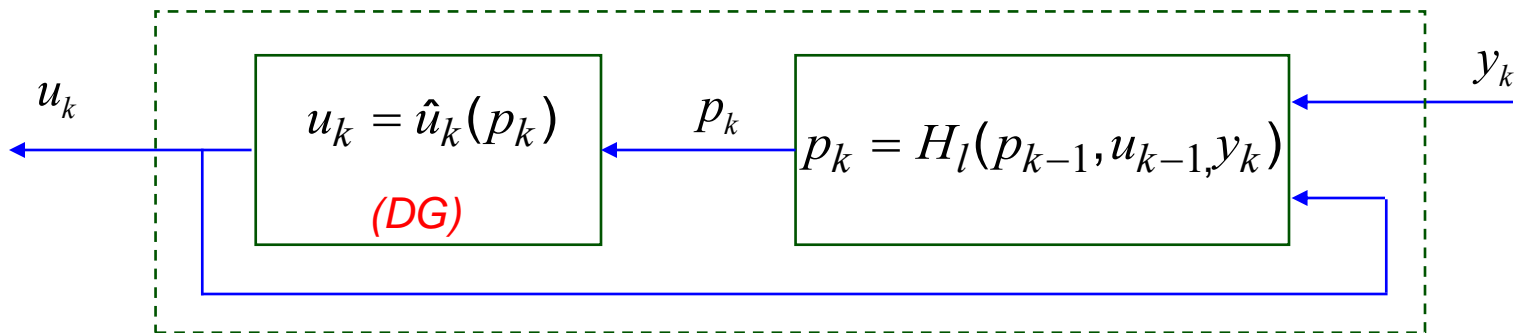


Solution of the Randomized Problem



Statistics Required

$\varepsilon \rightarrow 0$



Statistics Not Required (depends on bounds only)

- **Key analytical methods:**

The equivalence of three problems:

- **Output feedback robust control**
- **Partially observed deterministic dynamic game**
- **Partially observed risk-sensitive stochastic control**

- Nonlinear partially observed
- Set valued partially observed
- Finite Automata partially observed

1. M.R. James, J.S. Baras, R.J. Elliott, “Risk-Sensitive Control and Dyn. Games for PO Discrete-Time Nonl. Systems”, *IEEE TACON*, 1994.
2. M.R. James and J.S. Baras, “Robust H_∞ Output Feedback Control for Nonlinear Systems”, *IEEE TACON*, 1995.
3. M.R. James and J.S. Baras, “PO Differential Games, Infinite Dimensional HJI Equations, and Nonlinear H_∞ Control”, *SICON* 1996.
4. J.S. Baras and M.R. James, “Robust and Risk-Sensitive Output Feedback Control for FSM and HMM”, *J. Math. Syst., Est., Control*, 1997.
5. J.S. Baras and N.S. Patel, “Robust Control of Set-Valued Discrete Time Dynamical Systems”, *IEEE TACON*, 1998.
6. J. Baras, M. Rabi, “Maximum Entropy Models, Dynamic Games, and Robust Output Feedback Control for Automata”, *IEEE CDC*, 2005.
7. J. Baras, “Maximum Entropy Models, Dynamic Games and Robust Output Feedback Control of Nonlinear Systems”, *IEEE CDC*, 2006.

From Robust to Intelligent Control

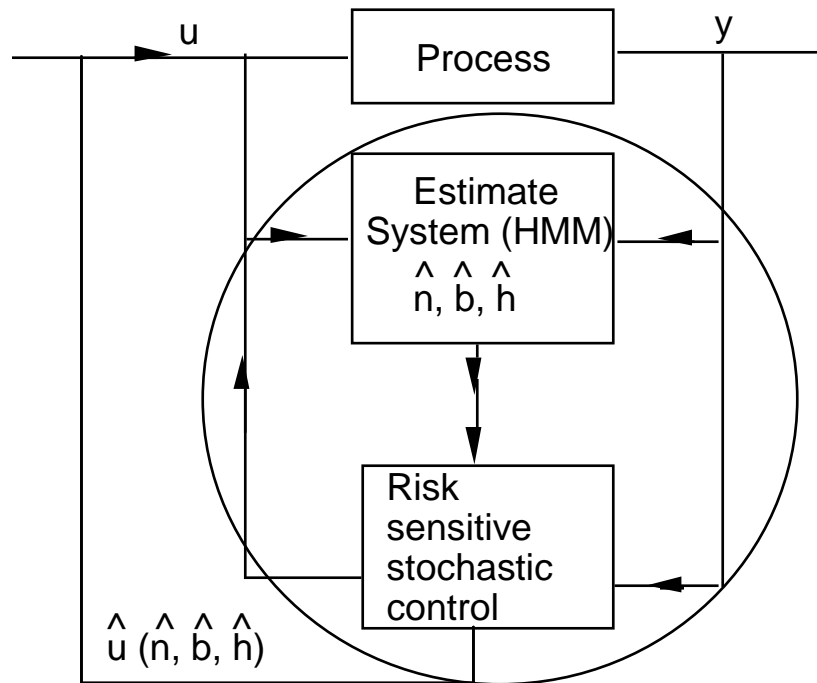
John S. Baras

**Department of Electrical Engineering and
Institute for Systems Research
University of Maryland at College Park**

**March 5, 1997
LIDS, MIT**

Unknown Models -- Learning

- When we do not have models (i.e. f , b , etc.)

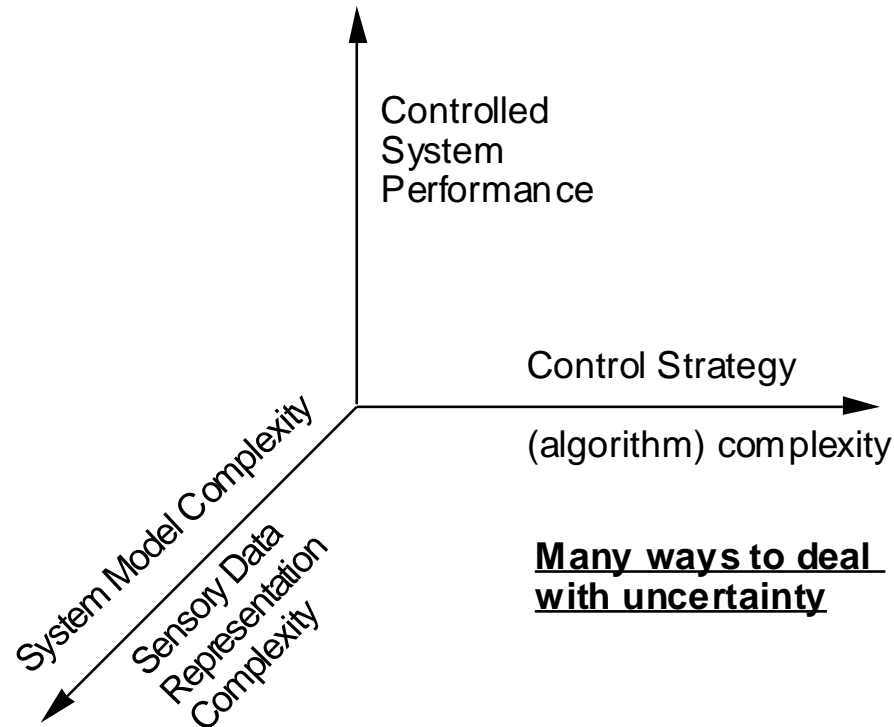


Not necessary?

Replace by
“Approximate DP”
to compute info-state
and value function!

- Control cost and model complexity cost combined
- Trade off: complexity vs. performance
- Uncertainty causes learning

Intelligent Control and Complexity



At least “3-dimensions”

Trade-offs must be considered

Examples: speech coding, speech understanding, image understanding, autonomous navigation

Intelligent Control and Complexity

- **Combine model learning:** intelligent control
- Risk sensitive control of HMM with unknown models
 θ = set of parameters of HMM, including order
- Metric for model complexity

$$MDL(k) = -\log P(y^n | \hat{\theta}) - \log \pi(\hat{\theta}) + \frac{k}{2} \log n$$

k = "length" of data

- What is the interpretation of γ in this context?
- In general solution computationally intractable
 - Approximation of info-state evolution
 - A dynamic game DP; approximation of value function

Primarily interested in "feature" - based and/or compact representations/approximations (RNN, basis fncs)

Risk-Sensitive Control with Unknown Models: Learning Information State Dynamics

- Information state dynamics not known!

Iterative learning of information state: related to features?

Must run faster than learning W

Is this a “good” way to partition handling the uncertainty?

Information state does capture the notion of “states relevant” for control

BACK TO PRESENT TIME
From Robust Control to Adversarial
and Robust Machine Learning

Extensions of Robust and Risk Sensitive Control Theory

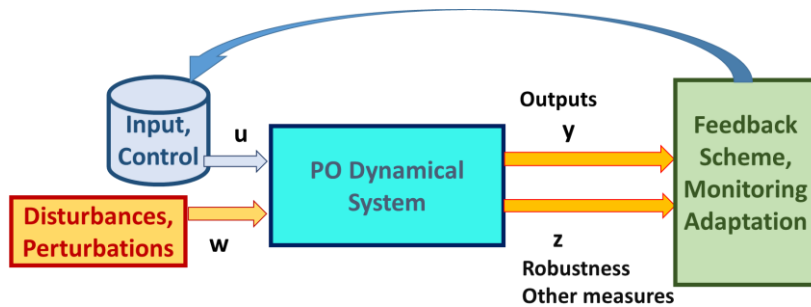


Fig.: Robust Output Feedback Control – General H_∞ ; “Four Block Problem”

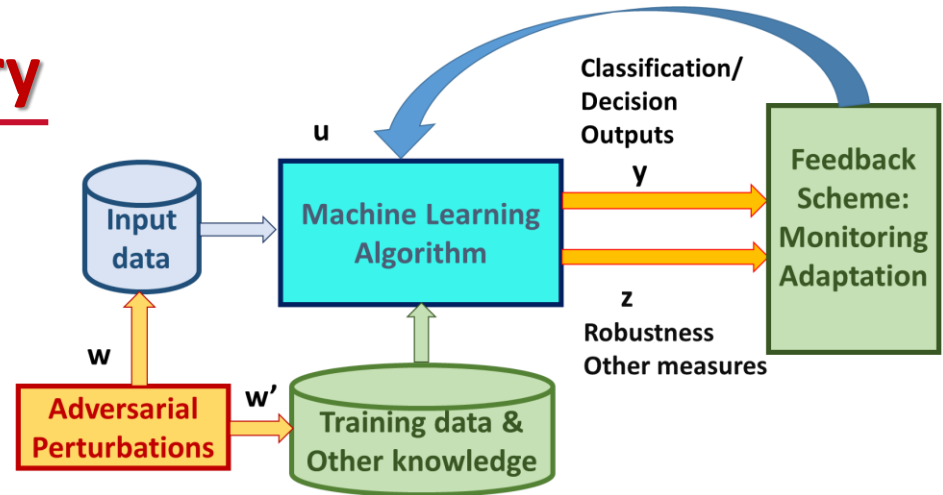


Fig.: Generalized framework to formulate and analyze robustness: Select ML algorithm, feedback u , to optimize classification/decision performance, while minimizing effect of perturbations w and w' on robustness and other measures z

- **Solution:** two coupled Hamilton-Jacobi-Bellman (HJB) equations (one on-line, computing the novel *information state*, and another off-line, for the **decision u**)
- **Complete equivalence of three previously unrelated problems:** general nonl. robust output feedback problem, a dynamic game with two players, a stochastic control problem with metric the expectation of the exponential of an integral-type perform. Measure.
- Deeper understanding of some key randomizations employed as **max. entropy modeling**.
- Extension to other models of risk – relationships with **Prospect Theory**
- Extension to **Robust ML** and **Robust AI, Robust RL**
- Extension to **Multi-agent systems**

From Robust Control to Robust Reinforcement Learning

Is there a unified theory?

YES – Extending Robust Output
Feedback Control Theory –
Baras et al [1994-98, 2005-06]

PLUS

Using Risk-Utility rigorous duality

(Rockafellar and earlier works from mathematics of finance)

T. Rockafellar review article: “Risk and Utility in the Duality Framework of Convex Analysis”, 2018

Theory Unifies

- **Robustness in ML and RL**
- **Trusted Autonomy – safety and risk**
- **Trustworthy AI**
- **Composability**
- **.....**

Reinforcement Learning (RL)

- Standard (risk-neutral) RL algorithms follow from the **Expected Utility Theory** (Von Neumann–Morgenstern) – **MDP**

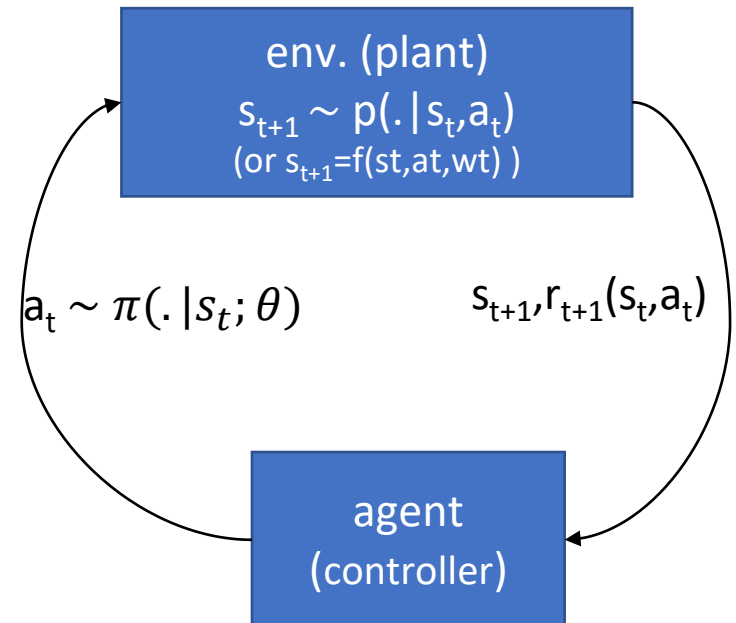
$$\mathcal{M} = (S, \mathcal{A}, p_0, P, r, \gamma)$$

$$\max_{\theta} J(\theta) := E[R], \text{ where } R = \sum_{t=0}^{|\tau|-1} \gamma^t r_t(s_t, a_t)$$

$$:= (s_1, a_1, s_2, a_2, \dots, a_{T-1}, s_T).$$

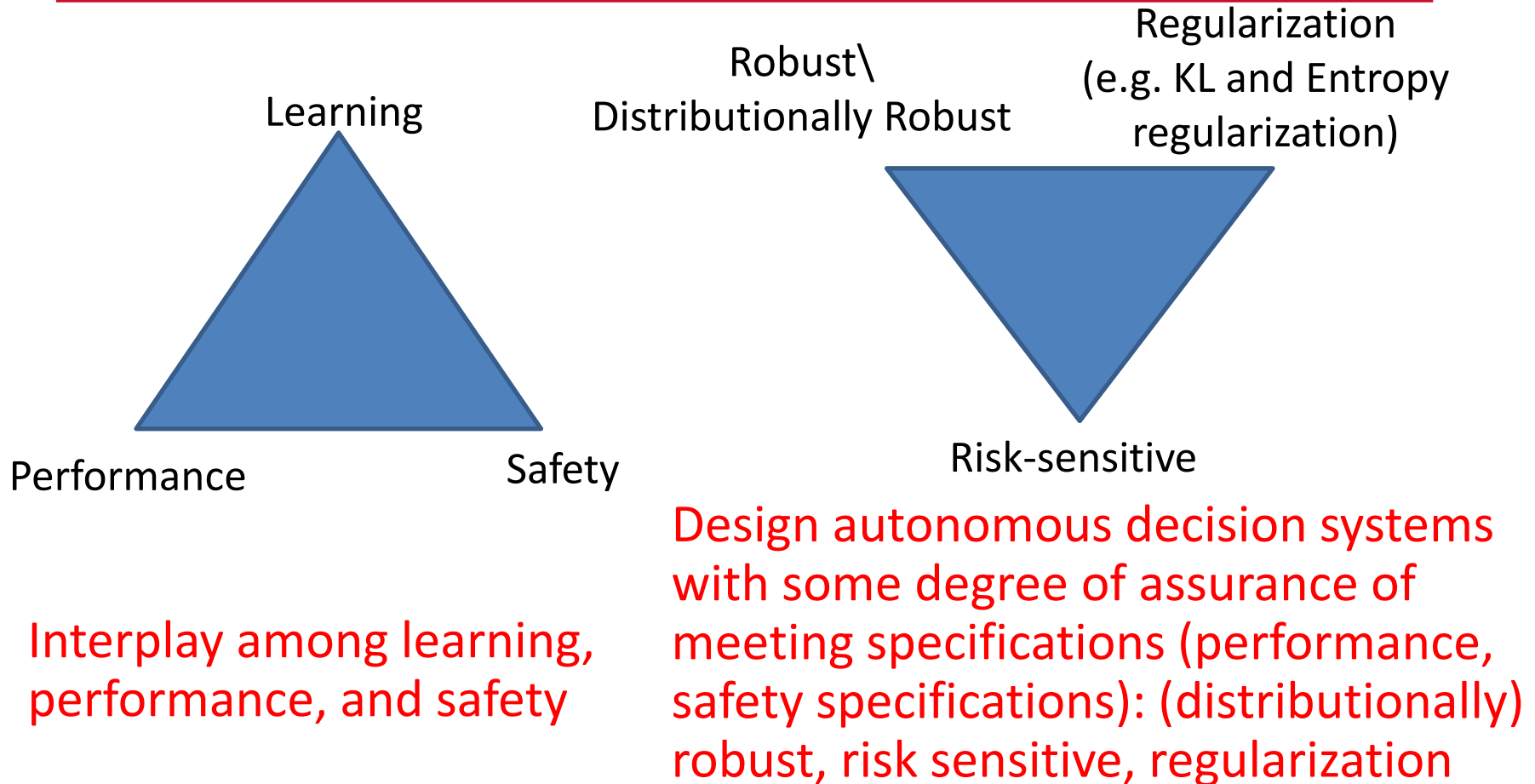
$$\rho_{\pi, P}(\tau) = p_0 \prod_{t=0}^{T-1} \pi(a_t | s_t; \theta) p(s_{t+1} | s_t, a_t) \quad (1)$$

- Prospect Theory** by (Tversky and Kahneman):
“Nobel Memorial Prize in Economic Sciences for his groundbreaking work in applying psychological insights to economic theory, particularly in the areas of judgment and decision-making under uncertainty.”
- Portfolio Optimization and risk:**
Modern portfolio theory (MPT), or mean-variance analysis,
Harry Max Markowitz is an American economist who received the 1989 John von Neumann Theory Prize and the 1990 Nobel Memorial Prize in Economic Sciences.



Known problems with (risk-neutral) RL: brittle. Highly sensitive to noise, etc.

Safe Learning for Autonomy: Robust and Risk Sensitive Control Approach



Risk Measures

$$J(R) = ?$$

① Standard (risk-neutral) Obj. ~ Expected Value

$$\mathbb{E}_{\pi, P}[R] \tag{2}$$

② Risk-Sensitive Obj

- Mean-Variance

$$\mathbb{E}_{\pi, P}[R] - \lambda \mathbb{V}[R] \tag{3}$$

- **Value-at-Risk (VaR)** and **Conditional-Value-at-Risk (CVaR)**

$$\text{CVaR}_p(R) = \mathbb{E}_{\pi, P}[R | R \leq \text{VaR}_p(R)], \tag{4}$$

$$\text{VaR}_p(R) = \inf\{r \in \mathbb{R} : P(R \leq r) > p\} \tag{5}$$

- **Entropic Risk Measure**

$$J_{l_\beta}(R) := (1/\beta) \log \mathbb{E}_{\pi, P}[e^{\beta R}] \tag{6}$$

Coherent and Convex Risk Measures

For all R and $R' \in \mathcal{R}$:

- ① **Monotonicity.** For $R \leq R'$,

$$J(R) \leq J(R')$$

- ② **Translation Invariance.** For $m \in \mathbb{R}$,

$$J(R + m) = J(R) + m$$

- ③ **Convexity.** For $0 \leq \alpha \leq 1$,

$$J(\alpha R + (1 - \alpha)R') \leq \alpha J(R) + (1 - \alpha)J(R')$$

A convex risk measure is called coherent if and only if in addition to the properties (1), (2), and (3), is positive homogeneous, i.e.,

- ④ **Positive Homogeneity.** For $\alpha \geq 0$,

$$J(\alpha R) = \alpha J(R)$$

Dual Representation of Coherent Risk Measures

Proposition (Coherent)

[FS02, FR02] A functional $J : \mathcal{R} \rightarrow \mathbb{R}$ is a coherent risk measure if and only if there exists a subset \mathcal{Q} of $\mathcal{M}_{1,f}$ such that

$$J = \sup_{Q \in \mathcal{Q}} E_Q[R]$$

Moreover, \mathcal{Q} can be chosen as a convex set for which the supremum is attained.

Remark

The set of all finitely additive set functions $Q : \mathcal{T} \rightarrow [0, 1]$ which are normalized to $Q[\mathcal{T}] = 1$.

Dual Representation of Convex Risk Measures

Proposition (Convex)

[FS02] Any convex risk measure $J : \mathcal{R} \rightarrow \mathbb{R}$ is of the form

$$J = \sup_{Q \in \mathcal{Q}} \left\{ E_Q[R] - D(Q) \right\}$$

where D is the minimal penalty function which represents J .

Coherent or Convex?

- **Mean-Variance** is neither Coherent nor Convex.

$$J(R) := \mathbb{E}_{\pi, P}[R] - \lambda \mathbb{V}(R)$$

- **VaR** is not Coherent, but **CVaR** is Coherent.

$$\text{CVaR}_p(R) = \mathbb{E}_{\pi, P}[R | R \leq \text{VaR}_p(R)],$$

$$\text{VaR}_p(R) = \inf\{r \in \mathbb{R} : P(R \leq r) > p\}$$

- **Entropic Risk Measure** is **Convex** (but not Coherent).

$$J_{l_\beta}(R) := (1/\beta) \log \mathbb{E}_{\pi, P}[e^{\beta R}]$$

Risk-Sensitive RL Using Exponential Criteria

Entropic Risk Measure

$$J_{\beta}(\pi) := \frac{1}{\beta} \log \mathbb{E}_{\pi, P} \left[e^{\beta R} \right]$$

positive and negative risk parameters β result in risk-seeking/optimistic ($\beta > 0$) and risk-averse/pessimistic ($\beta < 0$) behavior.

$$\frac{1}{\beta} \log \mathbb{E}_{\pi, P} \left[e^{\beta R} \right] = \mathbb{E}_{\pi, P} [R] + \frac{\beta}{2} \mathbb{V} [R] + \mathcal{O}(\beta^2) \quad (\text{Taylor series})$$

- Interpretation of optimizing exponential criteria by examining it through two theoretical frameworks:
 - ① Large Deviation Theory [NB22a]
 - ② Theory of dual representation of convex risk measures [NB21c]

Risk-Sensitive RL Using Exponential Criteria

Entropic Risk Measure: Large Deviation Theory and Asymptotic Interpretation

Theorem

[NB22a] The maximization of the risk-sensitive exponential criterion, i.e.,

$$J_{l_\beta}(\pi) := \frac{1}{\beta} \log \mathbb{E}[e^{\beta R}]$$

is equivalent to

$$\operatorname{argmax}_{\pi} J_{l_\beta}(\pi) = \lim_{T \rightarrow \infty} \operatorname{argmin}_{\pi} \mathbb{P}[R_T < \psi], \quad \beta < 0 \text{ (risk-averse/pessimistic)}$$

$$\operatorname{argmax}_{\pi} J_{l_\beta}(\pi) = \lim_{T \rightarrow \infty} \operatorname{argmax}_{\pi} \mathbb{P}[R_T > \psi], \quad \beta > 0 \text{ (risk-seeking/optimistic)}$$

Remark

Rate of decay of the left/right tail of the performance metric R .

Risk-Sensitive RL Using Exponential Criteria

Entropic Risk Measure: Duality and Game Theoretic Interpretation

Theorem

[NB21c] The maximization of the risk-sensitive exponential criterion, i.e.,

$$J_{l_\beta} := \frac{1}{\beta} \log \mathbb{E}[e^{\beta R}]$$

is equivalent to

$$\sup_{\pi} \inf_{\hat{\pi}} \left\{ \mathbb{E}[R(\tau)] - \frac{T}{\beta} D_{\text{KL}}(\pi_{\hat{\theta}}(\cdot|s_t), \pi_{\theta}(\cdot|s_t)) \right\}, \beta < 0 \text{ (risk-averse/pessimistic)} \quad (7)$$

$$\sup_{\pi} \sup_{\hat{\pi}} \left\{ \mathbb{E}[R(\tau)] + \frac{T}{\beta} D_{\text{KL}}(\pi_{\hat{\theta}}(\cdot|s_t), \pi_{\theta}(\cdot|s_t)) \right\}, \beta > 0 \text{ (risk-seeking/optimistic)} \quad (8)$$

where $D(Q, P)$ is the KL divergence between the distributions P and Q and is given by $D_{\text{KL}}(P, Q) = \mathbb{E}_Q \left[\log \frac{dQ}{dP} \right]$ if $Q \ll P$ and infinity otherwise.

Risk-Sensitive RL Using Exponential Criteria

Policy Robustness: Definition of Robustness

Definition

Let $\pi(\theta)$ be a given policy and ρ_θ be its associated trajectory distribution given by (1) with transition probabilities P . In addition, let $\hat{\rho}$ be a trajectory distribution generated by $\pi(\theta)$ given a perturbed system of transition probabilities \hat{P} . The policy $\pi(\theta)$ is (ξ, δ, ϵ) -robust if, for $\delta, \epsilon > 0$, and under the condition $D(\hat{\rho}, \rho_\theta) \leq \epsilon$, it holds that

$$\mathbb{P}_{\tau \sim \hat{\rho}} [R(\tau(\theta)) > \xi] \geq 1 - \delta(\xi, \epsilon), \quad (9)$$

where $D(\cdot, \cdot)$ represents the KL divergence.

Risk-Sensitive RL Using Exponential Criteria

Policy Robustness: Robustness Gurantees

Theorem

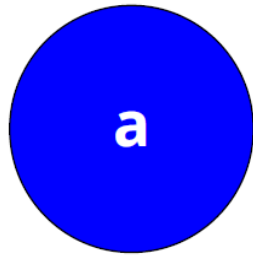
[NMB22] Let $\pi(\theta^*)$ be an optimal policy with respect to J_{l_β} , i.e., $\pi(\theta^*) = \operatorname{argmax}_\theta J_{l_\beta}(\theta)$, and ρ_{θ^*} be its associated trajectory distribution given by (1) with transition probabilities P . In addition, let $\hat{\rho}$ be a trajectory distribution generated by $\pi(\theta)$ given a perturbed system of transition probabilities \hat{P} such that $D(\hat{\rho}, \rho_{\theta^*}) \leq \epsilon$. Then the following inequalities hold:

$$\mathbb{P}_{\tau \sim \hat{\rho}} [R(\tau) \leq \xi] \leq \frac{R_{\max}}{R_{\max} - \xi} \left(1 - \frac{1}{R_{\max}} J_{l_\beta}^* + \frac{\epsilon}{|\beta| R_{\max}} \right), \quad \beta < 0 \text{ (risk-averse)}, \quad (10)$$

$$\mathbb{P}_{\tau \sim \hat{\rho}} [R(\tau) \geq \xi] \leq \frac{1}{\xi} J_{l_\beta}^* + \frac{\epsilon}{\beta \xi}, \quad \beta > 0 \text{ (risk-seeking)}, \quad (11)$$

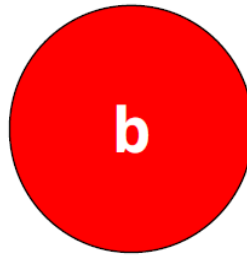
where $J_{l_\beta}^* = \frac{1}{\beta} \ln \mathbb{E}_{\tau \sim \rho_{\theta^*}} [\exp(\beta R(\tau))]$.

Robustness, Risk Sensitivity, and Regularization



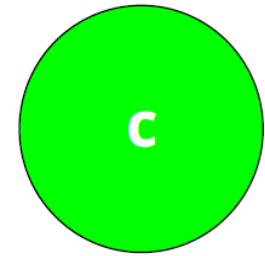
(a) **Risk-Sensitive**

$$\max_{\pi \in \Pi} \frac{1}{\beta} \mathbb{E}_{x \sim \pi} [\exp(\beta R(x))]$$



(b) **Distributionally Robust**

$$\max_{\pi \in \Pi} \inf_{\rho \in \Psi} \mathbb{E}_{x \sim \pi, \zeta \sim \rho} [R(x, \zeta)]$$



(c) **Regularization**

$$\max_{\theta} \mathbb{E}_{\tau \sim \rho_{\theta}} [R(\tau(\theta))] - \frac{1}{\beta} D(\rho_{\theta}, \bar{\rho}_{\theta})$$

Figure: There is an intricate connection between risk-sensitive, distributionally robust, and regularized objectives.

Robustness, Risk Sensitivity, and Regularization

$$J_{l_\beta}(\pi) := \frac{1}{\beta} \log \mathbb{E}_{\tau \sim \rho_{\pi, P}} \left[e^{\beta R_T(\tau)} \right] \quad (\text{Entropic Risk Measure})$$

$$J_{DR}(\pi) := \min_{\hat{P} \in \mathcal{U}(P)} \mathbb{E}_{\tau \sim \rho_{\pi, \hat{P}}} \left[R_T(\tau) \right] \quad (\text{Distributionally Robust})$$

$$J_{KL}^\lambda(\theta, \theta_0) := \mathbb{E}_{\tau \sim \rho_\theta} \left[R(\tau) - \lambda D_{KL} \left(\pi_\theta(\cdot | s_t), \pi_{\theta_0}(\cdot | s_t) \right) \right] \quad (\text{Regularization})$$

$$J_{ent}(\theta) := \mathbb{E}_{\tau \sim \rho_\theta} \left[R(\tau) \right] + \lambda \mathbb{E}_{s_t \sim p(s_{t+1} | s_t, a_t)} \left[\sum_{t=1}^T \mathcal{H}^\pi(\cdot | s_t) \right]$$

Remark

Note that for a choice of uniform distribution as the reference policy, the KL-regularized objective is equivalent to the maximum entropy objective (up to a constant). Therefore, we only consider the more general case of KL-regularized objective from hereon.

Robustness, Risk Sensitivity, and Regularization

Coherent Entropic Risk Measure [FK11] – A bridge to distributionally robust

$$J_{l_\beta}(\pi) := \frac{1}{\beta} \log \mathbb{E}_{\tau \sim \rho_{\pi, P}} \left[e^{\beta R_T(\tau)} \right]$$

$$J_c^{\beta, \alpha}(\pi) := J_{l_\beta}(\pi) - \frac{1}{|\beta|} \ln(\alpha) \quad 0 < \alpha < 1$$

$$J_{c-}^{\alpha}(\pi) := \sup_{\beta < 0} J_c^{\beta, \alpha}(\pi)$$

$$J_{c-}^{\alpha}(\pi) := \inf_{\hat{\rho} \in \{\hat{\rho} \ll \rho : D(\hat{\rho}, \rho) < -\ln \alpha\}} \mathbb{E}_{\tau \sim \hat{\rho}} \left[R_T(\tau) \right] \quad (\text{Entropic Value at Risk [AJ12]})$$

* It is evident that the KL-regularized and KL-constrained algorithms such **TRPO** and **PPO** are attempts to iteratively optimize the convex and coherent risk-sensitive criterion.

Robustness, Risk Sensitivity, and Regularization

Equivalences: Distributionally Robust

Theorem

[NB22a] The coherent risk-sensitive exponential criterion with a positive risk parameter $\beta < 0$ (risk-aversion) or $\beta > 0$ (risk-seeking) is equivalent to a distributionally robust objective with the uncertainty set

$$\mathcal{U}(\mathcal{P}) = \left\{ \hat{\mathcal{P}} : \mathbb{E}_{\tau \sim \rho_{\pi, \hat{\mathcal{P}}}} \left[D(\hat{\mathcal{P}}, \mathcal{P}) \right] \leq \frac{-\ln \alpha}{T} \right\}.$$

That is to say,

$$J_{c-}^{\alpha} := \min_{\hat{\mathcal{P}} \in \mathcal{U}(\mathcal{P})} \mathbb{E}_{\tau \sim \rho_{\pi, \hat{\mathcal{P}}}} \left[R_T(\tau) \right], \quad J_{c+}^{\alpha} = \max_{\hat{\mathcal{P}} \in \mathcal{U}(\mathcal{P})} \mathbb{E}_{\tau \sim \rho_{\pi, \hat{\mathcal{P}}}} \left[R_T(\tau) \right]$$

where J_{c+}^{α} and J_{c-}^{α} are the coherent risk-sensitive exponential criterion for $\beta > 0$ and $\beta < 0$, respectively.

Robustness, Risk Sensitivity, and Regularization

Equivalences: Regularization

Recall $J_{l_\beta}(\pi) := \frac{1}{\beta} \log \mathbb{E}_{\tau \sim \rho_{\pi, P}} \left[e^{\beta R_T(\tau)} \right]$. By the dual representation theorem of Convex risk measures

$$J_{l_\beta}(\theta) = \sup_{\hat{\theta}} \left\{ \mathbb{E}_{\tau \sim \rho_{\hat{\theta}}} [R(\tau)] - \frac{1}{\beta} D_{\text{KL}}(\rho_{\hat{\theta}}(\tau), \rho_{\theta}(\tau)) \right\}$$

By noting the definition of the trajectory distribution, we have

$$D_{\text{KL}}(\rho_{\hat{\theta}}(\tau), \rho_{\theta}(\tau)) = T \mathbb{E}_{\tau \sim \rho_{\hat{\theta}}} \left[D_{\text{KL}}(\pi_{\hat{\theta}}(\cdot | s_t), \pi_{\theta}(\cdot | s_t)) \right]$$

Thus,

$$J_{l_\beta}(\theta) = \sup_{\hat{\theta}} \mathbb{E}_{\tau \sim \rho_{\hat{\theta}}} \left[R(\tau) - \frac{T}{\beta} D_{\text{KL}}(\pi_{\hat{\theta}}(\cdot | s_t), \pi_{\theta}(\cdot | s_t)) \right]$$

$$J_{l_\beta}(\theta) = \sup_{\hat{\theta}} J_{\text{KL}}^{\frac{T}{\beta}}(\hat{\theta}, \theta) \quad (12)$$

Robustness, Risk Sensitivity, and Regularization

Developing risk-sensitive RL algorithms using regularization equivalence

Theorem

[NB21d] The maximization of the exponential criteria $J_{l_\beta}(\theta)$ with a positive risk parameter $\beta > 0$ is equivalent to the maximization of the KL-regularized objective $J_{KL}(\hat{\theta}, \theta)$ jointly over the policy parameters $\hat{\theta}$ and the reference policy parameters θ , that is,

$$\operatorname{argmax}_{\theta} J_{l_\beta}(\theta) = \operatorname{argmax}_{\hat{\theta}, \theta} J_{KL}^{\frac{T}{\beta}}(\hat{\theta}, \theta)$$

where $J_{KL}^{\frac{T}{\beta}}(\hat{\theta}, \theta) = \mathbb{E}_{\tau \sim \rho_{\hat{\theta}}} \left[R(\tau) - \frac{T}{\beta} D_{KL}(\pi_{\hat{\theta}}(\cdot | s_t), \pi_{\theta}(\cdot | s_t)) \right]$ is the KL-regularized objective with reference policy parameter θ and the regularization weight T/β with T being the time horizon.

Risk-Sensitive RL Algorithms Using MDP

REINFORCE – A Policy Gradient Algorithm

$$J(\theta) := \mathbb{E}_{\tau \sim \rho_\theta} [R]$$

$$\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)} \quad (\text{Gradient Ascent})$$

$$\nabla J(\theta) \propto \mathbb{E}_{\pi_\theta} \left[R \sum_{t=0}^{|\tau|-1} \nabla \log \pi_\theta(a_t | s_t) \right] \quad (\text{Policy Gradient Theorem (Sutton et. al.)})$$

$$\nabla J(\theta) \propto \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{|\tau|-1} R_t \nabla \log \pi_\theta(a_t | s_t) \right]$$

$$\theta_{k+1} = \theta_k + \alpha R_t \frac{\nabla \pi_\theta(a_t | s_t)}{\pi_\theta(a_t | s_t)}. \quad (\text{REINFORCE (Williams)})$$

Risk-Sensitive RL Algorithms Using MDP

REINFORCE – A Policy Gradient Algorithm (using baseline)

To reduce variance

$$\nabla J(\theta) \propto \mathbb{E}_{\pi_{\theta}} \left[\sum_{t=0}^{|\tau|-1} \left(R_t - b(s_t) \right) \nabla \log \pi_{\theta}(a_t | s_t) \right]$$

As we will discuss, a particularly convenient property of using exponential criteria is that it alleviates the need for such approaches [NB21c].

Risk-Sensitive RL Algorithms Using MDP

Policy Gradient Algorithms (using function approximation)

Let the reward-to-go $R_t := \sum_{t'=t}^{|\tau|-1} \gamma^{t'-t} r(s_{t'}, a_{t'})$ and the value function $V^{\pi_\theta}(s_t) := \mathbb{E}_{\pi_\theta}[R_t | s_t]$. Then, we have

$$V^{\pi_{\theta^*}}(s_t) = \mathbb{E}_{\pi_{\theta^*}} \left[r(s_t, a_t) + \gamma V^{\pi_{\theta^*}}(s_{t+1}) \mid s_t \right] \quad (\text{Bellman's equation})$$

where $a_t \sim \pi_{\theta^*}(\cdot | s_t)$.

Bellman's equation is a **contraction** map. This has led to stochastic approximation algorithms that try to asymptotically minimize the mean-squared error

$$\min_{\theta} \mathbb{E}_{\pi_{\theta}} \left[\| r(s_t, a_t) + \gamma V^{\pi_{\theta}}(s_{t+1}) - V^{\pi_{\theta}}(s_t) \|^2 \mid s_t \right]$$

Risk-Sensitive RL Algorithms Using MDP

Temporal-Difference Methods (actor-critic algorithms)

$$\begin{cases} \theta_{t+1} = \theta_t + \alpha \left(\hat{R}_t - V(s_t; w_t) \right) \frac{\nabla \pi_{\theta_t}(a_t | s_t)}{\pi_{\theta_t}(a_t | s_t)} \\ w_{t+1} = w_t - \bar{\alpha} \nabla J_c(s_t; w_t, \theta_t) \end{cases} \quad (\text{Risk-neutral TD})$$

$$J_c(s_t; w_t, \theta_t) := \|\hat{R}_t - V(s; w_t)\|^2$$

$$\hat{R}_t := r(s_t, a_t) + \gamma V(s_{t+1}, w_t) \simeq R_t$$

The **'actor'** implements a policy gradient algorithm based on a function approximation of the (risk-neutral) reward-to-go, estimated by the **'critic'** based on Bellman's equation associated with (risk-neutral) dynamic programming.

Risk-Sensitive RL Algorithms Using MDP

Risk-Sensitive REINFORCE [NB21a, NMB22]

$$J_{\beta}(\theta) := \mathbb{E}_{\pi_{\theta}} \left[\beta e^{\beta R} \right]$$
$$\nabla J_{\theta}(\theta) \propto \frac{1}{\beta} \mathbb{E}_{\tau \sim p_{\theta}} \left[\sum_{t=0}^{|\tau|-1} e^{\beta R_t} \nabla \log \pi_t(\theta) \right]$$
$$\theta_{t+1} = \theta_t + \frac{\alpha}{\beta} e^{\beta R_t} \frac{\nabla \pi(a_t | s_t; \theta)}{\pi(a_t | s_t; \theta)} \quad (\text{Risk-Sensitive REINFORCE (Noorani \& Baras)})$$

Remark

- The update rule is not proportional to the reward-to-go $R_t := \sum_{t'=t}^{|\tau|-1} \gamma^{t'-t} r(s_{t'}, a_{t'})$, but to the exponential

$$\beta e^{\beta R_t} = \beta \prod_{t'=t}^{|\tau|-1} \exp\{\gamma^{t'-t} \beta r(s_{t'}, a_{t'})\}$$

- The risk-sensitive objective can be understood to provide a natural baseline. This has been shown in [NB21b] and holds for the temporal-difference case as well.

Risk-Sensitive RL Algorithms Using MDP

Risk-Sensitive Actor-Critic

Define the risk-sensitive value function of a policy π as

$$V_{\beta}^{\pi}(s_t) := \beta \mathbb{E} \left[e^{\beta \sum_{l=t}^{\infty} \gamma^{l-t} r(s_l, a_l)} | s_t \right], \quad a_l \sim \pi(\cdot | s_l).$$

We further define:

$$\bar{V}_{\beta}^{\pi}(s_t) := \frac{1}{\beta} V_{\beta}^{\pi}(s_t) = \mathbb{E} \left[e^{\beta \sum_{l=t}^{\infty} \gamma^{l-t} r(s_l, a_l)} | s_t \right]$$

where $a_l \sim \pi(\cdot | s_l)$ and by definition, $\bar{V}_{\beta}^{\pi}(\cdot) \geq 0$. The following relationship holds:

$$V_{\beta}^*(s_t, w_t) = \max_a e^{\beta r(s_t, a)} \mathbb{E} \left[e^{(V_{\beta}^*(s_{t+1}, w_t))^{\gamma}} | s_t \right].$$

Risk-Sensitive RL Algorithms Using MDP

Risk-Sensitive (Online) Actor-Critic [NMB22, NMB23]

The following actor-critic learning approach can be constructed [NMB23]:

$$\begin{cases} \theta_{t+1} = \theta_t + \alpha |\beta| (R_t^\beta - \bar{V}_\beta(s_t; w_t)) \frac{\nabla \pi(a_t | s_t; \theta_t)}{\pi(a_t | s_t; \theta_t)} \\ w_{t+1} = w_t - \bar{\alpha} \nabla J_r(s_t; w_t, \theta_t) \end{cases} \quad (\text{Risk-Sensitive TD})$$

$$R_t^\beta = \exp\{\beta r(s_t, a_t) + \gamma \ln \bar{V}_\beta(s_{t+1}; w_t)\}$$

$$J_r(s_t; w_t, \theta_t) = \|e^{\beta r(s_t, a_t) + \gamma \ln \bar{V}_\beta(s_{t+1}; w_t)} - \bar{V}_\beta(s_t; w_t)\|^2$$

The **'actor'** implements a policy gradient algorithm based on a function approximation of the exponential of the reward-to-go, estimated by the **'critic'** based on the **multiplicative Bellman's equation** associated with risk-sensitive dynamic programming.

Risk-Sensitive RL Algorithms Using MDP

Algorithm: REINFORCE

Algorithm REINFORCE

- 1: **Input:** a differentiable policy parametrization $\pi(a|s; \theta)$.
 - 2: **Algorithm parameters:** step-size $\alpha > 0$, discount factor $\gamma > 0$, and the risk parameter β .
 - 3: **Initialization:** Initialize policy parameters $\theta \in \mathbb{R}^d$ (e.g. to 0).
 - 4: **while True do**
 - 5: **Generate an episode following the policy $\pi(\cdot|\cdot; \theta)$, i.e., $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t; \theta)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$, generating a sequence of state-actions $s_0, a_0, \dots, s_{|\tau|-1}, a_{|\tau|-1}$**
 - 6: **for $t = 0$ to $|\tau| - 1$ do**
 - 7: $\hat{R} \leftarrow \sum_{t'=t}^{|\tau|-1} \gamma^{t'-t} r_{t'}$
 - 8: $\theta_{t+1} \leftarrow \theta_t + \alpha \gamma^t \hat{R} \nabla \log \pi(a_t|s_t; \theta)$
 - 9: **end for**
 - 10: **end while**
-

Risk-Sensitive RL Algorithms Using MDP

Algorithm: REINFORCE with baseline

Algorithm REINFORCE with Baseline

- 1: **Input:** a differentiable policy parametrization $\pi(a|s; \theta)$.
 - 2: **Algorithm parameters:** step-sizes $\alpha > 0$ and $\bar{\alpha} > 0$, discount factor $\gamma > 0$, and the risk parameter β .
 - 3: **Initialization:** Initialize policy parameters $\theta \in \mathbb{R}^d$ (e.g. to 0) and value parameters $w \in \mathbb{R}^{d'}$.
 - 4: **while True do**
 - 5: **Generate an episode following the policy $\pi(\cdot|\cdot; \theta)$, i.e., $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t; \theta)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$, generating a sequence of state-actions $s_0, a_0, \dots, s_{|\tau|-1}, a_{|\tau|-1}$**
 - 6: **for $t = 0$ to $|\tau| - 1$ do**
 - 7: $\hat{R} \leftarrow \sum_{t'=t}^{|\tau|-1} \gamma^{t'-t} r_{t'}$
 - 8: $\theta_{t+1} \leftarrow \theta_t + \alpha \gamma^t (\hat{R} - V(s_t; w_t)) \nabla \log \pi(a_t|s_t; \theta)$
 - 9: $w_t = w_t + \bar{\alpha} \gamma^t (\hat{R} - V(s_t; w_t)) \nabla V(s_t; w_t)$
 - 10: **end for**
 - 11: **end while**
-

Risk-Sensitive RL Algorithms Using MDP

Algorithm: Online Actor-Critic

Algorithm Online Actor-Critic

- 1: **Input:** a differentiable policy parametrization $\pi(a|s; \theta)$.
 - 2: **Algorithm parameters:** step-sizes $\alpha > 0$ and $\bar{\alpha} > 0$, discount factor $\gamma > 0$, and the risk parameter β .
 - 3: **Initialization:** Initialize policy parameters $\theta \in \mathbb{R}^d$ and value parameters $w \in \mathbb{R}^{d'}$ (e.g. to 0) (e.g. to 0).
 - 4: **while True do**
 - 5: **for** $t = 0$ **to** $|\tau| - 1$ **do**
 - 6: **Starting at an initial state** s_0 , **take an action by following the current policy** $a_t \sim \pi(\cdot|s_t; \theta)$ **and observe the successor state** $s_{t+1} \sim p(\cdot|s_t, a_t)$, **and the reward** r_t
 - 7: $\hat{R} \leftarrow r_t + \gamma V(s_{t+1}; w_t)$
 - 8: $\theta_{t+1} \leftarrow \theta_t + \alpha \gamma^t (\hat{R} - V(s_t; w_t)) \nabla \log \pi(a_t|s_t; \theta)$
 - 9: $w_{t+1} = w_t + \bar{\alpha} \gamma^t (\hat{R} - V(s_t; w_t)) \nabla V(s_t; w_t)$
 - 10: **end for**
 - 11: **end while**
-

Risk-Sensitive RL Algorithms Using MDP

Algorithm: Risk-sensitive REINFORCE [NB21a, NMB22]

Algorithm Risk-sensitive REINFORCE

- 1: **Input:** a differentiable policy parametrization $\pi(a|s; \theta)$.
 - 2: **Algorithm parameters:**
step-size $\alpha > 0$, discount factor $\gamma > 0$, and the risk parameter β .
 - 3: **Initialization:** Initialize policy parameters $\theta \in \mathbb{R}^d$ (e.g. to 0).
 - 4: **while True do**
 - 5: **Generate an episode following the policy** $\pi(\cdot|\cdot; \theta)$,
 i.e., $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t; \theta)$ **and** $s_{t+1} \sim p(\cdot|s_t, a_t)$,
 generating a sequence of state-actions $s_0, a_0, \dots, s_{|\tau|-1}, a_{|\tau|-1}$
 - 6: **for** $t = 0$ **to** $|\tau| - 1$ **do**
 - 7: $\hat{R} \leftarrow \sum_{t'=t}^{|\tau|-1} \gamma^{t'-t} r_{t'}$
 - 8: $\theta_{t+1} \leftarrow \theta_t + \alpha \gamma^t \frac{1}{\beta} e^{\beta \hat{R}} \nabla \log \pi(a_t|s_t; \theta_t)$
 - 9: **end for**
 - 10: **end while**
-

Risk-Sensitive RL Algorithms Using MDP

Algorithm: Risk-sensitive Online Actor-Critic [NMB22, NMB23]

Algorithm Risk-sensitive Online Actor-Critic

- 1: **Input:** a differentiable policy parametrization $\pi(a|s; \theta)$.
 - 2: **Algorithm parameters:**
step-sizes $\alpha > 0$ and $\bar{\alpha} > 0$, discount factor $\gamma > 0$, and the risk parameter β .
 - 3: **Initialization:** Initialize policy parameters $\theta \in \mathbb{R}^d$
and value parameters $w \in \mathbb{R}^{d'}$ (e.g. to 0).
 - 4: **while True do**
 - 5: **for** $t = 0$ **to** $|\tau| - 1$ **do**
 - 6: **Starting at an initial state** s_0 ,
 take an action by following the current policy $a_t \sim \pi(\cdot|s_t; \theta)$
 and observe the successor state $s_{t+1} \sim p(\cdot|s_t, a_t)$, **and the reward** r_t
 - 7: $\hat{R}_\beta \leftarrow \beta r_t + \gamma \ln \bar{V}_\beta(s_{t+1}; w_t)$
 - 8: $\theta_{t+1} \leftarrow \theta_t + \alpha \gamma^t \frac{1}{|\beta|} (e^{\hat{R}_\beta} - \bar{V}_\beta(s_t; w_t)) \nabla \log \pi(a_t|s_t; \theta)$
 - 9: $w_{t+1} \leftarrow w_t + \bar{\alpha} \gamma^t (e^{\hat{R}_\beta} - \bar{V}_\beta(s_t; w_t)) \nabla \bar{V}_\beta(s_t; w_t)$
 - 10: **end for**
 - 11: **end while**
-

Risk-Sensitive RL Algorithms Using MDP

Simulation results (Training and Testing behavior): Risk-Sensitive REINFORCE (Double Pendulum)

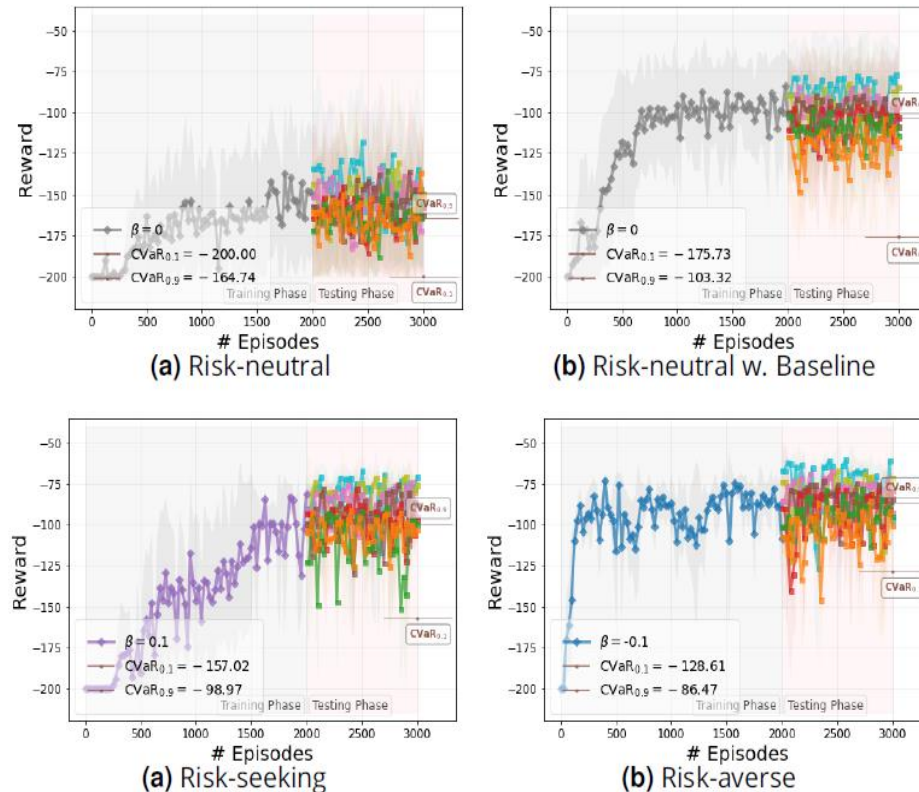


Figure: Training and testing behavior of the risk-neutral REINFORCE (Alg. 1) and risk-neutral REINFORCE with baseline (Alg. 2) algorithms against the proposed risk-sensitive R-REINFORCE algorithm (Alg. 4) for $\beta = -0.1$ and $\beta = +0.1$ in the Acrobot problem. Average reward, $CVaR_{0.1}$, and $CVaR_{0.9}$ values (for $l = 1.0$) are computed over 10 independent training and testing runs with different random seeds.

Risk-Sensitive RL Algorithms Using MDP

Simulation results (Robustness): Risk-Sensitive REINFORCE (Double Pendulum)

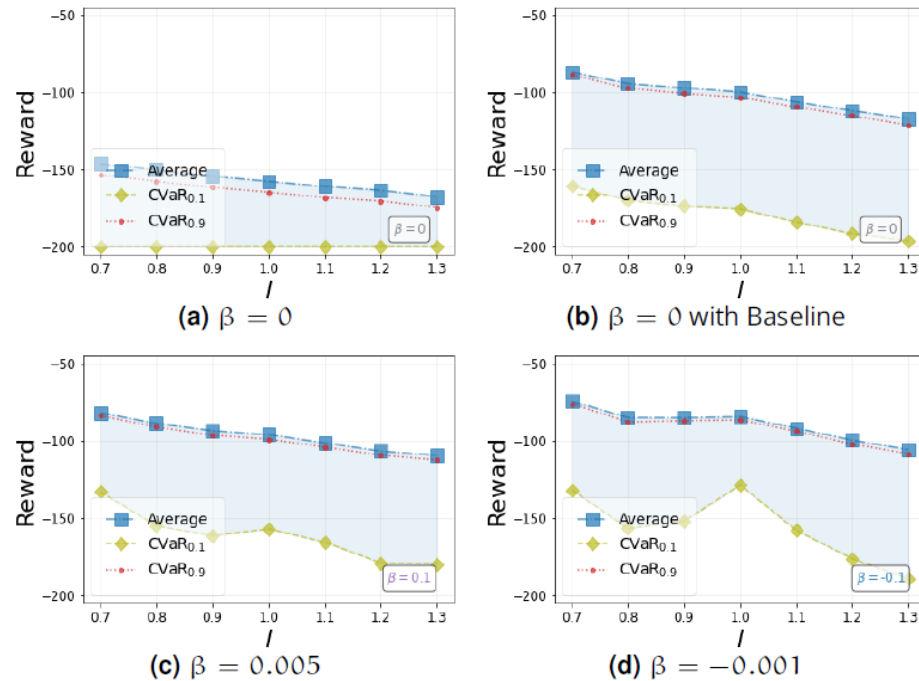
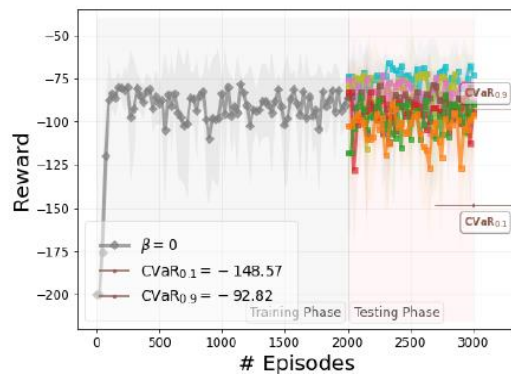


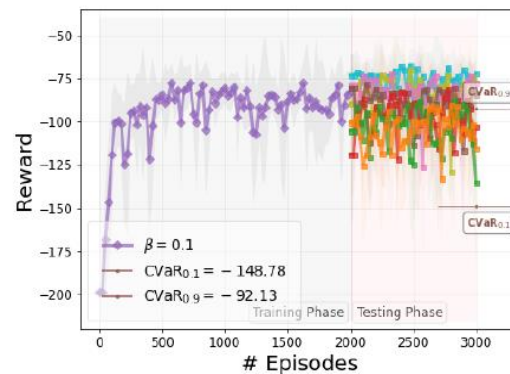
Figure: Robustness of risk-neutral REINFORCE (Alg. 1), risk-neutral REINFORCE with baseline (Alg. 2), and risk-sensitive R-REINFORCE (Alg. 4) algorithms in the Acrobot environment with respect to varying pole length. The training environment is modeled with pole length $l = 1.0$. The testing environments have perturbed pole length values of $l \in [0.7, 1.3]$. Average reward, CVaR_{0.1}, and CVaR_{0.9} values are computed over 10 independent training and testing runs with different random seeds.

Risk-Sensitive RL Algorithms Using MDP

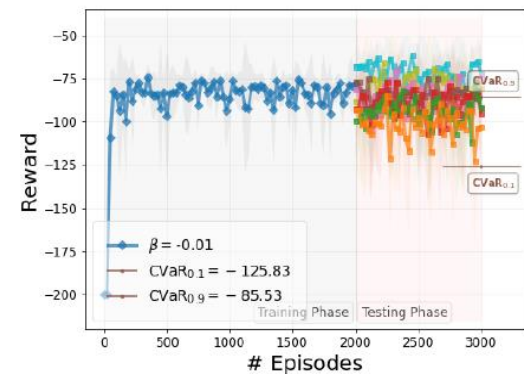
Simulation results (Training and Testing behavior): Risk-Sensitive Actor-Critic (Double Pendulum)



(a) Risk-neutral



(b) Risk-seeking

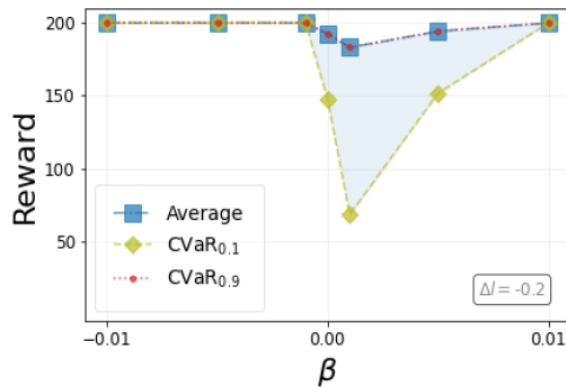


(c) Risk-averse

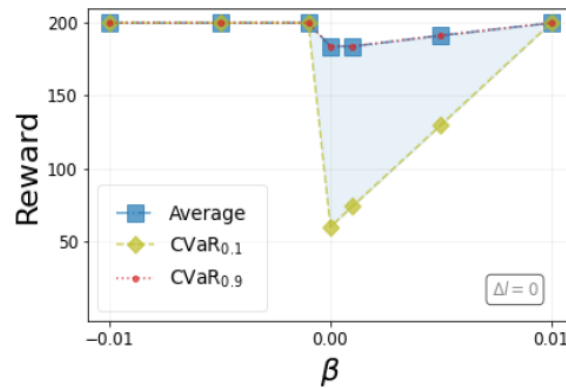
Figure: Training and testing behavior of the risk-neutral Online Actor-Critic (OAC) (Alg. 3) algorithm against the proposed risk-sensitive R-AC algorithm (Alg. 5) for $\beta = -0.01$ and $\beta = +0.1$ in the Acrobot problem. Average reward, $\text{CVaR}_{0.1}$, and $\text{CVaR}_{0.9}$ values (for $l = 1.0$) are computed over 10 independent training and testing runs with different random seeds.

Risk-Sensitive RL Algorithms Using MDP

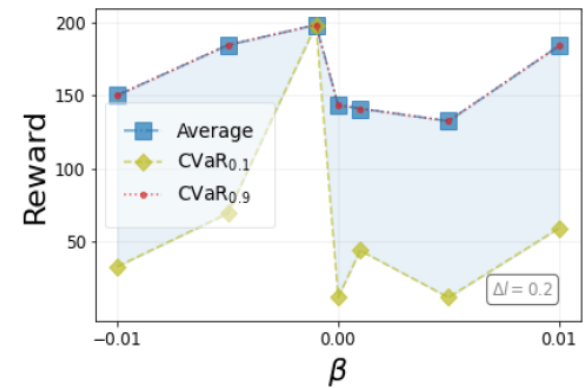
Simulation results (Robustness): Risk-Sensitive Actor-Critic (Double Pendulum)



(a) $l = 0.3$



(b) $l = 0.5$



(c) $l = 0.7$

Figure: Sensitivity analysis of the risk-sensitive R-AC algorithm (Alg. 5) with respect to the risk-sensitive parameter $\beta \in [-0.01, 0.01]$ in the Cart-Pole problem. $\beta = 0$ corresponds to the risk-neutral Online Actor-Critic (OAC) (Alg. 3). The training environment is modeled with pole length $l = 0.5$. Average reward, CVaR_{0.1}, and CVaR_{0.9} values for testing environments with $l \in \{0.3, 0.5, 0.7\}$ are computed over 10 independent training and testing runs with different random seeds.

Risk-Sensitive Safety Filters

- The idea is to decouple optimality and safety by independently determining safe and optimal control laws. Before applying an optimal, but potentially unsafe control input to the real system, its safety is checked, such that a safe control input can be chosen instead.

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\omega}_k), \quad (23)$$

$$\boldsymbol{\pi}_{\text{safe}}^*(\mathbf{x}) = \underset{\mathbf{u} \in \mathbb{U}}{\operatorname{argmin}} \|\boldsymbol{\pi}^*(\mathbf{x}) - \mathbf{u}\| \quad (24)$$

$$\text{such that } \mathbf{u} \text{ is safe.} \quad (25)$$

- When the dynamics of the systems are known to exhibit a control-affine structure, control barrier functions (CBF) can be effectively employed to address this challenge.

Risk-Sensitive Safety Filters

- 1 **Risk-sensitive safety conditions:** In order to ensure the probabilistic satisfaction of state constraints, we introduce cost functions that allow us to express safety through risk-sensitive conditions on the cumulative cost along system trajectories. These conditions reveal an intuitive relationship between the risk-aversion and safety probability.
- 2 **Safe policies and value functions through RL:** Based on these results, we develop an approach for determining safe policies and corresponding safety value functions using common techniques from RL. The success of the proposed approach is shown to be guaranteed under weak assumptions relating to the controllability properties of the system dynamics.
- 3 **Inhibitory control through safety filters:** By enforcing the satisfaction of the derived safety conditions with the learned value function online, we obtain a risk-sensitive safety filter. Moreover, we prove it to inherit probabilistic safety guarantees from the safe policy obtained through RL.

Risk-Sensitive Safety Filters

Risk-sensitive safety conditions

Proposition

[LNHB23] Consider a cost function $c(\cdot)$ satisfying the conditions that

$$c(x) \geq \hat{c} \quad \forall x \in \mathbb{X}_{\text{unsafe}}. \quad (26)$$

If there exist constants $\xi, \beta \in \mathbb{R}_+$ with $\xi < \bar{\xi}$, such that

$$\mathbb{R}_\beta[V_\pi(x^+)] \leq \xi, \quad \forall x \in \mathbb{V}_\pi^{\bar{\xi}}, \quad \mathbb{R}_\beta[C] := \frac{1}{\beta} \log(\mathbb{E}[\exp(\beta C)]) \quad (27)$$

holds for $x^+ = f(x, \pi(x), \omega)$, then, $\pi(\cdot)$ is δ -safe, i.e., $\mathcal{P}(f(x, \pi(x), \omega) \in \mathbb{V}) \geq 1 - \delta$, on $\mathbb{V}_\pi^{\bar{\xi}}$ with

$$\delta = \exp(\beta(\xi - \bar{\xi})). \quad (28)$$

This result provides a straightforward condition, which merely requires the evaluation of the risk operator and the computation of the cumulative cost. Moreover, it offers a simple expression for the probability of safety.

Risk-Sensitive Safety Filters

Safe backup policies via RL ($\pi_{safe} = \arg \min_{\pi \in \Pi} E_x[V_\pi(x)]$)

Theorem

[LNHB23] Consider a cost function $c(\cdot)$ satisfying (26) and assume that there exist a policy $\tilde{\pi}(\cdot)$ and constants $\theta_1, \theta_2 \in \mathbb{R}_+$ with $\theta_1 < 1/(1-\gamma)$ such that

$$V_{\tilde{\pi}}(x) \leq \theta_1 c(x) + \theta_2, \quad \forall x \in \mathbb{X} \quad (29)$$

is satisfied. Moreover, assume there exist constants $\theta_3, \theta_4 \in \mathbb{R}_{0,+}$ such that

$$V_{\tilde{\pi}}(x) \geq \theta_3 c(x) + \theta_4, \quad \forall x \in \mathbb{X} \quad (30)$$

holds for all policies $\pi(\cdot)$. If

$$\hat{c} > \frac{\theta_2}{\theta_3(\theta_1(\gamma - 1) + 1)} - \frac{\theta_4}{\theta_3} \quad (31)$$

holds, then, the policy (??) is δ^* -safe on \mathbb{V}_{ξ^*} with $\delta^* = \exp(\beta^*(\xi^* - \bar{\xi}))$, where

$$\beta^*, \xi^* = \underset{\beta \in \mathbb{R}_+, \xi \in \mathbb{R}_+}{\operatorname{argmin}} \exp(\beta(\xi - \bar{\xi})) \quad \text{s.t. } \xi < \bar{\xi} \quad (27) \text{ holds.} \quad (32a)$$

Risk-Sensitive Safety Filters

Risk-Sensitive Inhibitory Control for Safe RL

we employ the **risk-sensitive filter**

$$\pi_{\text{safe}}^*(x) = \operatorname{argmin}_{u \in \mathcal{U}} \|\pi^*(x) - u\| \quad (33a)$$

$$\text{s.t. } \mathbb{R}_\beta[V_{\pi_{\text{safe}}}(f(x, u, \omega))] \leq \xi^* \quad (33b)$$

Theorem

[LNHB23] Consider a cost function $c(\cdot)$ satisfying (26) and a threshold \hat{c} , for which (31) holds. Moreover, assume that there exists a policy $\tilde{\pi}(\cdot)$ satisfying (29) with $\theta_1 < 1/(1-\gamma)$ for all $x \in \mathbb{X}_{\text{safe}}$. Then, the safety filtered policy (33) is δ^* -safe on $\mathbb{V}_{\pi_{\text{safe}}}^{\xi^*}$ with $\delta^* = \exp(\beta^*(\xi^* - \bar{\xi}))$, where β^* and ξ^* are defined in (32).

Risk-Sensitive Safety Filters

Simulation results

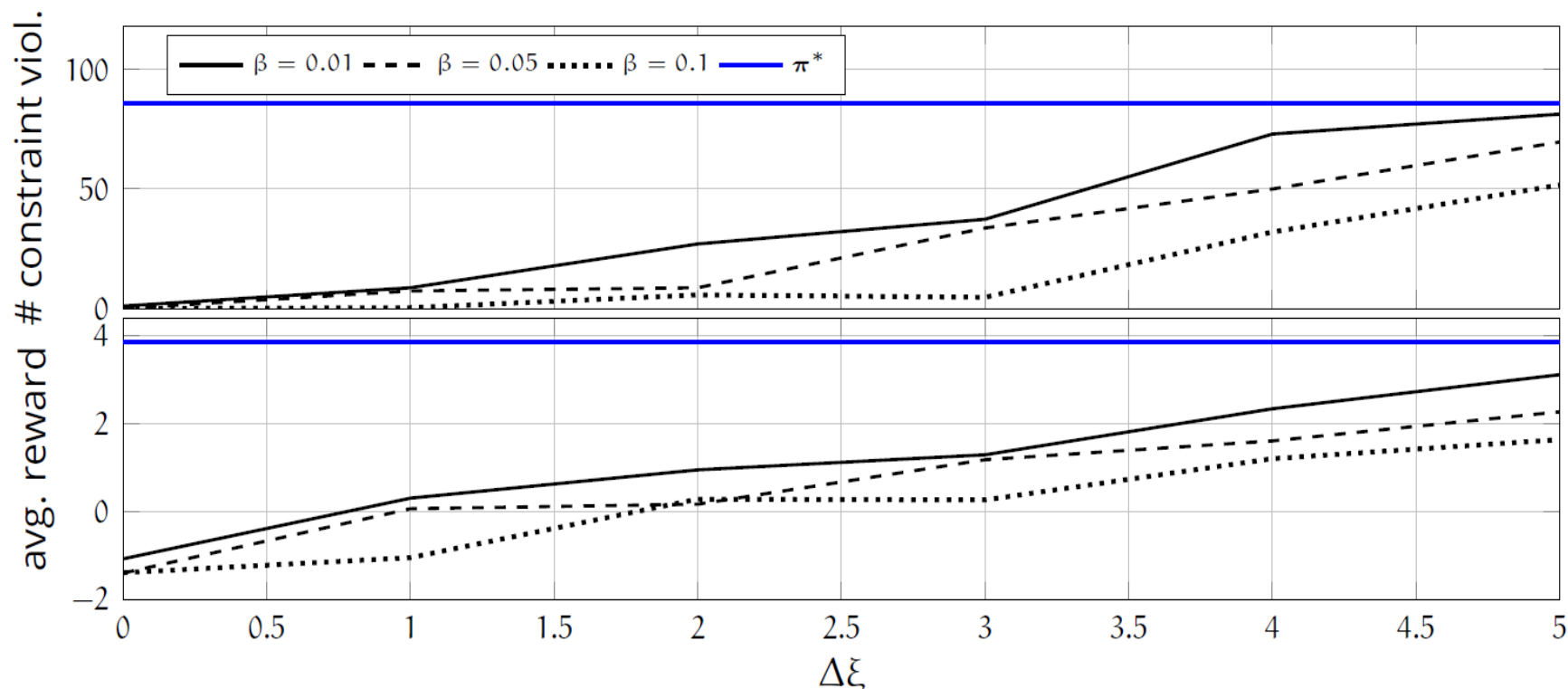


Figure: Number of constraint violations and average rewards in dependency on the safety constraint threshold $\xi = 521 + \Delta\xi$ and the risk-sensitivity β . Reducing β and increasing ξ have a similar effect of admitting more risky behavior in the response inhibition, such that the number of constraint violations and the average reward increase.

Risk-Sensitive Safety Filters

Summary of results

- Inspired by the psychological concept of inhibitory control, this paper proposes a risk-sensitive method for rendering arbitrary policies safe.
- This method is based on the introduction of cost functions, such that state constraints can be expressed in terms of value functions.
- We show that this formulation allows us to employ standard RL techniques for obtaining policies that their only goal is to ensure safety.
- Based on the determined safe policies and corresponding value functions, a risk-sensitive safety constraint is employed to enforce the satisfaction of state constraints online. Thereby, risk-sensitive inhibitory control is realized and its effectiveness is demonstrated in simulations.

References

- Amir Ahmadi-Javid, *Entropic Value-at-Risk: A New Coherent Risk Measure*, Journal of Optimization Theory and Applications **155** (2012), no. 3, 1105–1123.
- Hans Föllmer and Thomas Knispel, *Entropic Risk Measures: Coherence vs. Convexity, Model Ambiguity and Robust Large Deviations*, Stochastics and Dynamics **11** (2011), no. 02n03, 333–351.
- Marco Frittelli and Emanuela Rosazza Gianin, *Putting Order in Risk Measures*, Journal of Banking & Finance **26** (2002), no. 7, 1473–1486.
- Hans Föllmer and Alexander Schied, *Convex Measures of Risk and Trading Constraints*, Finance and stochastics **6** (2002), no. 4, 429–447.
- Matthew R James, John S Baras, and Robert J Elliott, *Risk-sensitive Control and Dynamic Games for Partially Observed Discrete-time Nonlinear Systems*, IEEE transactions on automatic control **39** (1994), no. 4, 780–792.

References

Armin Lederer, Erfaun Noorani, Sandra Hirche, and John S Baras, *Risk-sensitive inhibitory control for safe reinforcement learning*, 2023 62nd IEEE Conference on Decision and Control (CDC), IEEE, 2023.

Erfaun Noorani and John S Baras, *Risk-sensitive reinforce: A monte carlo policy gradient algorithm for exponential performance criteria*, 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 1522–1527.

_____, *Risk-sensitive reinforce: A monte carlo policy gradient algorithm for exponential performance criteria*, 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 1522–1527.

_____, *Risk-sensitive reinforcement learning and robust learning for control*, 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 2976–2981.

References

- _____, *Risk-sensitive reinforcement learning and robust learning for control*, 2021 60th IEEE Conference on Decision and Control (CDC), IEEE, 2021, pp. 2976–2981.
- _____, *Embracing risk in reinforcement learning: The connection between risk-sensitive exponential and distributionally robust criteria*, 2022 American Control Conference (ACC), IEEE, 2022, pp. 2703–2708.
- _____, *From regularization to risk-sensitivity-and back again*, IFAC-PapersOnLine **55** (2022), no. 15, 33–38.
- _____, *A probabilistic perspective on risk-sensitive reinforcement learning*, 2022 American Control Conference (ACC), IEEE, 2022, pp. 2697–2702.
- _____, *Risk-attitudes, trust, and emergence of coordination in multi-agent reinforcement learning systems: A study of independent risk-sensitive reinforce*, 2022 European Control Conference (ECC), IEEE, 2022, pp. 2266–2271.

References

Erfaun Noorani, Karl Johansson, and John S Baras, *Risk-sensitive reinforcement learning and robust learning for control*, 2023 62nd IEEE Conference on Decision and Control (CDC), IEEE, 2023, pp. ?-?

Erfaun Noorani, Christos Mavridis, and John Baras, *Risk-sensitive reinforcement learning with exponential criteria*, arXiv preprint arXiv:2212.09010 (2022).

Erfaun Noorani, Christos N Mavridis, and John S Baras, *Exponential td learning: A risk-sensitive actor-critic reinforcement learning algorithm*, 2023 American Control Conference (ACC), IEEE, 2023, pp. 2697–2702.

Alexander Peysakhovich and Adam Lerer, *Prosocial learning agents solve generalized stag hunts better than selfish ones*, Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (Richland, SC), AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 2043–2044.

Zhenggang Tang, Chao Yu, Boyuan Chen, Huazhe Xu, Xiaolong Wang, Fei Fang, Simon Du, Yu Wang, and Yi Wu, *Discovering Diverse Multi-Agent Strategic Behavior via Reward Randomization*, arXiv preprint arXiv:2103.04564 (2021).

Peter Whittle, *Risk-sensitive optimal control*, vol. 2, Wiley, 1990.

Exponential Loss for Deep Neural Networks (T. Poggio et al, PNAS 2020)

Theoretical issues in deep networks

Tomaso Poggio^{a,1}, Andrzej Banburski^a, and Qianli Liao^a

Exponential Loss Function

The standard approach to training deep networks is to use stochastic gradient descent to find the weights W_k that minimize the empirical exponential loss $L = \frac{1}{N} \sum_n e^{-y_n f(x_n)}$ by computing

$$\dot{W}_k = -\frac{\partial L}{\partial W_k} = \frac{1}{N} \sum_{n=1}^N y_n \frac{\partial f(W; x_n)}{\partial W_k} e^{-y_n f(W; x_n)} \quad [1]$$

on a given dataset $\{x_n, y_n\} \quad \forall n = 1, \dots, N$ with y binary. Since

Instead

B.1. Constrained minimization. Constrained optimization of the exponential loss, using Lagrange multipliers, minimizes $L = \frac{1}{N} \sum_n e^{-\rho y_n f(V; x_n)}$ under the constraint $\|V_k\| = 1$ which leads us to minimize

$$\mathcal{L} = \frac{1}{N} \sum_n e^{-\rho y_n f(V; x_n)} + \sum_k \lambda_k \|V_k\|^2 \quad [2]$$

with λ_k such that the constraint $\|V_k\| = 1$ is satisfied. We

Exponential Loss Function

B.2. Fixed ρ : minima. Gradient descent on \mathcal{L} for fixed ρ wrt V_k yields then the dynamical system

$$\dot{V}_k = \rho \frac{1}{N} \sum_n e^{-\rho y_n f(V; x_n)} y_n \left(\frac{\partial f(V; x_n)}{\partial V_k} - V_k f(V; x_n) \right) \quad [3]$$

because $\lambda_k = \frac{1}{2} \rho \frac{1}{N} \sum_n e^{-\rho y_n f(V; x_n)} y_n f(V; x_n)$, since $V_k^T \dot{V}_k = 0$, which in turn follows from $\|V_k\|^2 = 1$.

THEN

C.2. Summary theorem. The following theorem (informal statement) summarizes the main results on minimization of the exponential loss in deep ReLU networks:

Theorem 4. *Assume that separability is reached at time T_0 during gradient descent on the exponential loss; that is, $y_n f(W; x_n) > 0, \forall n$. Then unconstrained gradient descent converges in terms of the normalized weights to a solution that is under complexity control for any finite time. In addition, the following properties hold:*

Exponential Loss Function

In summary, there is an implicit regularization in deep networks trained on exponential-type loss functions, originating in the gradient descent technique used for optimization. The solutions are in fact the same as those that are obtained by regularized optimization. Convergence to a specific solution instead of another of course depends on the trajectory of gradient flow and corresponds to one of multiple infima of the loss (linear networks will have a unique minimum), each one being a margin maximizer. In general, each solution will show a different test performance. Characterizing the conditions that lead to the best among the margin maximizers is an open problem.

Advancing ML and AI and Applications

Foundations of AI and ML

- Rigorous Mathematics for Deep Networks – Universal Architecture emerging (“One Learning Algorithm Hypothesis”)
- Non von-Neumann computing – do not separate CPU from Memory – Synaptic NN, in-memory processing -- HTM
- Universal ML -- Integrate Deep NN and Synaptic NN
- Knowledge Representation and Reasoning: Integrate Knowledge Graphs and Semantic Vector Spaces
- Progressive Learning, Knowledge Compacting
- Link Machine Learning with Knowledge Representation and Reasoning
- Inspirations from neuroscience: attention, memory, time scales

Advancing AI and ML for Autonomy: our Approach

- Rigorous Mathematics for Deep Networks – Universal Architecture emerging

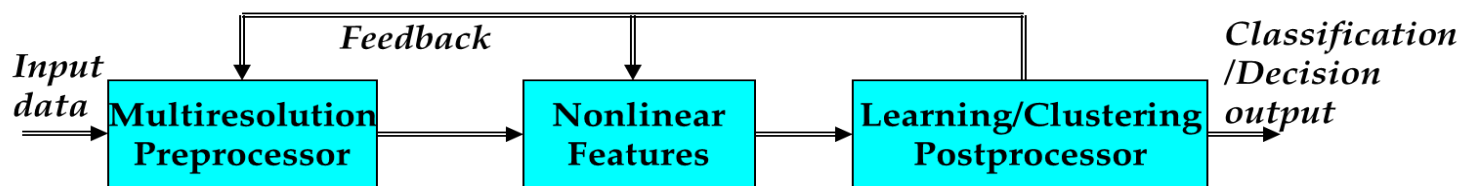


Fig.: Universal model and architecture abstraction of ML and DL algorithms

- **Inspired by the “One Algorithm Hypothesis”** – Andrew Ng *
- Non von-Neumann computing – do not separate CPU from Memory – Synaptic NN, in-memory processing -- HTM
- **Universal ML** -- Integrate Deep NN and Synaptic NN
- **Knowledge Representation and Reasoning**: Integrate Knowledge Graphs and Semantic Vector Spaces
- **Progressive Learning, Knowledge Compacting**
- **Link Machine Learning with Knowledge Representation and Reasoning**
- Inspirations from neuroscience

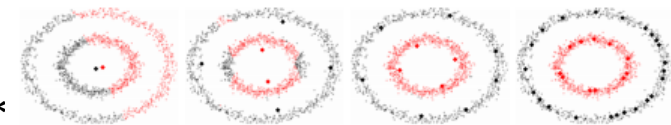
* A. Ng -- https://www.microsoft.com/en-us/research/wp-content/uploads/2013/01/andrew-ng_machinelearning.pdf

Progressive Learning with Deterministic Annealing Optimization

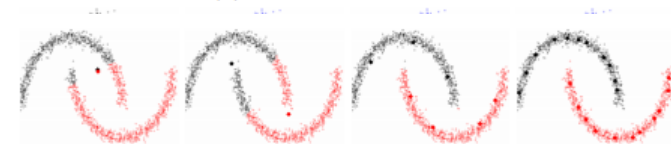
- Iterative machine learning algorithms:
 - What about **model complexity** and **hyper-parameter tuning**?
 - Novel dissimilarity measures – **Bregman divergences**

Progressively growing neural networks

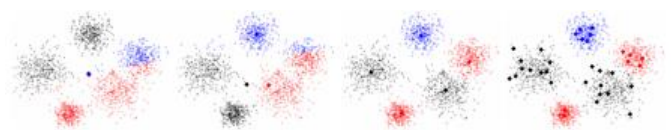
- Online deterministic annealing learning algorithm*
- Online, **gradient-free** training
- Progressively growing number of neurons
- Interpretable**, avoids poor local minima, **robust** wrt the initial conditions
- Memory efficiency, reduced computational complexity
- Control over the complexity-accuracy trade-off**



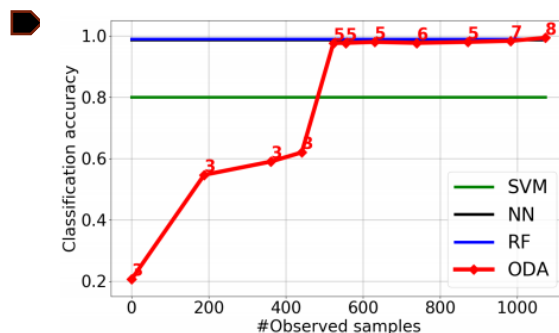
(a) Concentric circles.



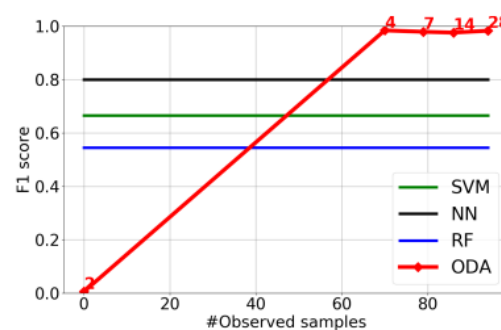
(b) Half moons.



(c) Gaussians.



(a) Gaussians.



(d) Credit Card (F1 score).

Most recent: Self-Organizing Maps w Bregman Divergence
[Mavridis, Raghavan, Baras, 2021]

*C. N. Mavridis and J. S. Baras, "Online Deterministic Annealing for Classification and Clustering," ArXiv (submitted to IEEE TNNLS)
*C. N. Mavridis and J. S. Baras, "Annealing Optimization for Progressive Learning with Stochastic Approximation," submitted to IEEE TACON

ODA – Supervised and Unsupervised Learning

- *Observations:* $X^N := \{x_i\}_{i=1}^N$, $x_i \in S$ realizations of a r.v. $X \in S$
- *Codevectors:* $\mu = \{\mu_i\}_{i=1}^M$, $\mu_i \in S$

Clustering Not Enough: $\min_{\mu} D(X, Q) := \mathbb{E}[d(X, Q)] = \int p(x) \sum_i p(\mu_i|x) d(x, \mu_i) dx$

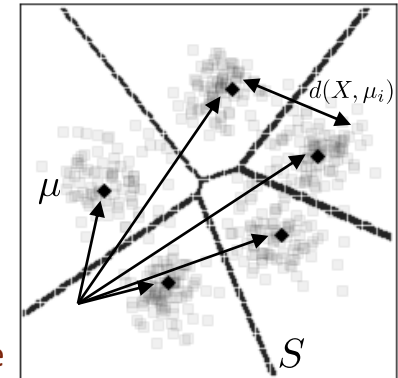
Online Deterministic Annealing

$$\min_{\mu} F_T := D - TH \quad \text{for decreasing values of } T.$$

where $H(X, Q) := \mathbb{E}[-\log P(X, Q)] = H(X) - \int p(x) \sum_i p(\mu_i|x) \log p(\mu_i|x) dx$

▪ Lagrange (Temperature) Coefficient T

- Controls Performance/Complexity Tradeoff
- Simulates Annealing Optimization (Temperature)
- Stochastic Approximation
 - **Simultaneous local system identification/reinforcement learning**
- Triggers Bifurcation
 - **Progressively adjust number of regions/codevectors**

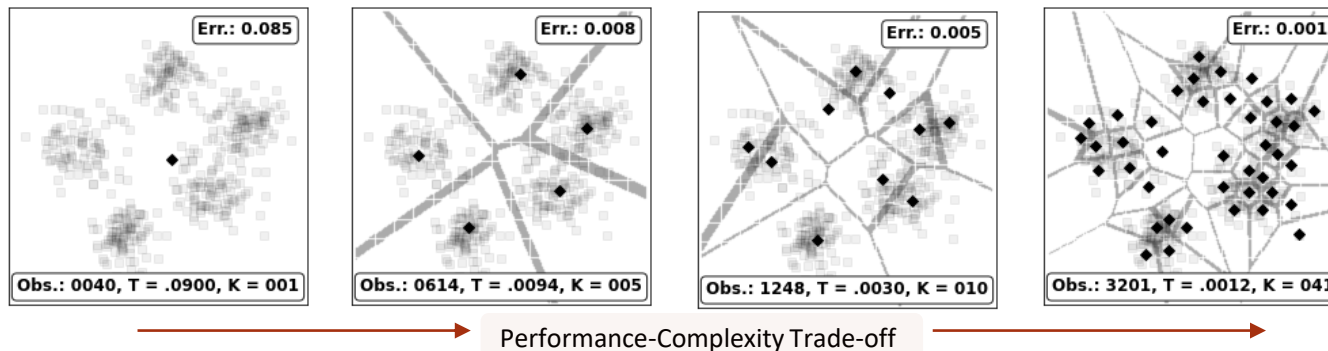


Adaptive
Robust
Progressive

ODA – Supervised and Unsupervised Learning

Bifurcation and the number of codevectors

- ▶ Sequentially solve: $\min F_{T_\infty} := D - T_\infty H$
...
 $\min F_{T_0} := D - T_0 H$, $T_i < T_{i+1}$: Decreasing Temperature
- ▶ **Remark.** As $T \rightarrow \infty$, we get $\mu_i = \mathbb{E}[f(X)]$, $\forall i$, i.e., one unique pseudo-input.
- ▶ **Remark.** As T is lowered below a critical value, a bifurcation phenomenon occurs, and the number of pseudo-inputs increases.

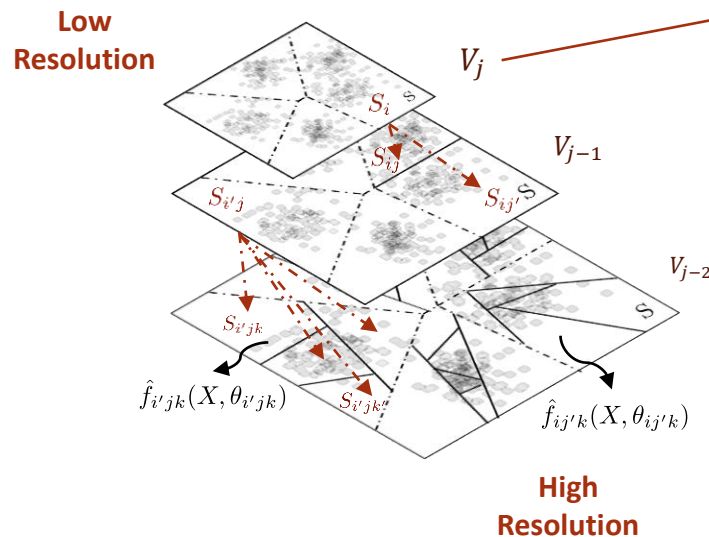


Mavridis, Baras, Online Deterministic Annealing for Classification and Clustering, IEEE TNNLS 2022.

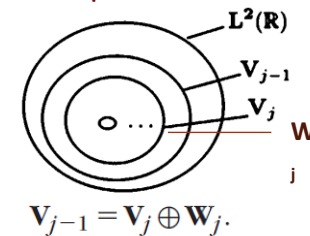
Mavridis, Baras, Annealing Optimization for Progressive Learning with Stochastic Approximation, IEEE TAC 2022.

Online Deterministic Annealing (VII)

Multi-Resolution Hierarchical Learning



Example: Group-convolution Wavelets



- **Constructive (Structured Representation)**
- **Provably Consistent**
- **Localization**
 - Emphasis on regions with high error
- **Asynchronous/Parallel Computation**
- **Reduced Complexity**

Mavridis, Baras, Multi-Resolution Online Deterministic Annealing: A Hierarchical and Progressive Learning Architecture [under review].


Mavridis, Baras, Towards the One Learning Algorithm Hypothesis: A System-theoretic Approach [under review].

ODA – Connection to Risk-Sensitive Optimization


➤ Jayne's Maximum Entropy Principle

- Most “Unbiased” estimator: each sub-problem induces “good” initial conditions for the next
- Duality (Legendre-type) and Regularization:

$$\frac{1}{\beta} \log \mathbb{E}_{P_\mu} [e^{\beta Z}] = \inf_{P_\nu \in \mathcal{P}(\Omega)} \left\{ \mathbb{E}_{P_\nu} [Z] - \frac{1}{\beta} D_{KL}(P_\nu, P_\mu) \right\}, \beta < 0$$


$$\min F_T \simeq \min \frac{1}{\beta} \log \mathbb{E} [e^{\beta D}], \beta = -\frac{1}{T}$$

Risk-Sensitivity


$$\frac{1}{\beta} \log \mathbb{E} [e^{\beta J}] = \mathbb{E} [J] + \frac{\beta}{2} \text{Var} [J] + O(\beta^2)$$

- Robustness w.r.t. initial conditions, input perturbations.

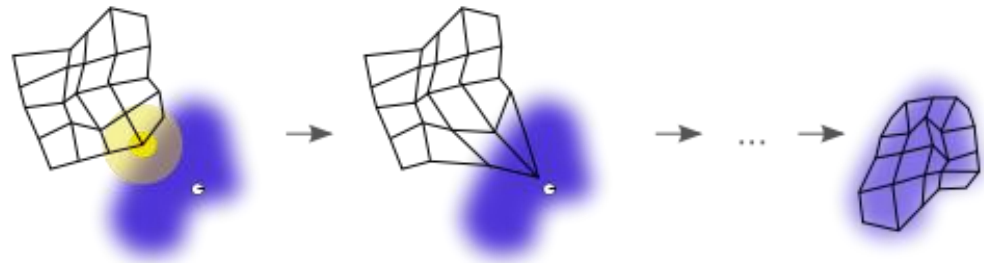
Mavridis et al., Risk Sensitivity and Entropy Regularization in Prototype-based Learning, IEEE MED 2022.

Reinforcement Learning Robot Control with Progressive State-Action Aggregation

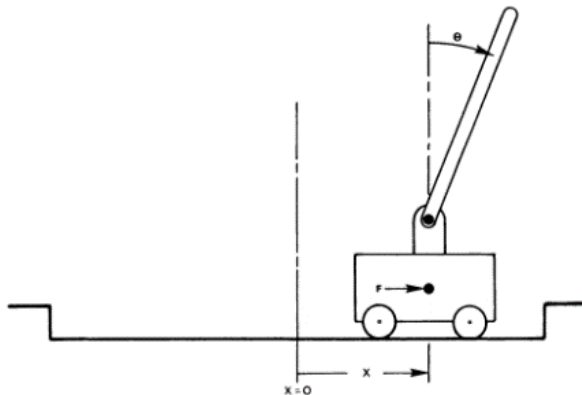
➤ Q-learning in infinite state/action spaces?

➤ **Adaptive State Aggregation***

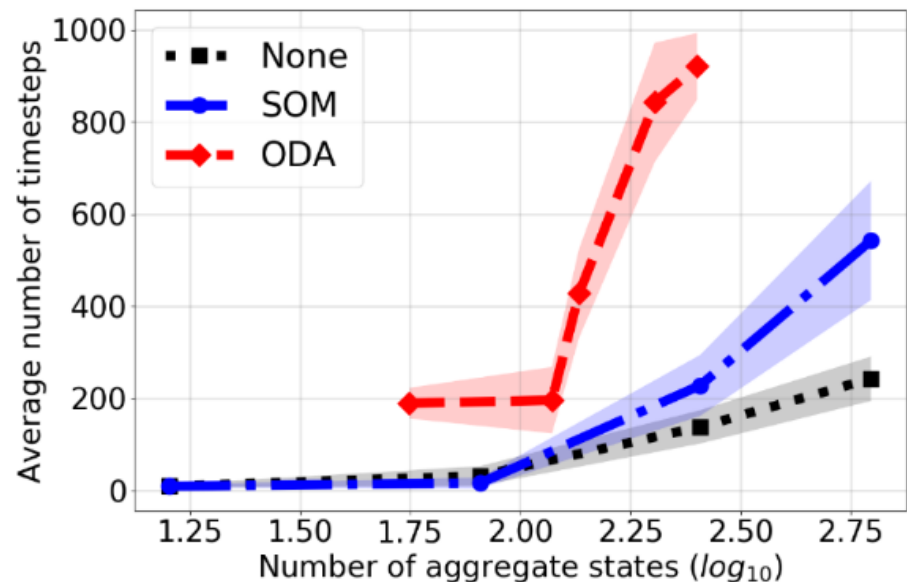
- ~~Ad hoc discretization~~ → Adaptive State Aggregation
- Towards **progressively growing/changing models****
- Memory efficiency, reduced computational complexity



➤ Inverted Pendulum Optimal Control



625 aggregate states → 5 bins/dimension !



* C. N. Mavridis and J. S. Baras, "Vector Quantization for Adaptive State Aggregation in Reinforcement Learning," ACC 2021

* C. N. Mavridis, N. Suriyarachchi and J. S. Baras, "Maximum-Entropy Progressive State Aggregation for Reinforcement Learning," CDC 2021

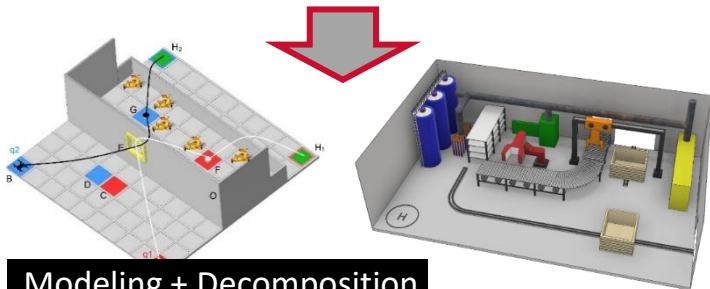
** C. N. Mavridis and J. S. Baras, "Online Deterministic Annealing for Classification and Clustering," ArXiv (submitted to TNNLS)

References

- Mavridis, Christos N., and John S. Baras. "Annealing optimization for progressive learning with stochastic approximation." IEEE Transactions on Automatic Control (2022).
- Mavridis, Christos N., and John S. Baras. "Online deterministic annealing for classification and clustering." IEEE Transactions on Neural Networks and Learning Systems (2022).
- Mavridis, Christos, and John Baras. "Multi-Resolution Online Deterministic Annealing: A Hierarchical and Progressive Learning Architecture." arXiv preprint arXiv:2212.08189 (2022) [submitted to IEEE Transaction on Signal Processing].
- Mavridis, Christos, and John Baras. "Towards the one learning algorithm hypothesis: A system-theoretic approach." arXiv preprint arXiv:2112.02256 (2021).
- Mavridis, Christos N., and John S. Baras. "Convergence of stochastic vector quantization and learning vector quantization with bregman divergences." IFAC-PapersOnLine 53.2 (2020): 2214-2219.
- Mavridis, Christos, Erfan Noorani, and John S. Baras. "Risk sensitivity and entropy regularization in prototype-based learning." 2022 30th Mediterranean Conference on Control and Automation (MED). IEEE, 2022.

Trustworthy Autonomy in Multi-agent Systems with Safe Learning: Approach/Results

Safety and Time-Critical Multi-robot missions



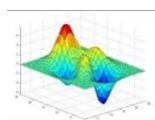
Modeling + Decomposition

Formal Methods

Temporal Logic
and Hybrid
A

Control Theory

Optimal Control

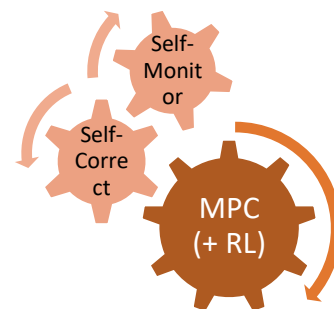
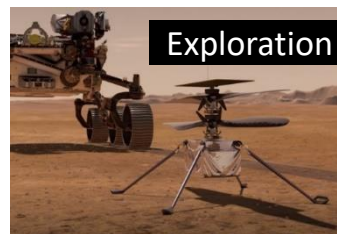


Optimization

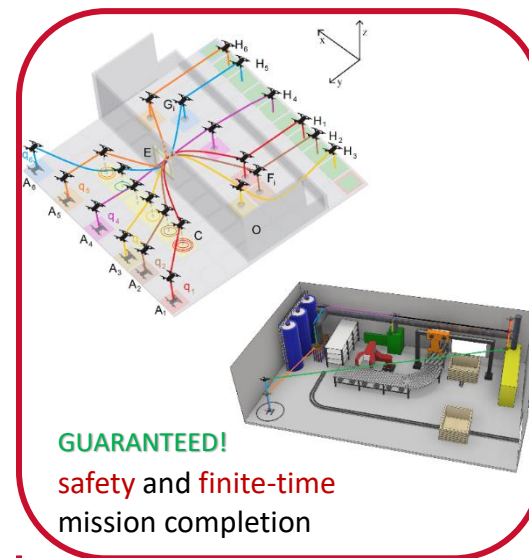
Mixed-Integer
Linear

$$\begin{aligned} & \text{maximize } c^T x \\ & \text{subject to } Ax + s = b, \\ & \quad s \geq 0, \\ & \quad x \geq 0, \\ & \quad \text{and } x \in \mathbb{Z}^n, \end{aligned}$$

Composable, real-time mission planning



Safe learning

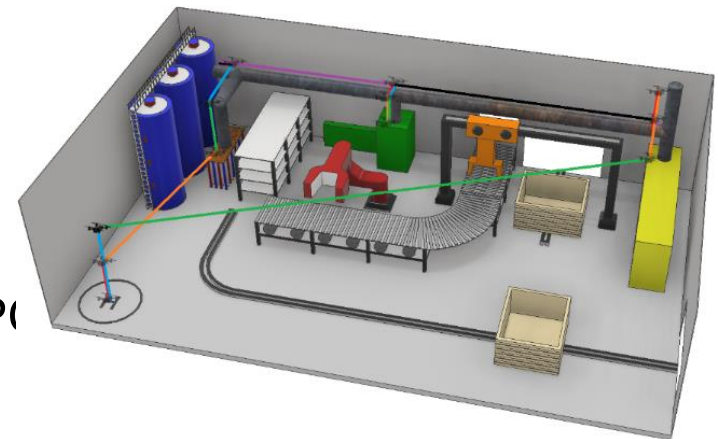
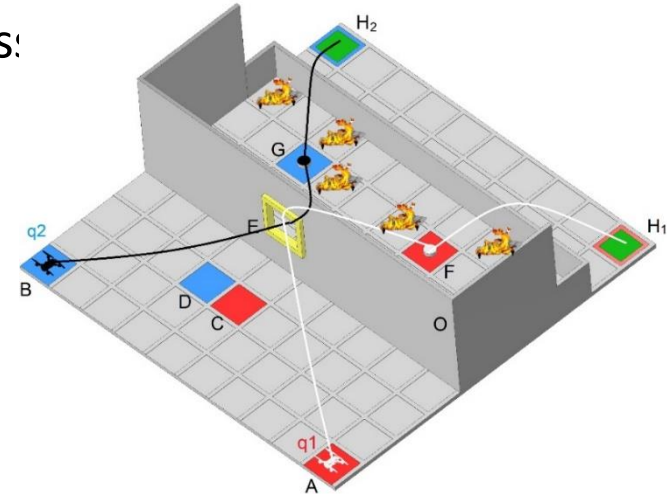


Real-time, fast algorithms

- [1] Fiaz and Baras, 2020. Fast, Composable Rescue Mission Planning for UAVs using Metric Temporal Logic, IFAC World Congress, 2020
- [2] Fiaz et al., 2021. Composable, Safe Mission Planning for UAV-Based Inspection tasks, for IEEE CS-L
- [3] Fiaz et al., 2021. Safe, Hybrid, Real-time Trajectory Planning for Quadrotors with Finite-Time Guarantees, for IEEE RA-L

Assured Autonomy in Multi-agent Systems with Safe Learning: Our Approach

- **Composable, hybrid mission planning** for multiagent systems
 - MTL specifications to represent complex mis:
 - Systematic decomposition into sub-tasks
 - Fast, optimization-based planning method
- **Assurances or guarantees**
 - Safety of all agents
 - Finite-time mission completion
 - Real-time performance (almost)
- **Dealing with uncertainty**
 - Self-monitoring for MTL sub-tasks
 - Self-correction using Event-triggered MP
 - Safe learning

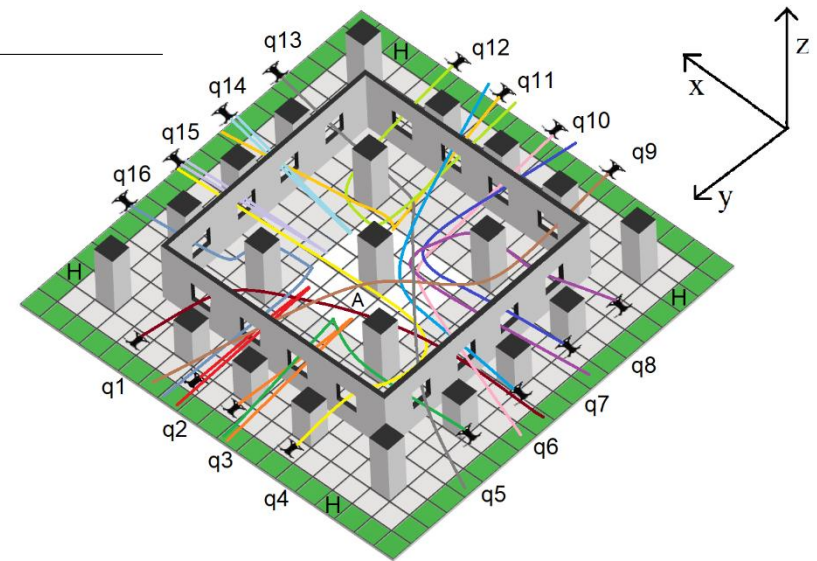
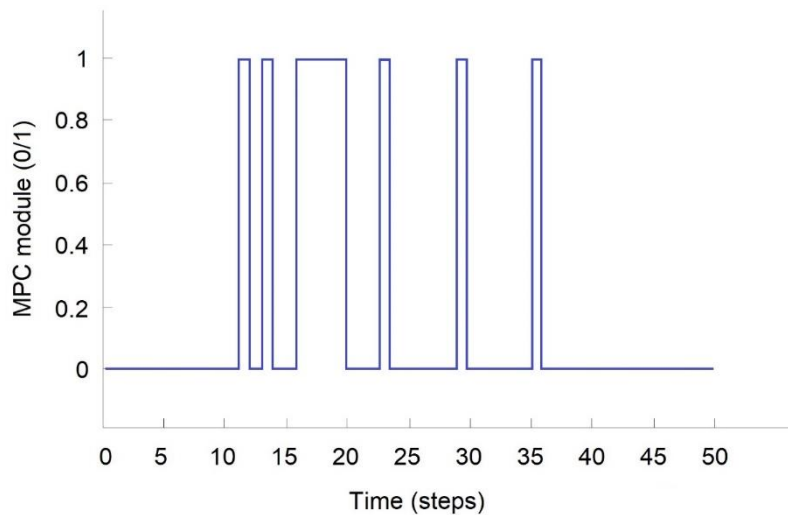


Simulation and results: Performance

- Example: Execution times for sub-task ϕ_{11}^k ():

Table 5.3: Timing analysis for the sub-tasks ϕ_i^k for the whole mission

Sub - task ϕ_i^k	Execution time without safe learning (steps)	Execution time with safe learning (steps)
$\phi_{11}^1 (H - A)$	$15 \leq 20$	$16 \leq 20$
$\phi_{11}^2 (A - H)$	$21 \leq 30$	$24 \leq 30$

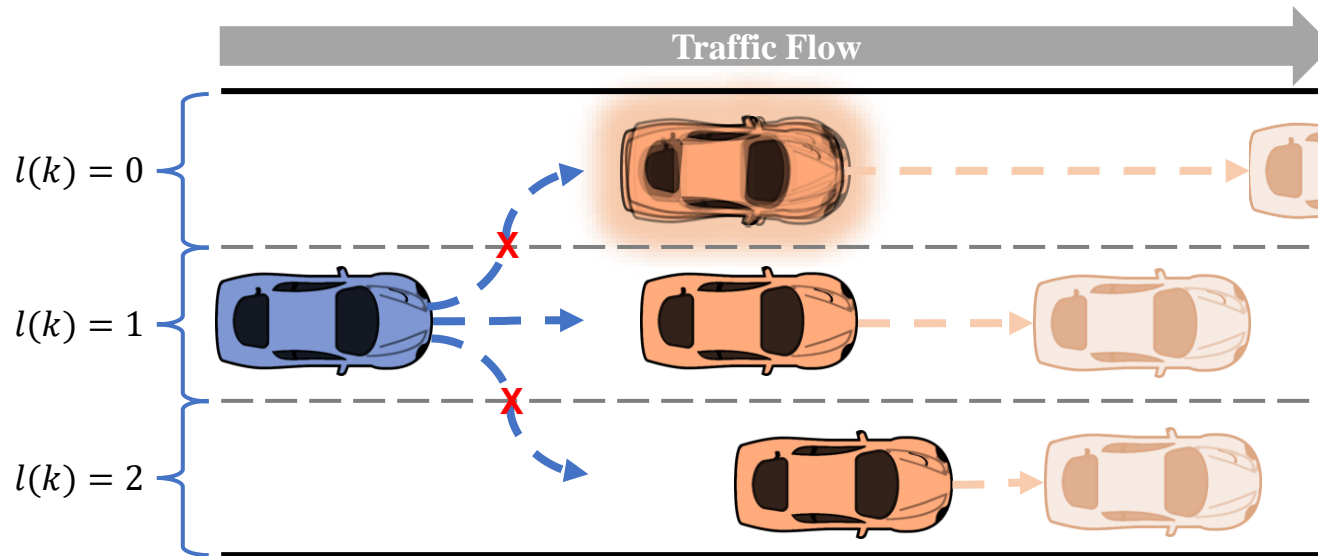


Several Contributions

- Composable, realtime, hybrid mission planning with safety and finite-time guarantees
 - UAV-based search and rescue scenario with evacuation
 - Search and rescue scenario with a team of ground robots in a leader-follower setting
 - UAV-based inspection tasks in a smart factory
- Safe learning mechanism for multiagent systems with MTL specifications
 - Self-monitoring for MTL sub-tasks
 - Self-correction with event-triggered MPC
 - UAV-based surveillance missions

Interaction & Risk Aware Approach (RMOP)

Which lane and at what speed to drive?



Major Challenges:

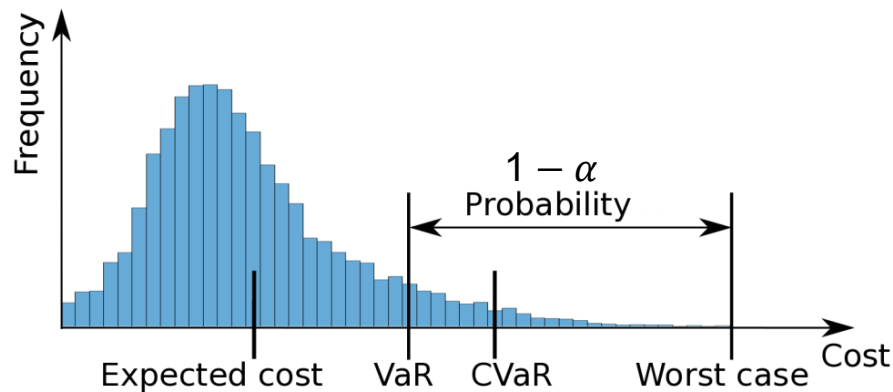
- Quantification of risk associated with each on-road agent
- Interaction and feasibility considerations for high-risk situations

Risk: “Likelihood and severity of the damage that the ego vehicle may suffer in the future”

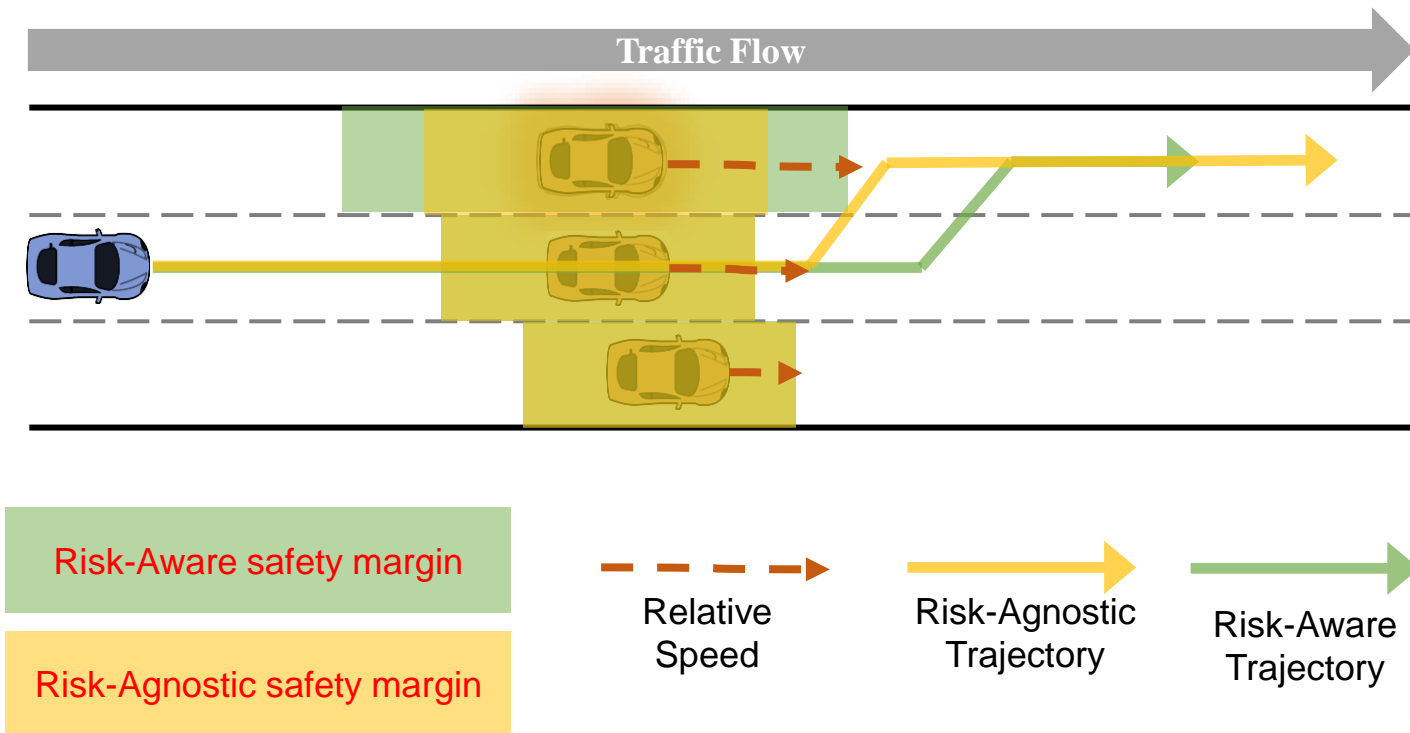
$$\rho: X \rightarrow \mathbb{R}$$

Vehicle-specific Risk: $\rho_i^k = \beta \cdot \text{CVaR}_{\alpha_s}(\mathcal{A}_i^k) + (1 - \beta) \cdot \text{CVaR}_{\alpha_d}(\mathcal{W}_i^k)$
Acceleration Ang. Velocity

Conditional Value-at-Risk:
$$\text{CVaR}_{\alpha}(\mathcal{X}) = \inf_{w \in \mathbb{R}} \mathbb{E} \left\{ w + \frac{[\mathcal{X} - w]^+}{1 - \alpha} \right\}$$

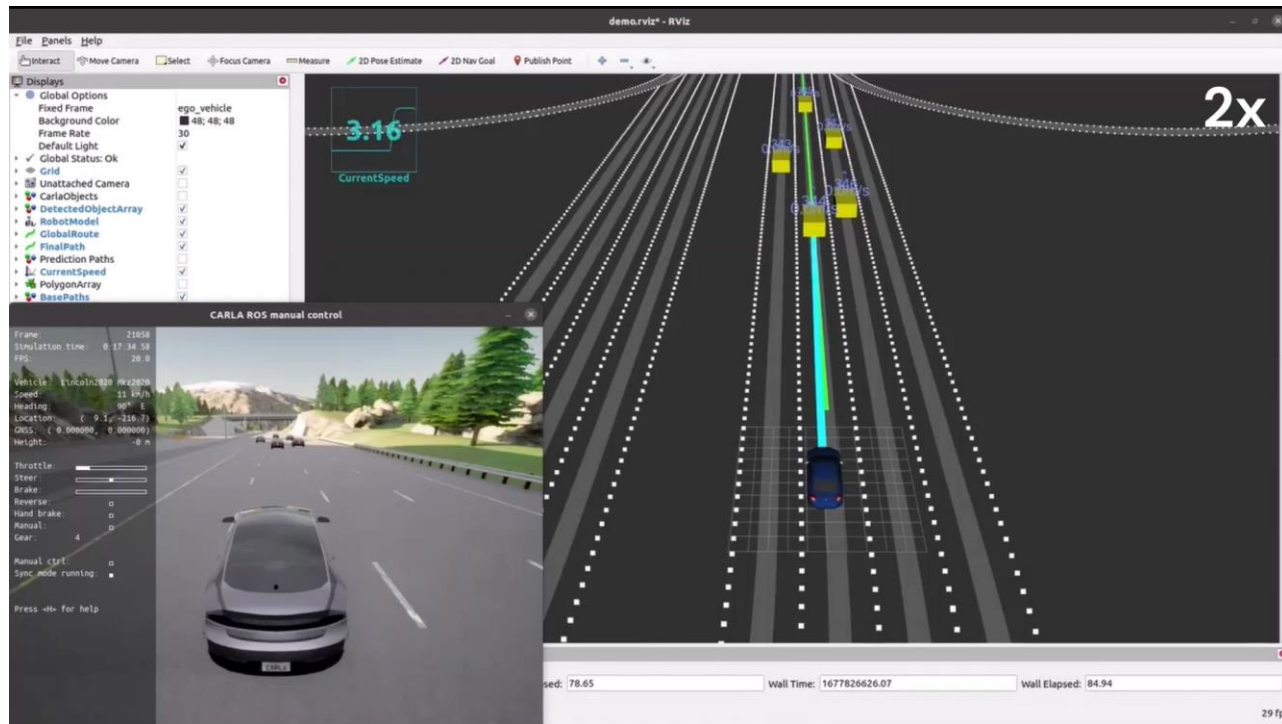


Tariq, Faizan M., et al. "RMOP: Risk-Aware Mixed-Integer Optimization-based Planning for Highway Navigation." IEEE Transactions on Intelligent Transportation Systems (2023). In Review.



Tariq, Faizan M., et al. "RMOP: Risk-Aware Mixed-Integer Optimization-based Planning for Highway Navigation." IEEE Transactions on Intelligent Transportation Systems (2023). In Review.

RMOP (Ours – Risk & Interaction Aware)



Tariq, Faizan M., et al. "RMOP: Risk-Aware Mixed-Integer Optimization-based Planning for Highway Navigation." IEEE Transactions on Intelligent Transportation Systems (2023). In Review.

Future Directions

- Tighten this theory
- Develop further general duality between performance measures and coherent risk measures
- Extend to multi-agent systems
- “Mathematize” Prospect Theory
- Design specialized hardware

Thank you!

baras@umd.edu

301-405-6606

<https://johnbaras.com/>

Questions?