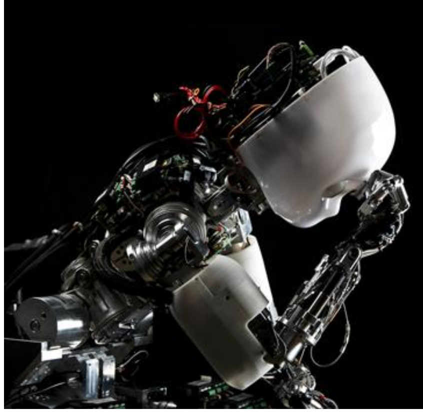
LCCC Symposium on Robotic Skill Learning and Cognition
Lund, 17-19 April 2012

Challenges in adapting imitation and reinforcement learning to compliant robots

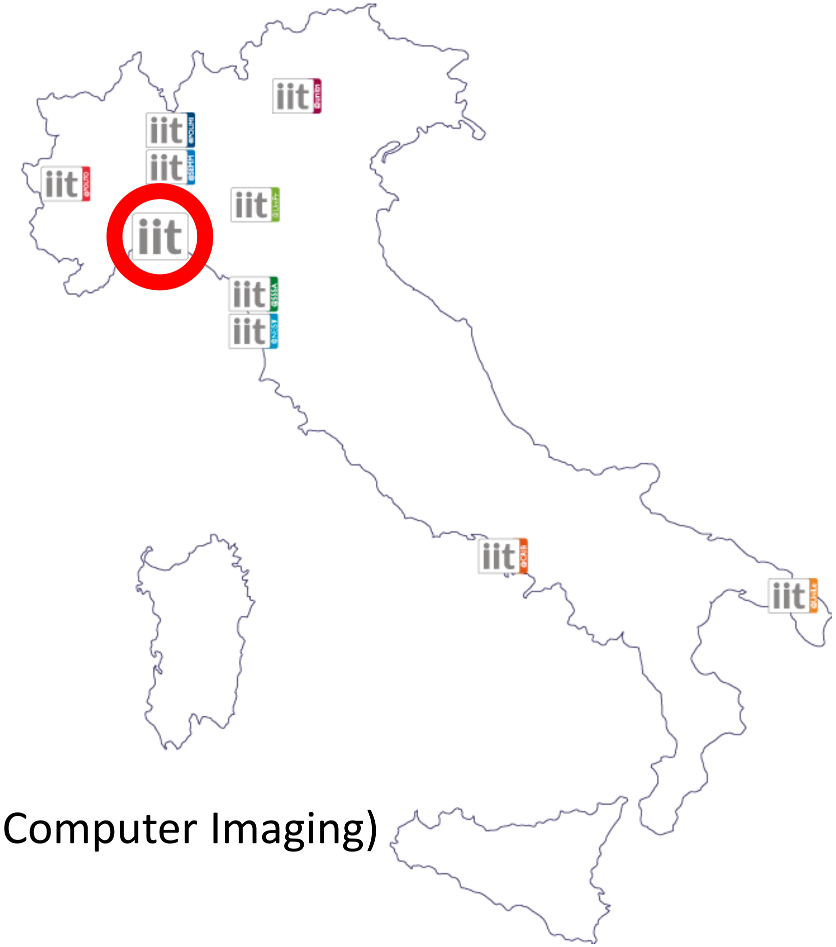
Sylvain Calinon
Learning & Interaction Group
Department of Advanced Robotics
Italian Institute of Technology (IIT)
www.programming-by-demonstration.org/learning-and-interaction/

Italian Institute of Technology (IIT)



iCub built at IIT

- Created in 2003, headquarters in Genova
- Group of nine universities as satellite research units
- Over 400 researchers (37 different countries, 200 working in the area of robotics)



Advanced Robotics
(Director: Darwin Caldwell)

Robotics, Brain and Cognitive Sciences
(Director: Giulio Sandini)

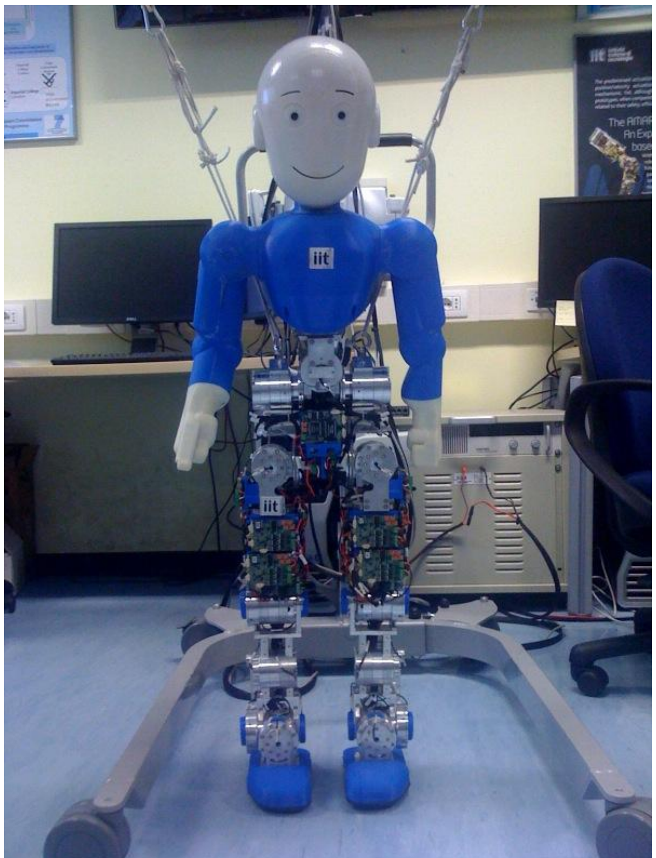
Drug Discovery and Development
(Director: Daniele Piomelli)

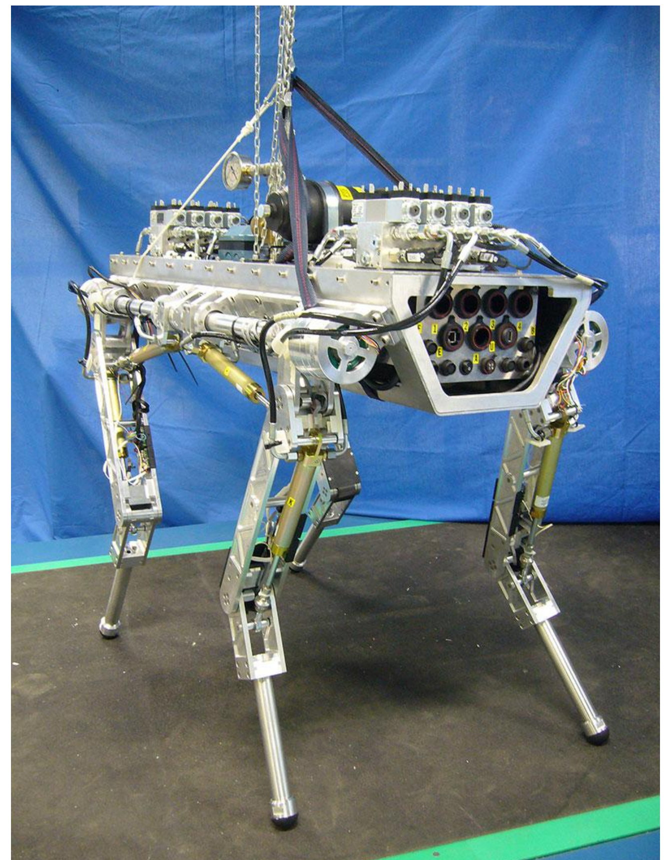
Neuroscience and Brain Technologies
(Director: Fabio Benfenati)

Nanobiotechnology
(Nanochemistry, Nanofabrication, Nanophysics, Computer Imaging)

Advanced Robotics Department (ADVR) @ IIT

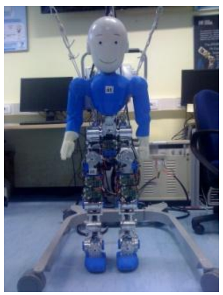
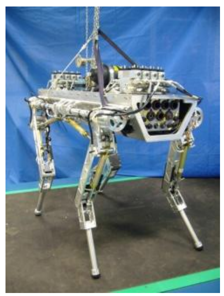
- Over 70 researchers (from 25 PhD students to 5 Full Professors).
- Multidisciplinary approach to design and control, such as the development of SEA-based **CoMan** and hydraulic **HyQ** robots.





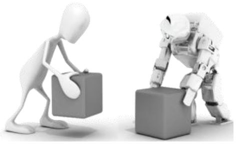
Advanced Robotics Department (ADVR) @ IIT

- Over 70 researchers (from 25 PhD students to 5 Full Professors).
- Multidisciplinary approach to design and control, such as the development of SEA-based **CoMan** and hydraulic **HyQ** robots.

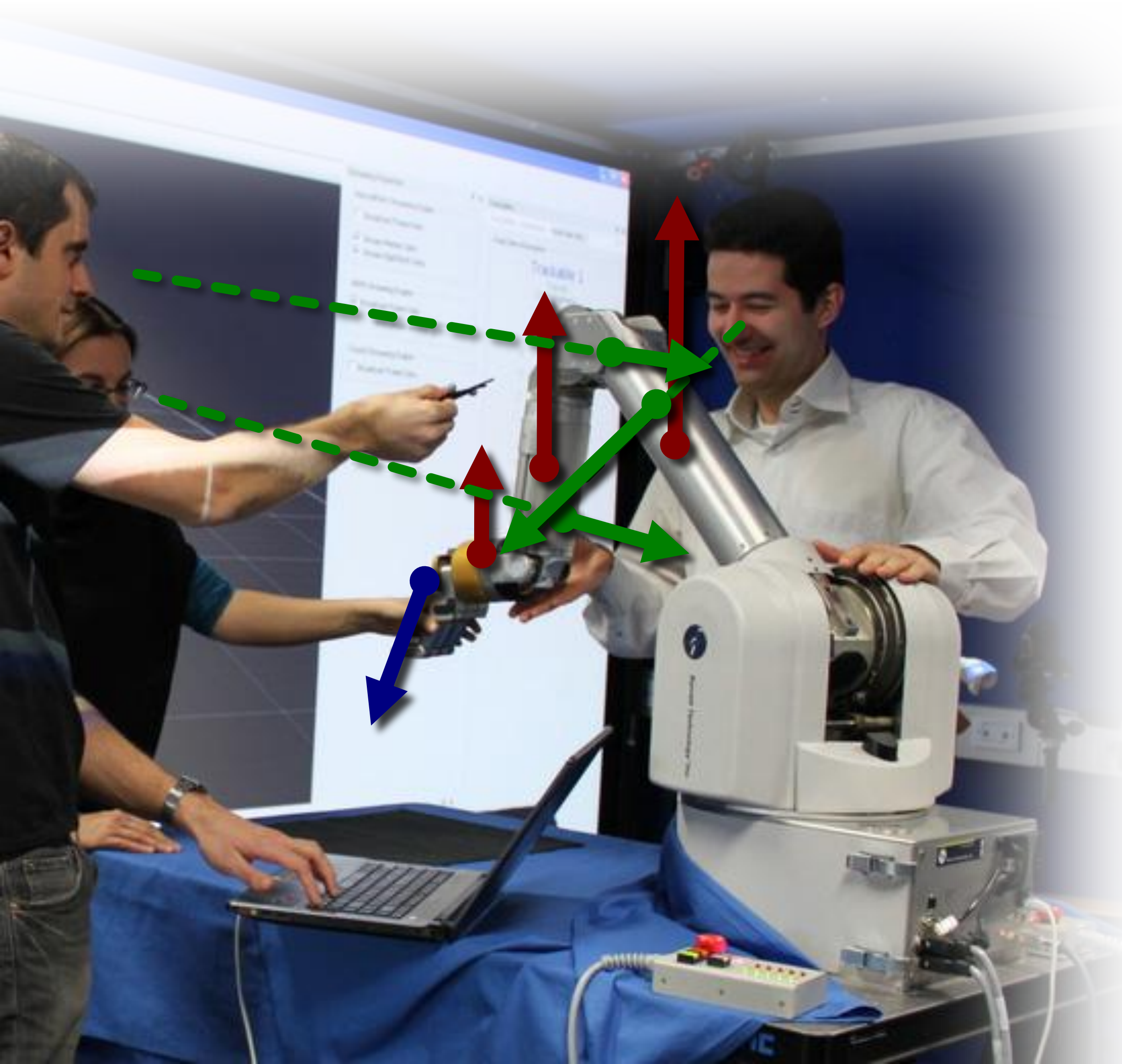

- ADVR resources include a **7-DOFs Barrett WAM** manipulator, a Barrett Hand, a **7-DOFs KUKA Lightweight Arm** and a 6-cameras **VICON** motion tracking system.
- EU research projects: RobotCub, Viactors, Octopus, Hands.DVI, Amarsi, **Saphari** **(2012)**, **Stiff-Flop** **(2012)** and **Pandora** **(2012)**.
- **Learning and Interaction Group** at ADVR created in 2009. (4 postdocs (2012), 5 PhD students)

Learning and Interaction Group @ ADVR-IIT



Compliant control for safe HRI

$$M(q)\ddot{q} + C(\dot{q}, q)\dot{q} + g(q) = \tau_G + \tau_T + \tau_O$$



Gravity compensation

$$\tau_G = \sum_{i=1}^{L} \mathbf{J}_{G,i}^{\top}(q) F_{G,i}$$

Task execution

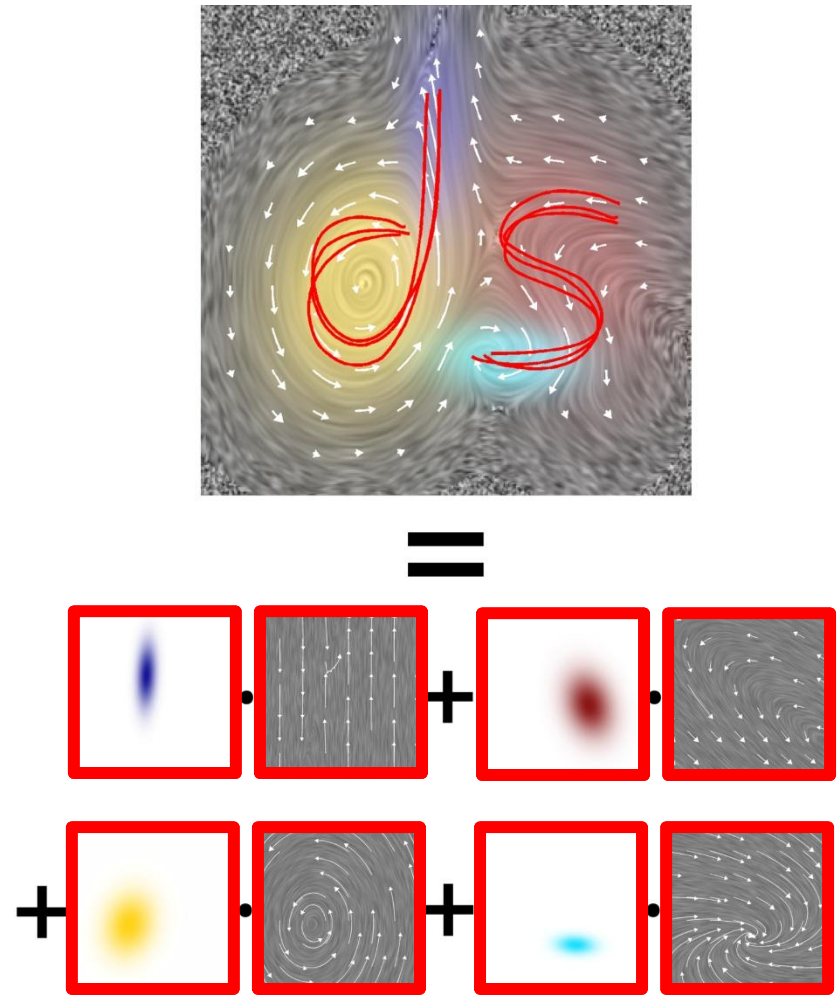
$$\tau_T = \mathbf{J}_T^{\top}(q) F_T$$

User avoidance

$$\tau_O = \mathbf{J}_O^{\top}(q) F_O$$

Flexible representation of skills through a superposition of basis flow fields

$$\dot{\boldsymbol{x}} = \sum_i \overbrace{h_i(\boldsymbol{x},t)}^{\text{scalar weight}} \; \overbrace{(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i)}^{\text{linear subsystem}}$$



Some examples:

- Gaussian Mixture Regression (GMR)
  [Calinon *et al*, IEEE RAM 17(2), 2010]
- Stable Estimator of Dynamical Systems (SEDS)
  [Khansari and Billard, IROS’10]
- Dynamic Movement Primitives (DMP)
  [Ijspeert *et al*, IROS’01][Hoffmann *et al*, ICRA’09]
- Correlated Dynamic Movement Primitives
  [Calinon, Sardellitti and Caldwell, IROS’10]
- Takagi-Sugeno (TS) fuzzy model
  [Takagi and Sugeno, IEEE Trans. SMC 15(1), 1985]

Dynamic Movement Primitives (DMP)

Core idea:

$$\tau \ddot{x} = \kappa^{\mathcal{P}}[x_T - x] - \kappa^{\mathcal{V}} \dot{x} + f(t), \quad f(t) = \sum_{i=1}^{K} h_i(t) f_i$$

Original formulation:

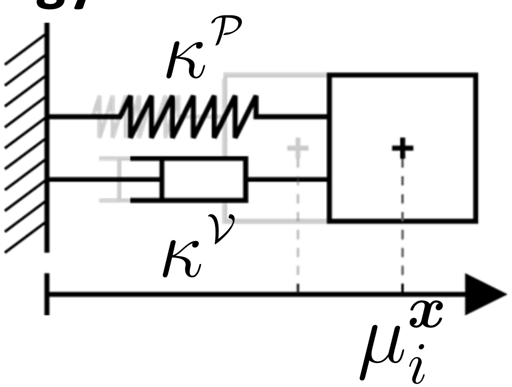
$$\tau \ddot{x} = \kappa^{\mathcal{P}}[x_T - x] - \kappa^{\mathcal{V}} \dot{x} + f(s), \quad f(s) = s\,[x_T - x_0] \sum_{i=1}^{K} h_i(s) f_i$$

$$\tau \dot{s} = -\alpha s$$

[A.J. Ijspeert, J. Nakanishi and S. Schaal, IROS'2001]

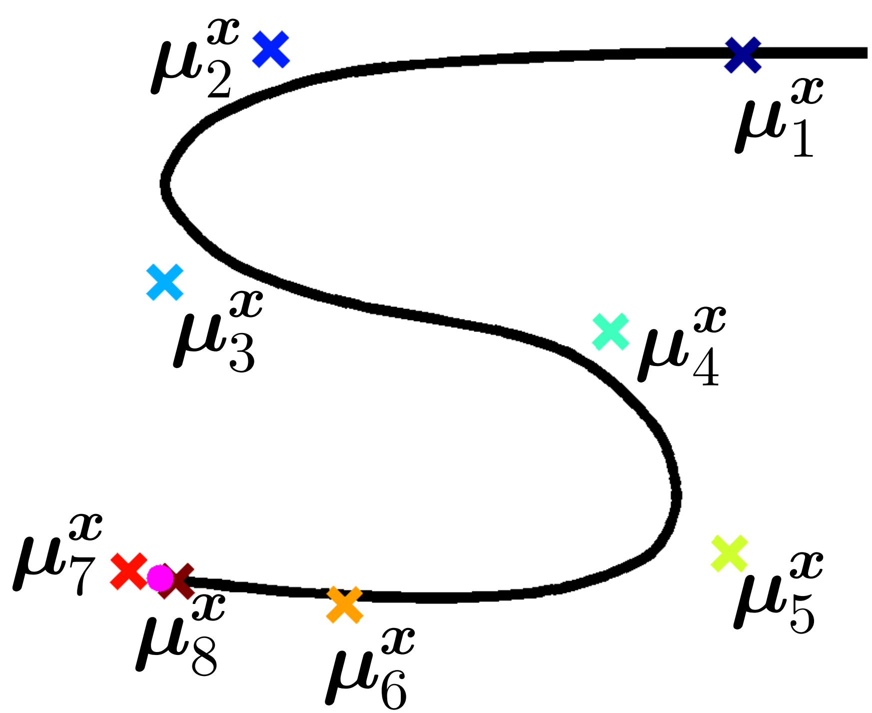
Variant of DMP based on mechanical springs analogy:

$$\ddot{\boldsymbol{x}} = \sum_{i=1}^{K} h_i(t) \Big[\kappa^{\mathcal{P}}(\boldsymbol{\mu}_i^{\boldsymbol{x}} - \boldsymbol{x}) - \kappa^{\mathcal{V}} \dot{\boldsymbol{x}} \Big]$$



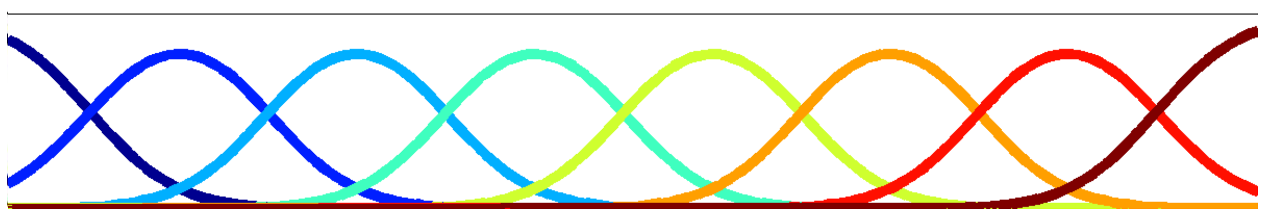
[H. Hoffmann, P. Pastor, D.H. Park and S. Schaal, ICRA'2009]
[S. Calinon, F. D'halluin, D.G. Caldwell and A. Billard, Humanoids'2009]

Extension of Dynamic Movement Primitives



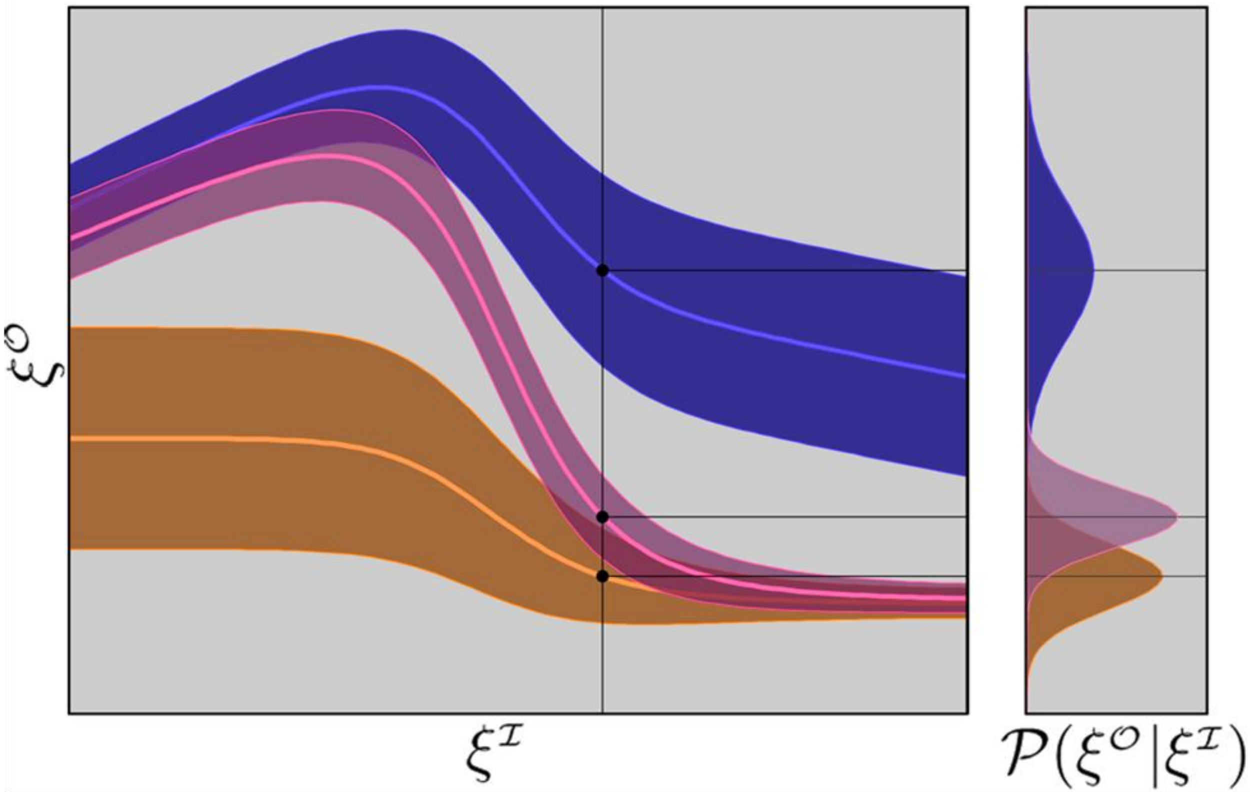
$\dot{s} = -\alpha s$

$h_i(t)$



$$\ddot{\boldsymbol{x}} = \sum_{i=1}^{K} h_i(t) \Big[\kappa^{\mathcal{P}} (\boldsymbol{\mu}_i^{\boldsymbol{x}} - \boldsymbol{x}) - \kappa^{\mathcal{V}} \dot{\boldsymbol{x}} \Big] = \kappa^{\mathcal{P}} (\hat{\boldsymbol{\mu}}^{\boldsymbol{x}} - \boldsymbol{x}) - \kappa^{\mathcal{V}} \dot{\boldsymbol{x}}$$

Gaussian Mixture Regression (GMR)



$$\xi^{\mathcal{I}} = t,\ \xi^{\mathcal{O}} = \frac{1}{\kappa^{\mathcal{P}}}\ddot{\boldsymbol{x}} + \frac{\kappa^{\mathcal{V}}}{\kappa^{\mathcal{P}}}\dot{\boldsymbol{x}} + \boldsymbol{x}$$

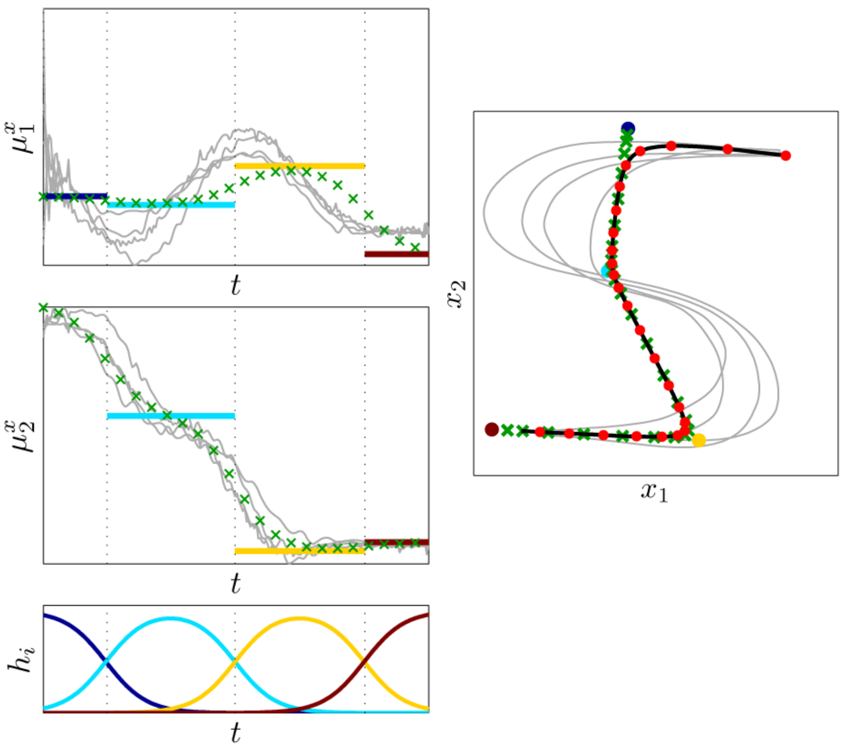
$\mathcal{P}(\xi^{\mathcal{I}}, \xi^{\mathcal{O}})$ encoded in GMM, $\mathcal{P}(\xi^{\mathcal{O}}|\xi^{\mathcal{I}})$ retrieved through GMR

[S. Calinon, F. Guenter and A. Billard, IEEE Trans. on SMC-B 37(2), 2007]

Extension of Dynamic Movement Primitives

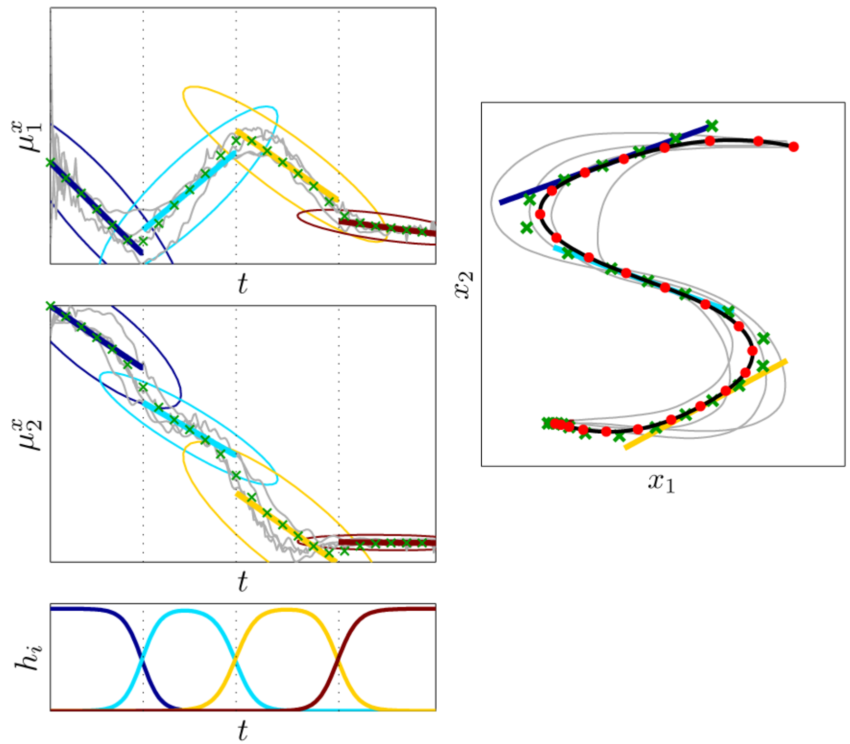
$$\ddot{\boldsymbol{x}} = \kappa^{\mathcal{P}}(\hat{\boldsymbol{\mu}}^{\boldsymbol{x}} - \boldsymbol{x}) - \kappa^{\mathcal{V}}\dot{\boldsymbol{x}}$$

DMP with WLS learning scheme



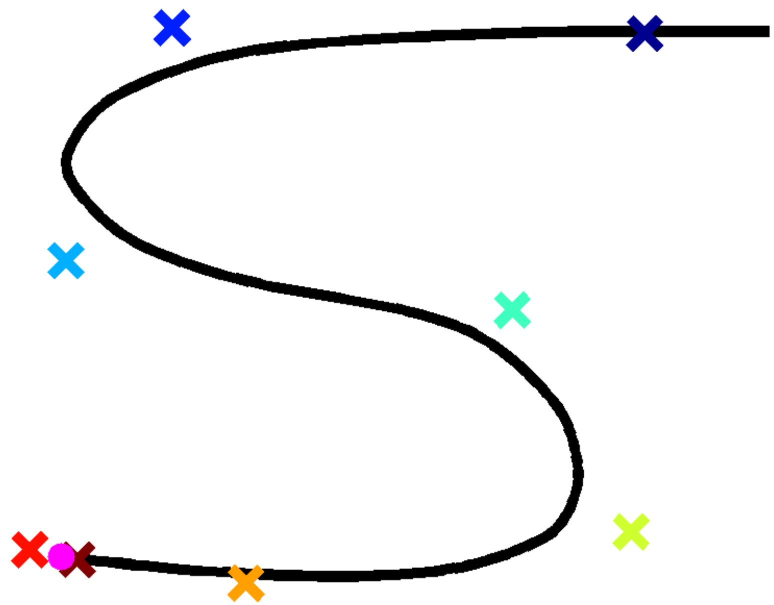
$$\hat{\boldsymbol{\mu}}^{\boldsymbol{x}} = \sum_{i=1}^{K} h_i(t)\ \boldsymbol{\mu}_{0,i}^{\boldsymbol{x}}$$

DMP with GMR learning scheme



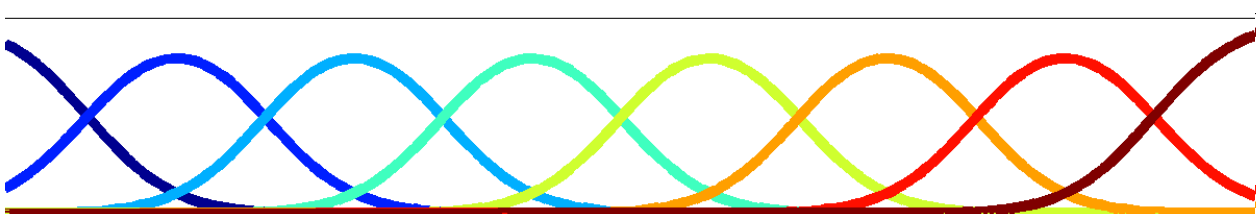
$$\hat{\boldsymbol{\mu}}^{\boldsymbol{x}} = \sum_{i=1}^{K} h_i(t) \left[\boldsymbol{\mu}_{1,i}^{\boldsymbol{x}}\ t + \boldsymbol{\mu}_{0,i}^{\boldsymbol{x}}\right]$$

Extension of Dynamic Movement Primitives



$\dot{s} = -\alpha s$

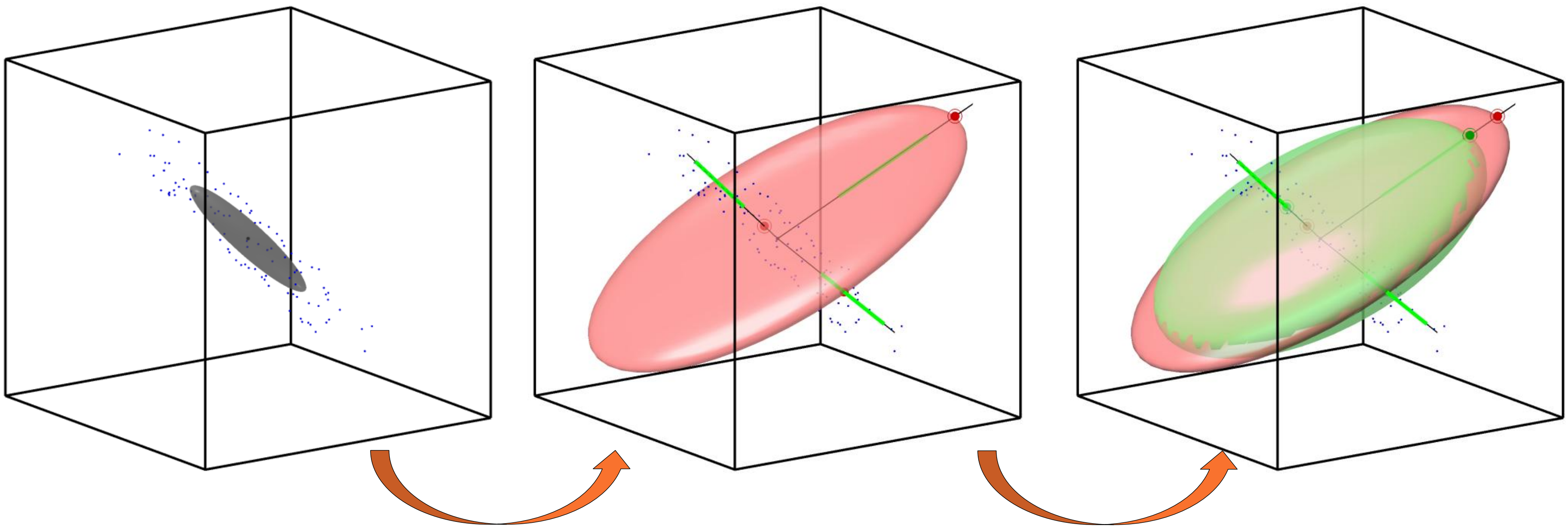
$h_i(t)$



$$\ddot{\boldsymbol{x}} = \sum_{i=1}^{K} h_i(t) \left[\mathbf{K}_i^{\mathcal{P}} (\boldsymbol{\mu}_i^{\boldsymbol{x}} - \boldsymbol{x}) - \kappa^{\mathcal{V}} \dot{\boldsymbol{x}} \right]$$

[Sylvain Calinon, Irene Sardellitti and Darwin Caldwell, IROS'2010]

Extension of Dynamic Movement Primitives



$$\mathbf{K}_i^{\mathcal{P}} = (\boldsymbol{\Sigma}_i^{\boldsymbol{x}})^{-1}$$

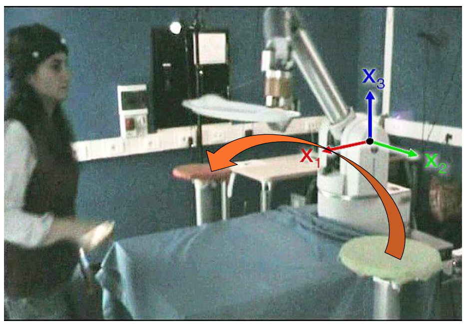
$$\mathbf{K}_i^{\mathcal{P}} = \boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{U}_i^{\top}$$

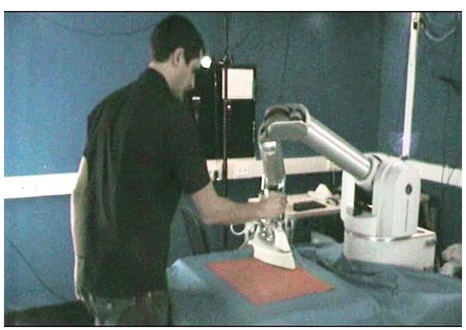
$$\boldsymbol{D}_i \in [\mathbf{K}_{\min}^{\mathcal{P}}, \mathbf{K}_{\max}^{\mathcal{P}}]$$

[Sylvain Calinon, Irene Sardellitti and Darwin Caldwell, IROS'2010]

Learning adaptive stiffness by extracting variability and correlation information

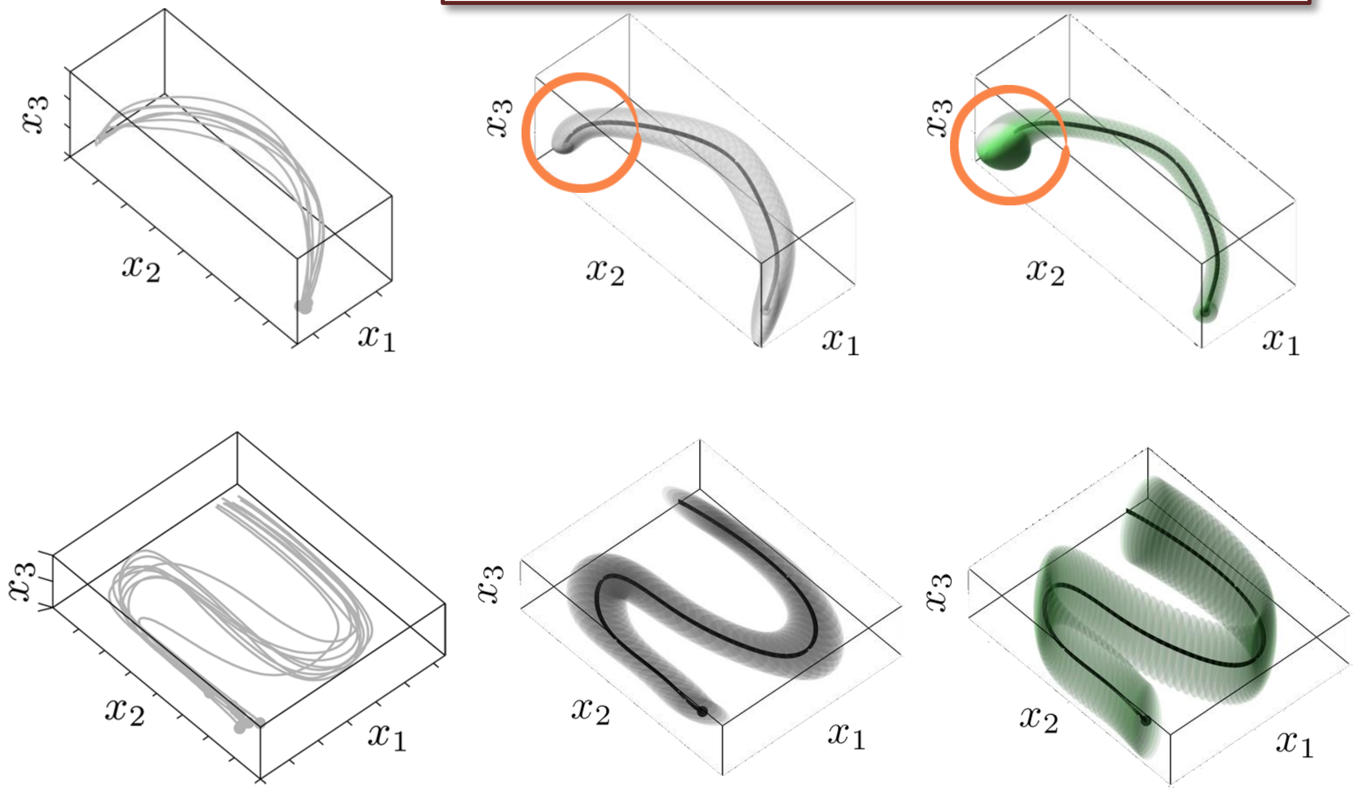
Tasks





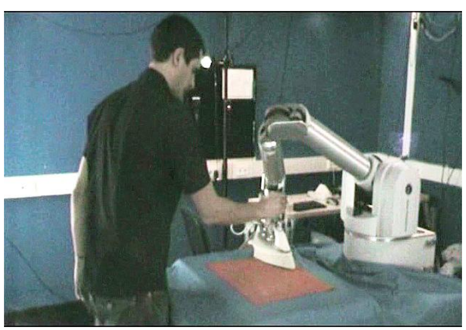
Multiple demonstrations

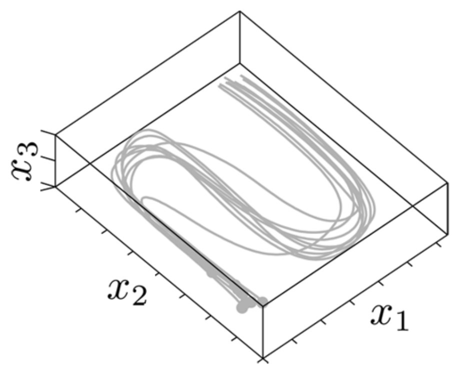
Invariant demonstrations
↓
High gains to track the desired position

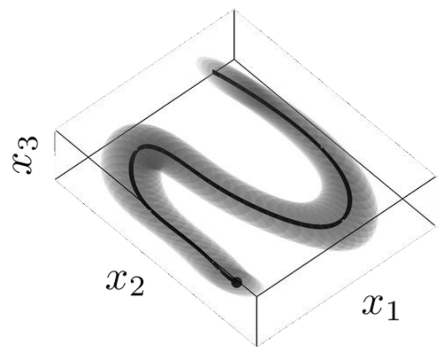


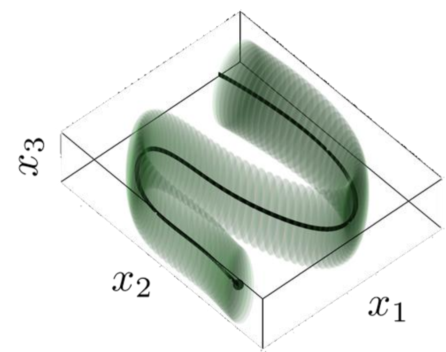
[Sylvain Calinon, Irene Sardellitti and Darwin Caldwell, IROS'2010]

Learning adaptive stiffness by extracting variability and correlation information







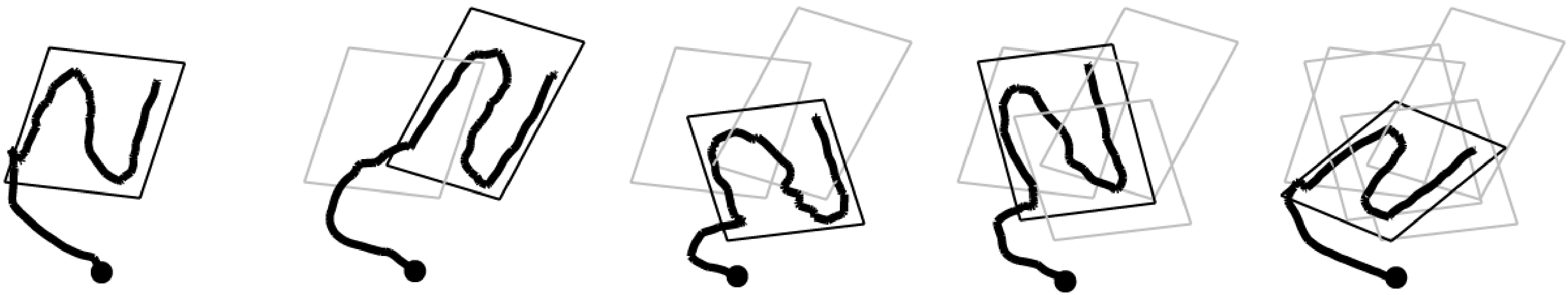


[Sylvain Calinon, Irene Sardellitti and Darwin Caldwell, IROS'2010]

Learning adaptive stiffness by extracting variability and correlation information

[Sylvain Calinon, Irene Sardellitti and Darwin Caldwell, IROS'2010]

Task-parameterized dynamical systems



Some examples:

- Based on **Parametric Hidden Markov Model (PHMM)**:

[Wilson and Bobick, IEEE Trans. on Pattern Analysis and Machine Intelligence 21(9), 1999]
[Krueger, Herzog, Baby, Ude and Kragic, IEEE Robotics & Automation Magazine 17(2), 2010]

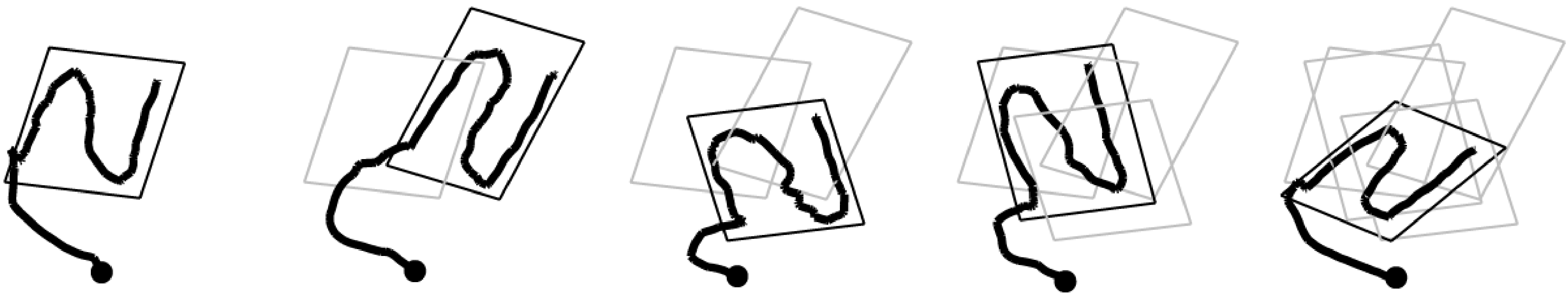
- Based on **Gaussian Mixture Regression (GMR)**:

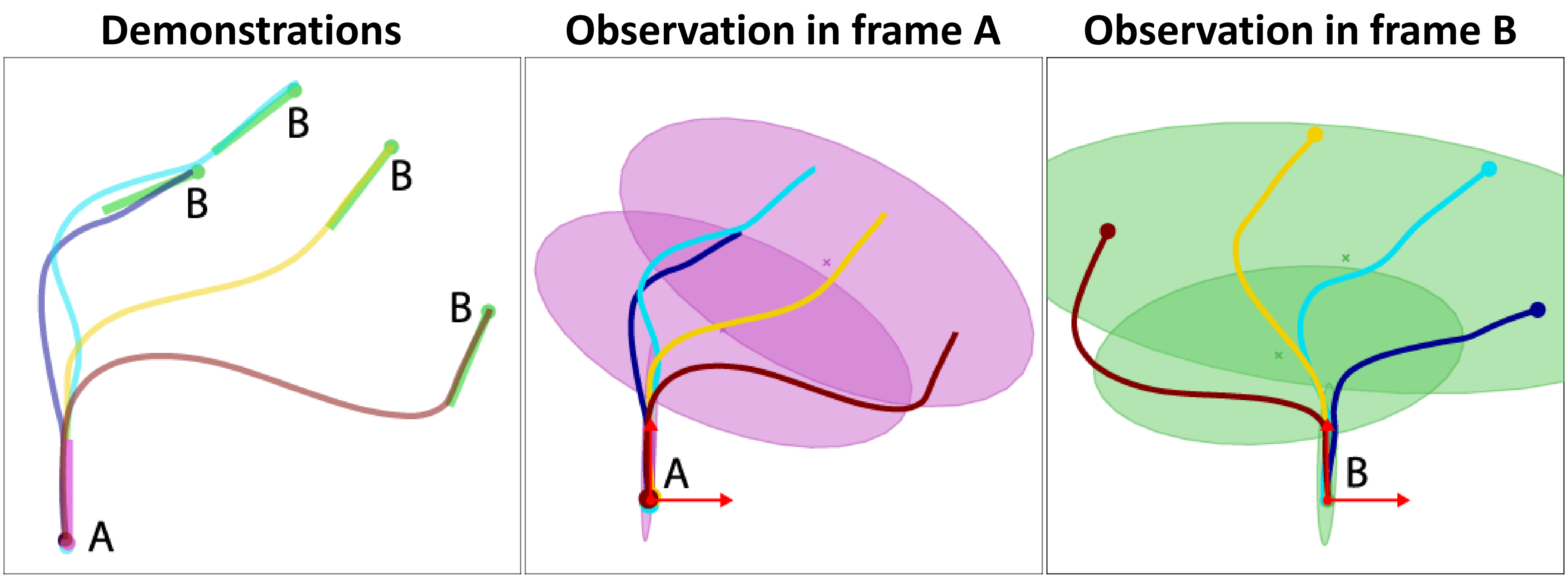
[Muehlig, Gienger, Hellbach, Steil and Goerick, ICRA'2009]
[Cederborg, Ming, Baranes and Oudeyer, IROS'2010]

- Based on **Dynamic Movement Primitives (DMP)**:

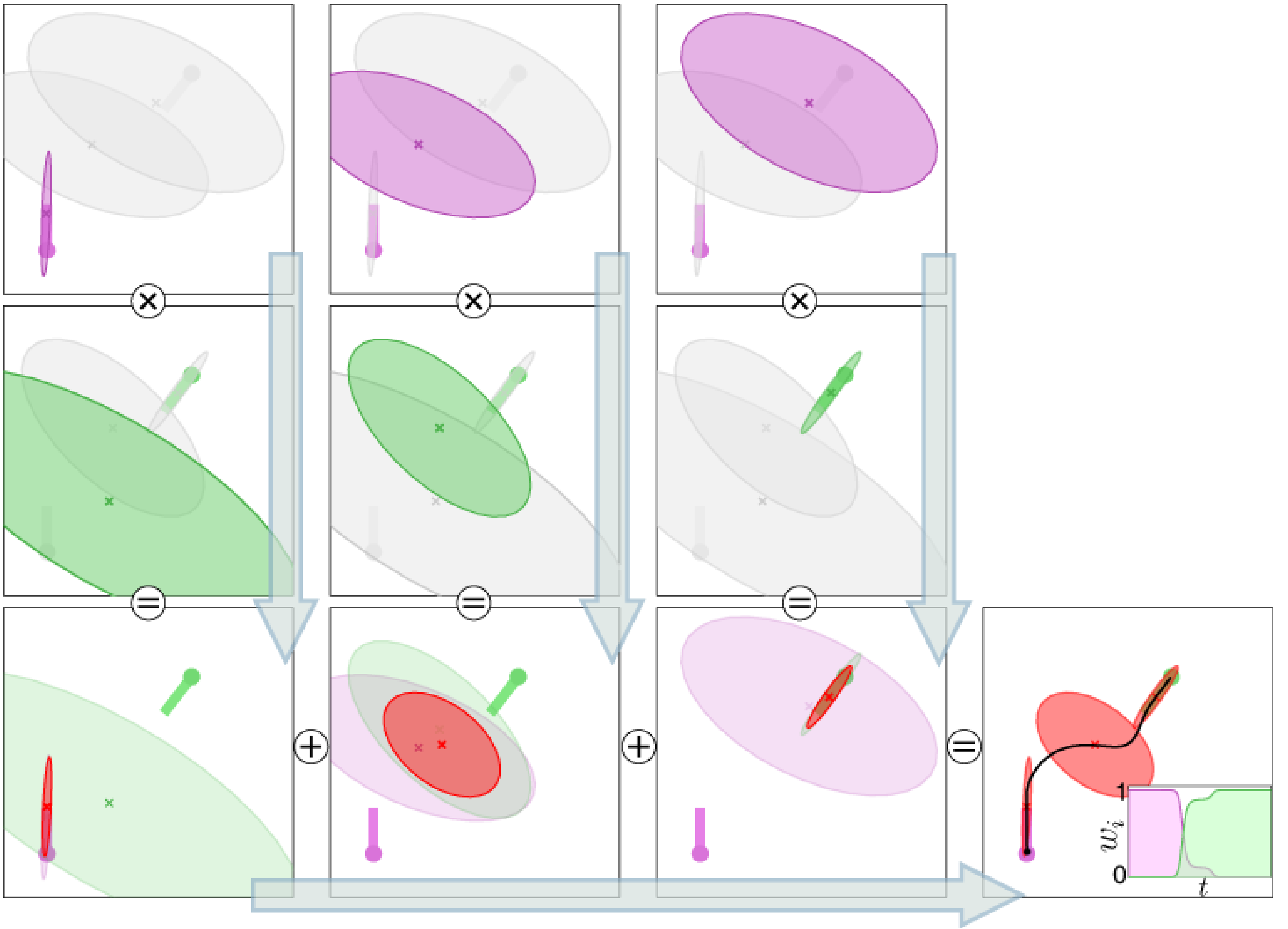
[Kober, Mohler and Peters, IROS'2008]
[Ude, Gams, Asfour and Morimoto, IEEE Trans. on Robotics 26(5), 2010]
[Matsubara, Hyon and Morimoto, Neural Networks 24(5), 2011]

Task-parameterized dynamical systems



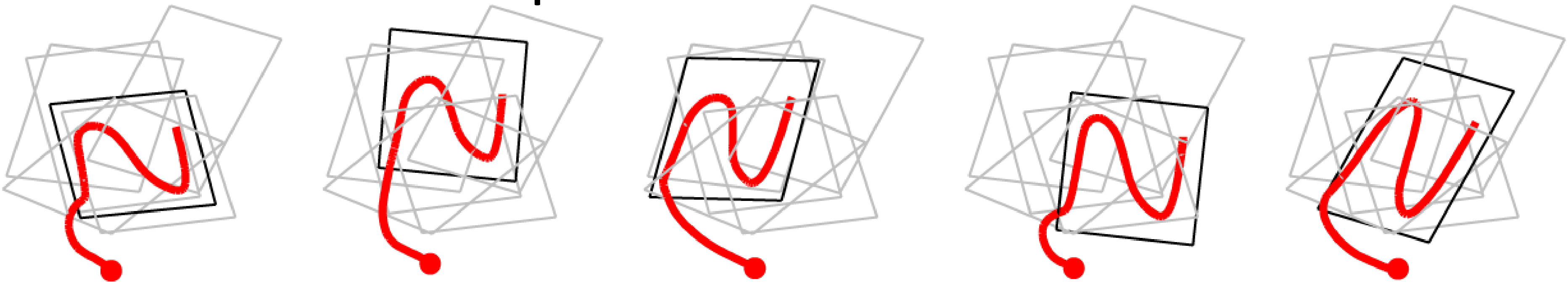


Task-parameterized dynamical systems

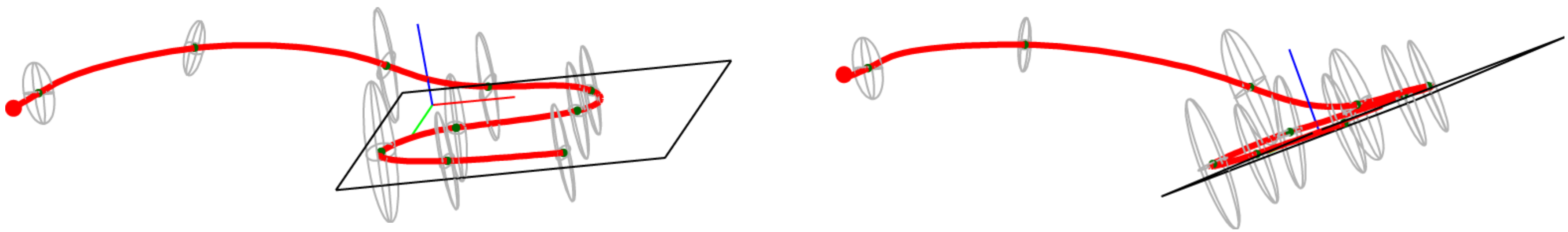


Task-parameterized dynamical systems

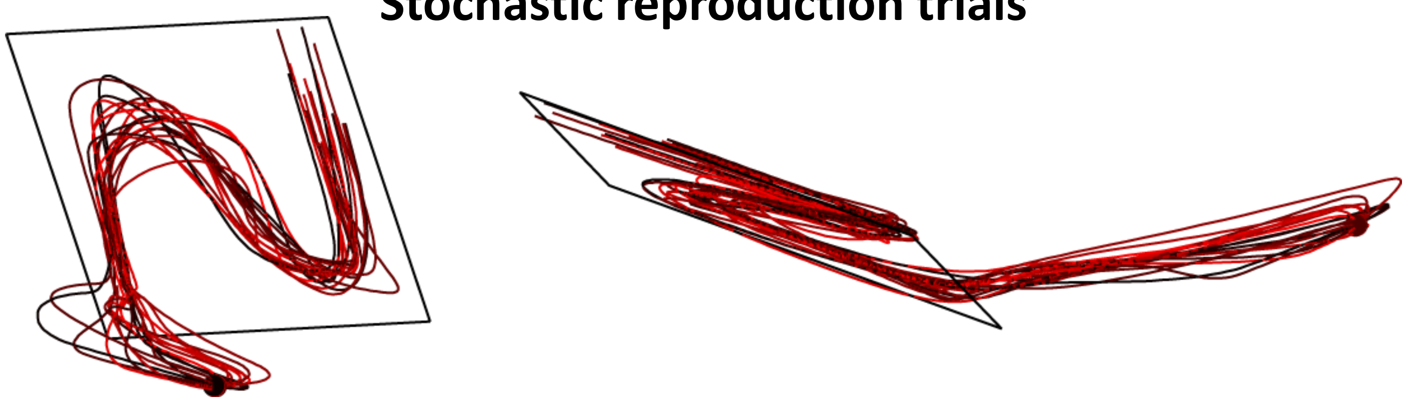
Reproductions in new situations



Stiffness ellipsoids at different time steps in the movement

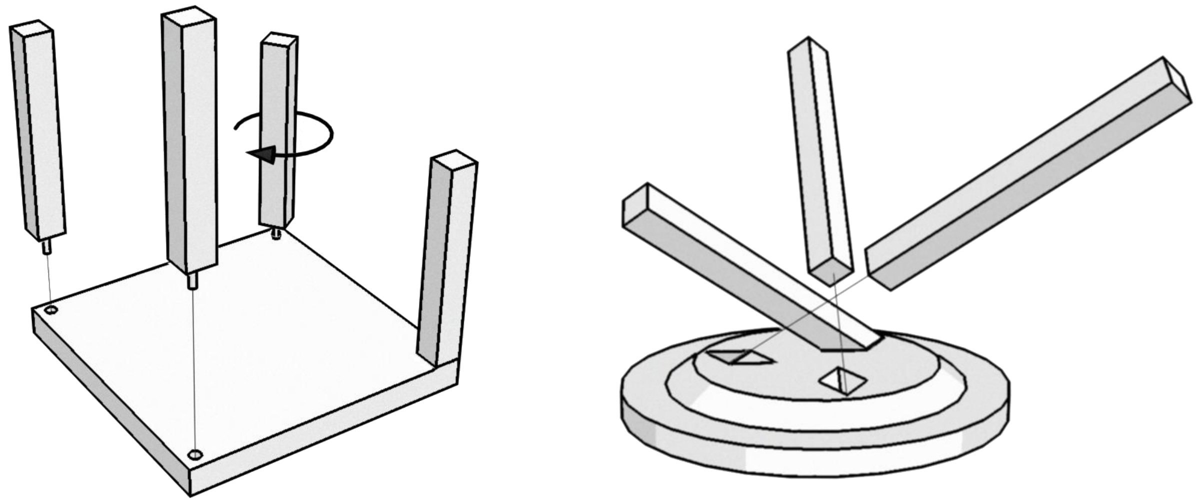


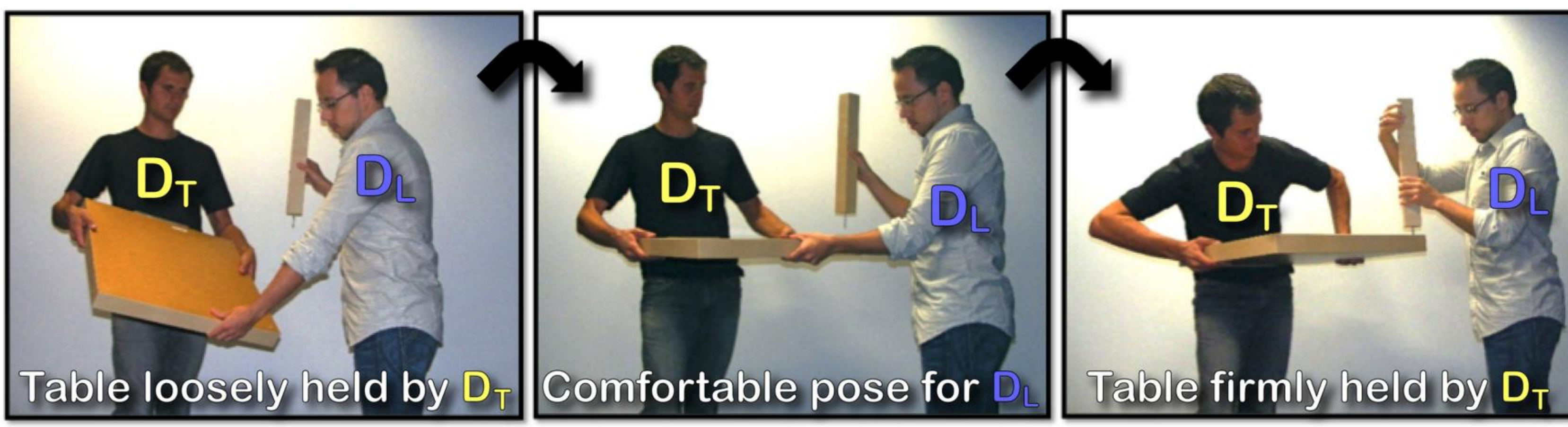
Stochastic reproduction trials



Extension to collaborative manipulation skills

Each assembly task is characterized by different sequences, positions and orientations of components, with haptic and movement patterns specific to the item to assemble.

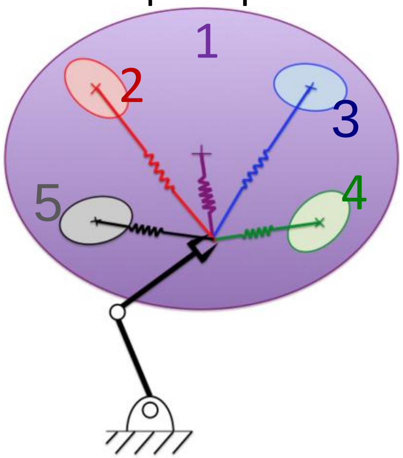




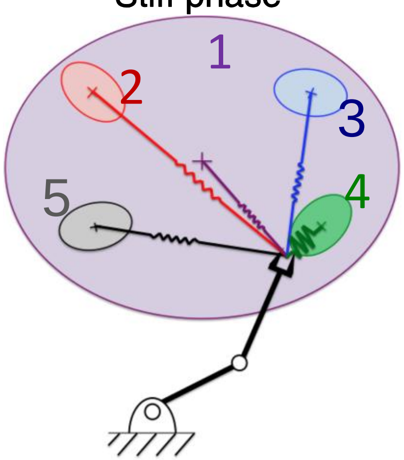
[Collaboration between IIT and IRI, UPC, Barcelona, Spain]

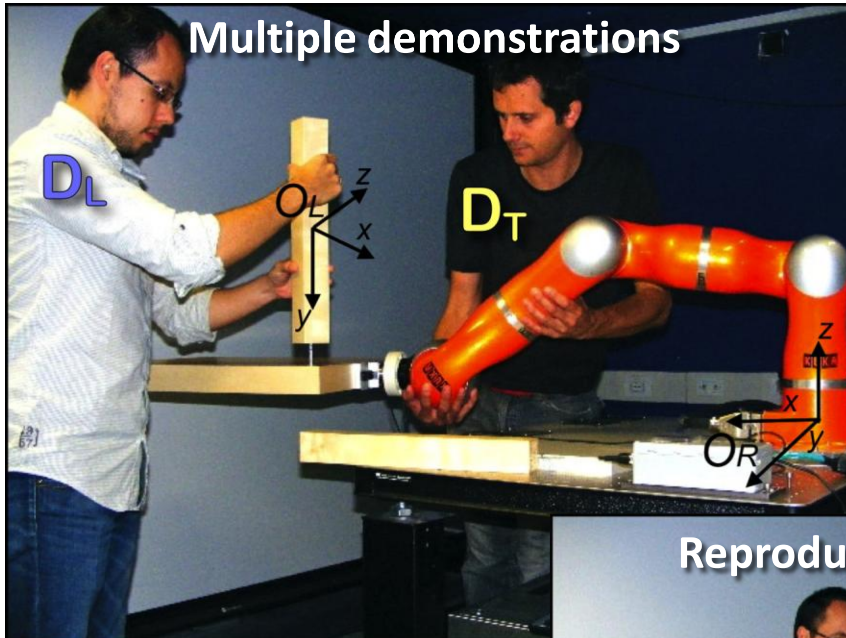
Extension to collaborative manipulation skills

Compliant phase



Stiff phase





[Collaboration between IIT and IRI, UPC, Barcelona, Spain]

EM-based Reinforcement Learning

[Petar Kormushev, Sylvain Calinon and Darwin Caldwell, IROS’2010]

EM-based Reinforcement Learning

[Petar Kormushev, Sylvain Calinon and Darwin Caldwell, IROS'2010]

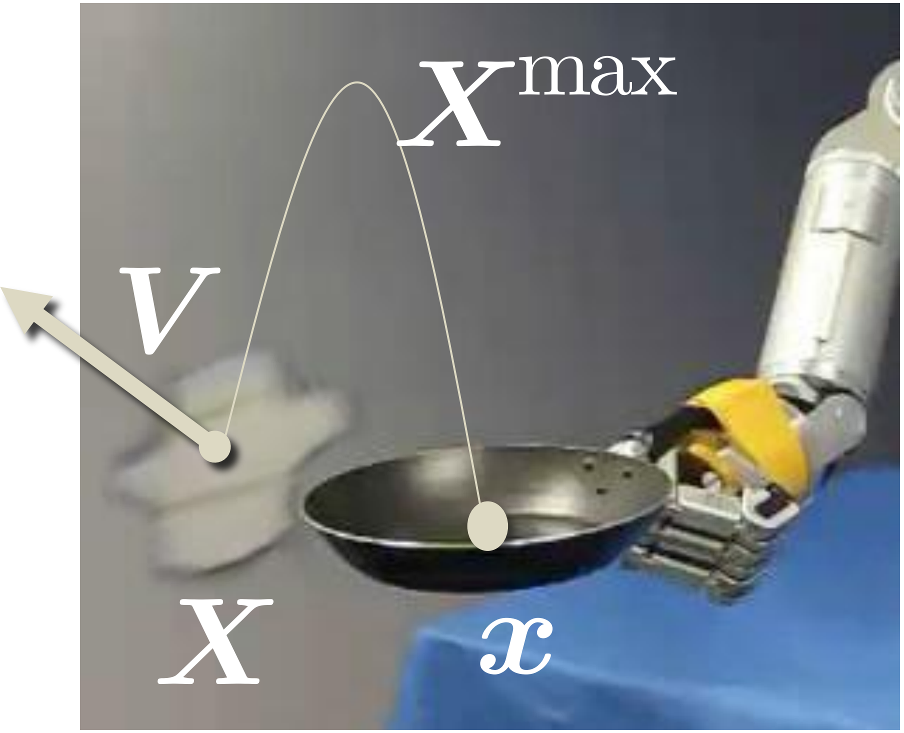
EM-based Reinforcement Learning

[Petar Kormushev, Sylvain Calinon and Darwin Caldwell, IROS'2010]

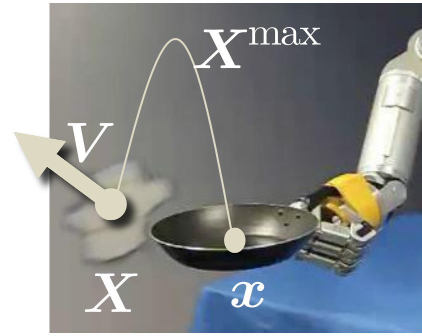
EM-based Reinforcement Learning

Episodic reward of policy $\boldsymbol{\Theta}_k$:

$$r(\boldsymbol{\Theta}_k) = \alpha_1 \frac{\arccos(\boldsymbol{V}_0 \boldsymbol{V}^\top)}{\pi} + \alpha_2 \exp(-|\boldsymbol{X} - \boldsymbol{x}|) + \alpha_3 \boldsymbol{X}_3^{\max}$$



EM-based Reinforcement Learning



Episodic reward of policy $\boldsymbol{\Theta}_k$:

$$r(\boldsymbol{\Theta}_k) = \alpha_1 \frac{\arccos(\boldsymbol{V}_0 \boldsymbol{V}^\top)}{\pi} + \alpha_2 \exp(-|\boldsymbol{X} - \boldsymbol{x}|) + \alpha_3 \boldsymbol{X}_3^{\max}$$

EM-based RL algorithm:

PoWER *(Policy learning by Weighting Exploration with the Returns)*

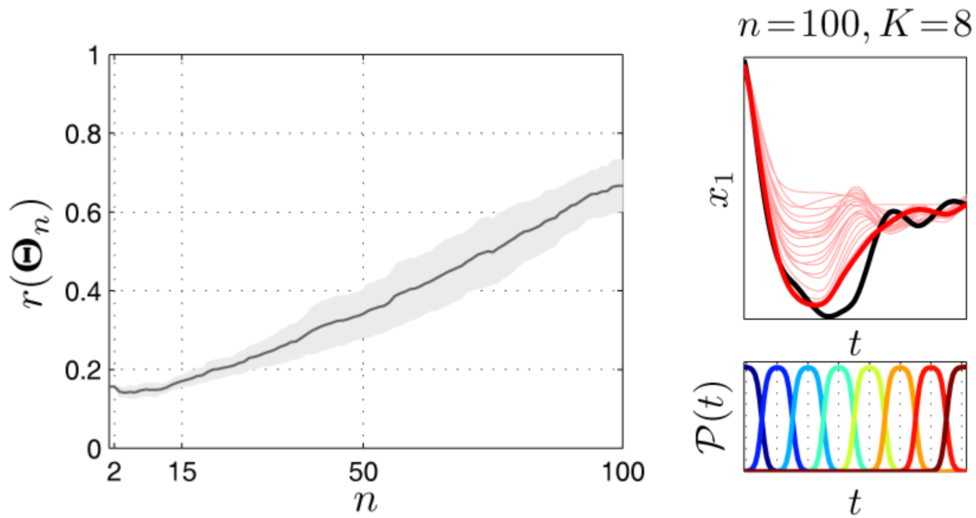
For an ordered set of policies $\{\boldsymbol{\Theta}_k\}_{k=1}^K$, with $r(\boldsymbol{\Theta}_1) \geq r(\boldsymbol{\Theta}_2) \geq \ldots,$ the update rule at each iteration *n* is defined as:

$$\boldsymbol{\Theta}^{(n)} = \boldsymbol{\Theta}^{(n-1)} + \frac{\sum_k^K r(\boldsymbol{\Theta}_k) \left[\boldsymbol{\Theta}_k - \boldsymbol{\Theta}^{(n-1)}\right]}{\sum_k^K r(\boldsymbol{\Theta}_k)}$$

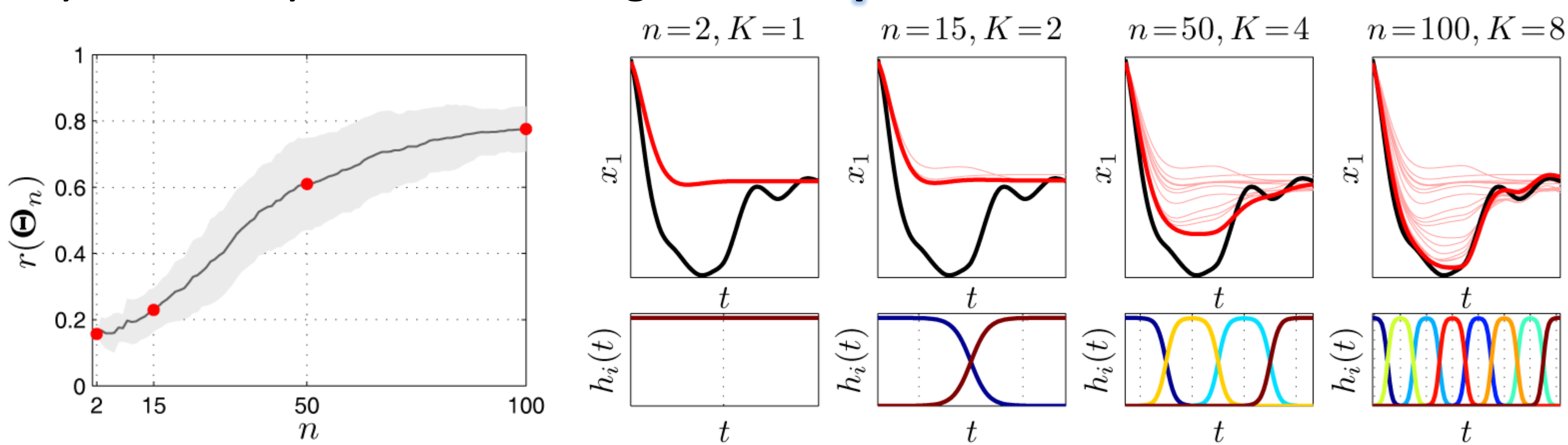
[J. Kober and J. Peters, IEEE RAM 17(2), 2010]

RL with adaptive resolution in the policy

Dynamical systems encoding with **fixed resolution**:

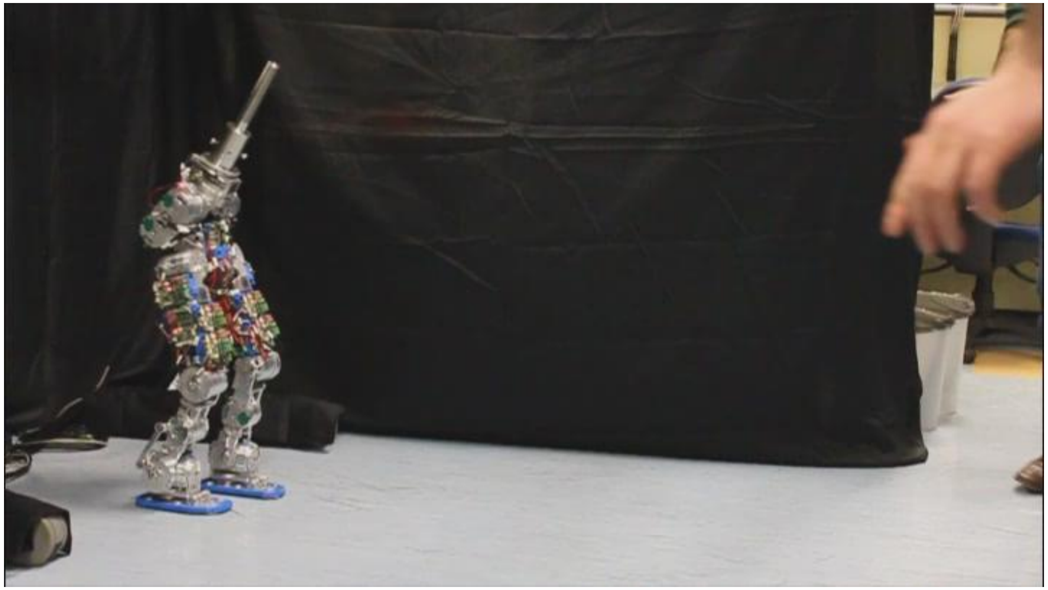


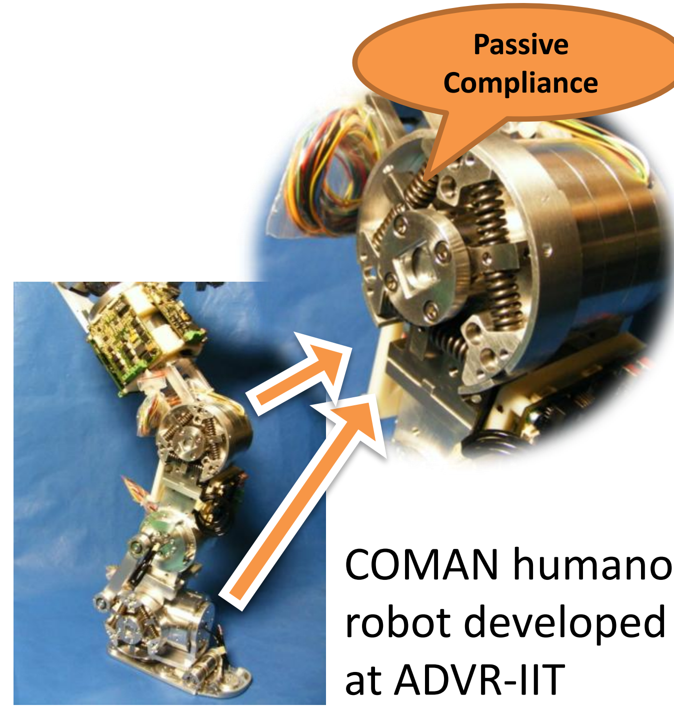
Dynamical systems encoding with **adaptive resolution**:



RL with adaptive resolution in the policy

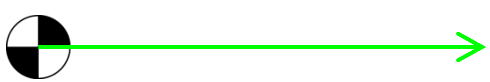
Conventional ZMP-based dynamic walking



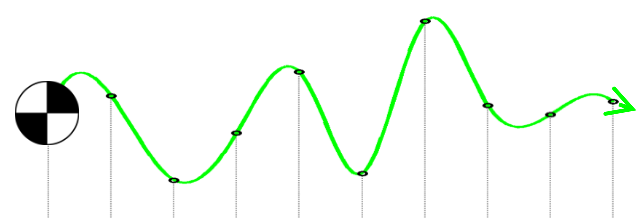


COMAN humanoid robot developed at ADVR-IIT

Fixed CoM height

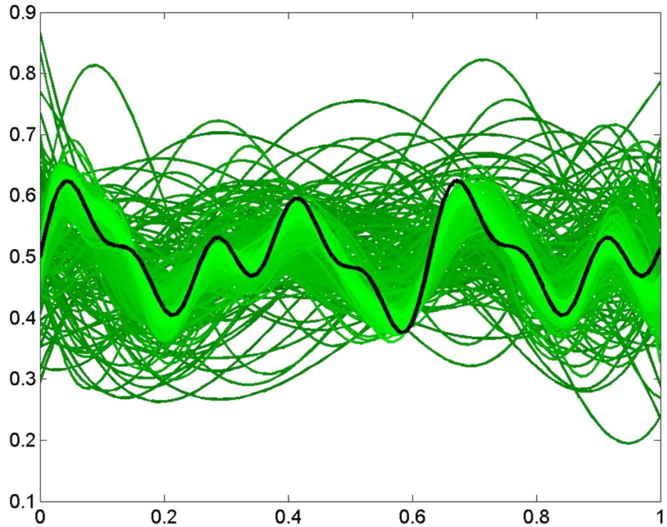


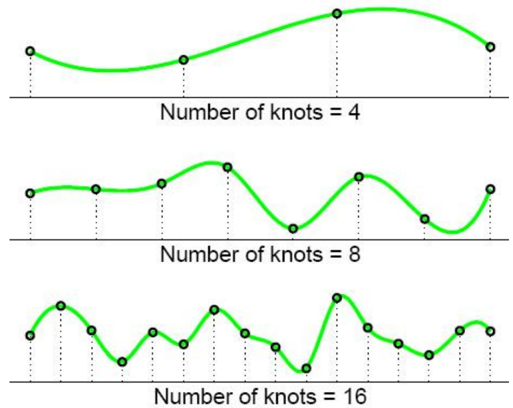
Variable CoM height



[P. Kormushev, B. Ugurlu, S. Calinon, N.G. Tsagarakis and D.G. Caldwell, IROS'2011]

RL with adaptive resolution in the policy





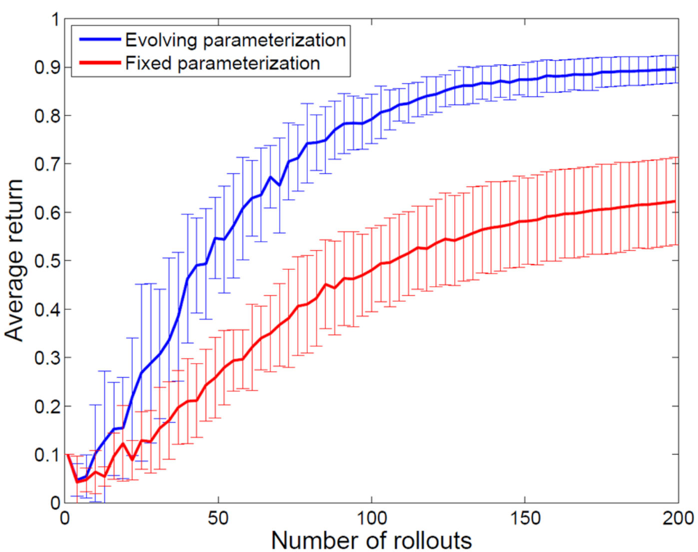
$$E_j(t_1, t_2) = \int_{t_1}^{t_2} I_j(t) U_j(t)\, dt$$

current voltage

$$E(\tau) = \frac{1}{c} \sum_{j \in J} E_j(t_1, t_2)$$

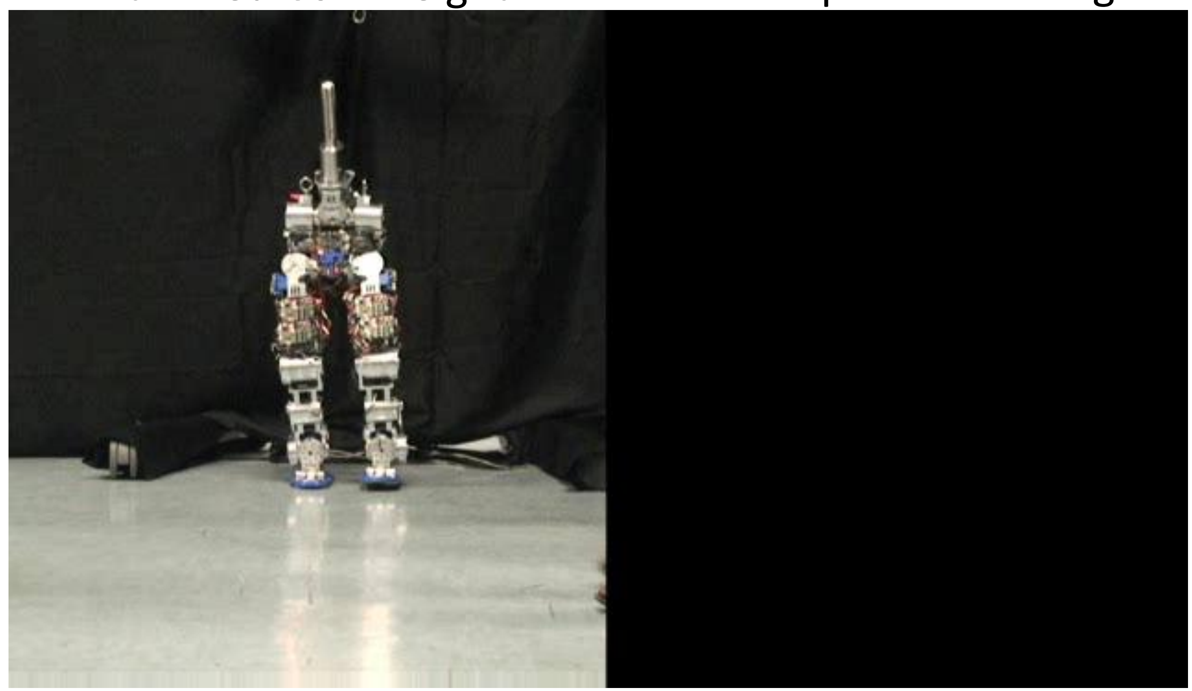
time interval

$$R(\tau) = e^{-kE(\tau)}$$ Return of a roll-out



With fixed CoM height

With adaptive CoM height



[P. Kormushev, B. Ugurlu, S. Calinon, N.G. Tsagarakis and D.G. Caldwell, IROS'2011]

Multidimensional rewards in EM-based RL

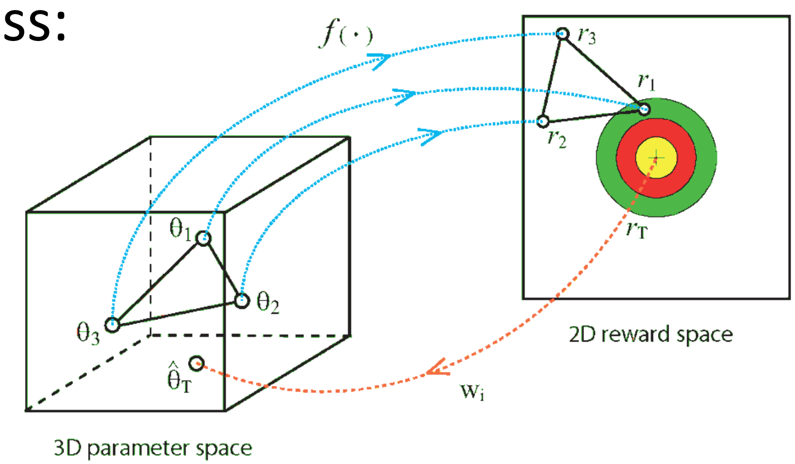
PoWER: [Kober and Peters, RAM 17(2), 2010]

$$r(\boldsymbol{\Theta}_k) = \alpha_1 r_1(\boldsymbol{\Theta}_k) + \alpha_2 r_2(\boldsymbol{\Theta}_k) + \alpha_3 r_3(\boldsymbol{\Theta}_k)$$

$$\boldsymbol{\Theta}^{(n)} = \boldsymbol{\Theta}^{(n-1)} + \frac{\sum_k^K r(\boldsymbol{\Theta}_k)\left[\boldsymbol{\Theta}_k - \boldsymbol{\Theta}^{(n-1)}\right]}{\sum_k^K r(\boldsymbol{\Theta}_k)}$$

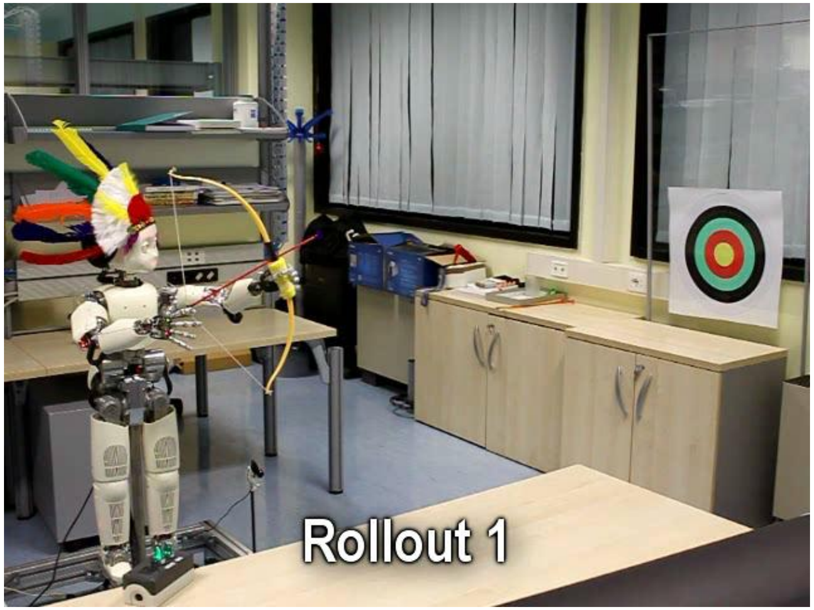
$$\boldsymbol{r}(\boldsymbol{\Theta}_k) = \begin{bmatrix} r_1(\boldsymbol{\Theta}_k) \\ r_2(\boldsymbol{\Theta}_k) \\ r_3(\boldsymbol{\Theta}_k) \end{bmatrix}$$

In some tasks, the desired outcome (maximum reward) is known, which can be exploited in the RL process:



[P. Kormushev, S. Calinon, R. Saegusa and G. Metta, Humanoids'2010]

ARCHER (Augmented Reward CHainEd Regression)



$$\boldsymbol{r}(\boldsymbol{\Theta}_k) = \begin{bmatrix} r_1(\boldsymbol{\Theta}_k) \\ r_2(\boldsymbol{\Theta}_k) \\ r_3(\boldsymbol{\Theta}_k) \end{bmatrix}$$

$$? = \sum_k^K \hat{w}_k \boldsymbol{\Theta}_k = \hat{\boldsymbol{W}} \boldsymbol{T}$$

$$\overbrace{\begin{bmatrix} \boldsymbol{\Theta}_1 \\ \boldsymbol{\Theta}_2 \\ \boldsymbol{\Theta}_3 \\ \vdots \end{bmatrix}}^{\boldsymbol{T}} \rightarrow \overbrace{\begin{bmatrix} \boldsymbol{r}(\boldsymbol{\Theta}_1) \\ \boldsymbol{r}(\boldsymbol{\Theta}_2) \\ \boldsymbol{r}(\boldsymbol{\Theta}_3) \\ \vdots \end{bmatrix}}^{\boldsymbol{R}}$$

$$\begin{bmatrix} ? \end{bmatrix} \rightarrow \begin{bmatrix} \boldsymbol{r}^{\max} \end{bmatrix}$$

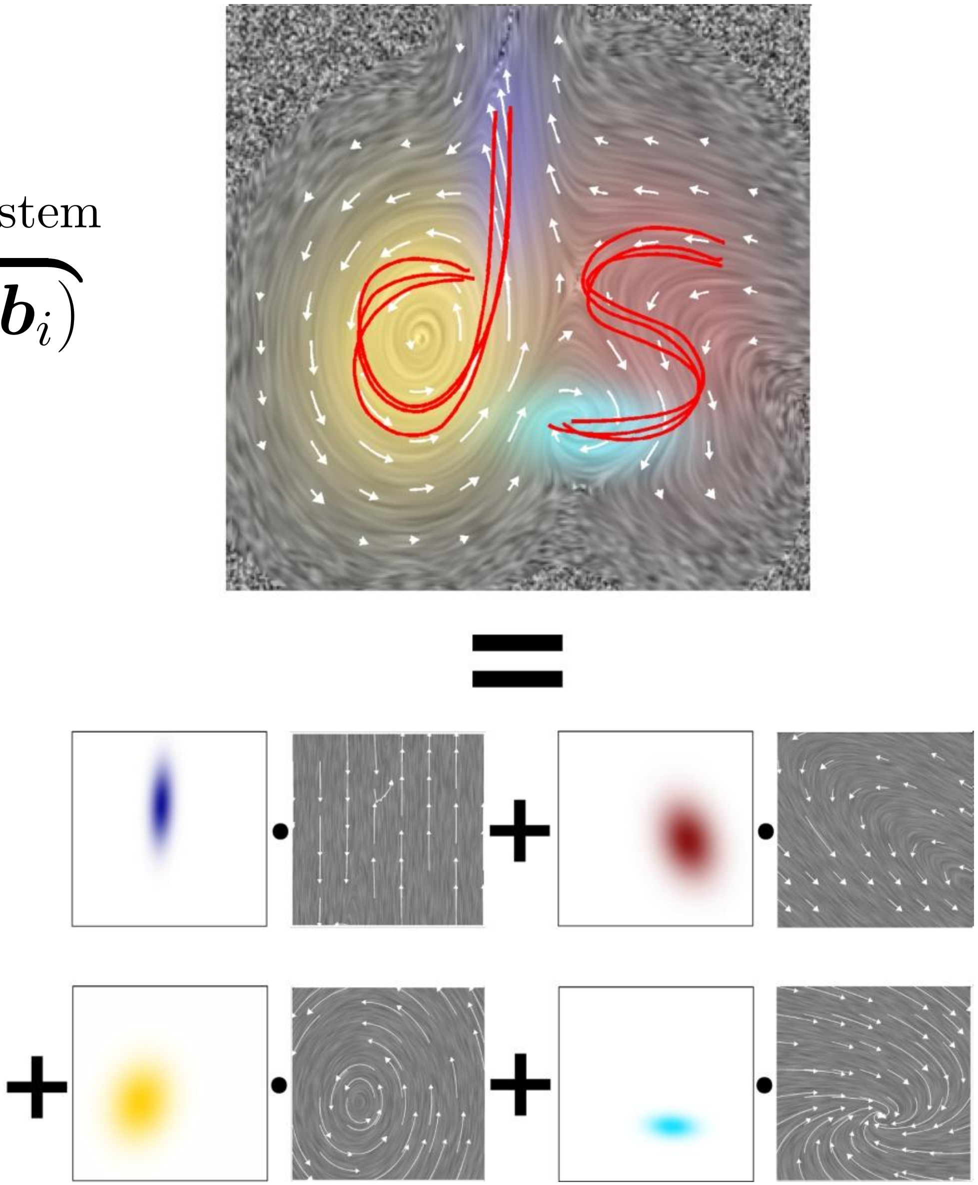
$$\boldsymbol{r}^{\max} = \sum_k^K w_k \boldsymbol{r}(\boldsymbol{\Theta}_k) = \boldsymbol{W}\boldsymbol{R}$$

$$\hat{\boldsymbol{W}} = \boldsymbol{r}^{\max} \boldsymbol{R}^{+}$$

[P. Kormushev, S. Calinon, R. Saegusa and G. Metta, Humanoids'2010]

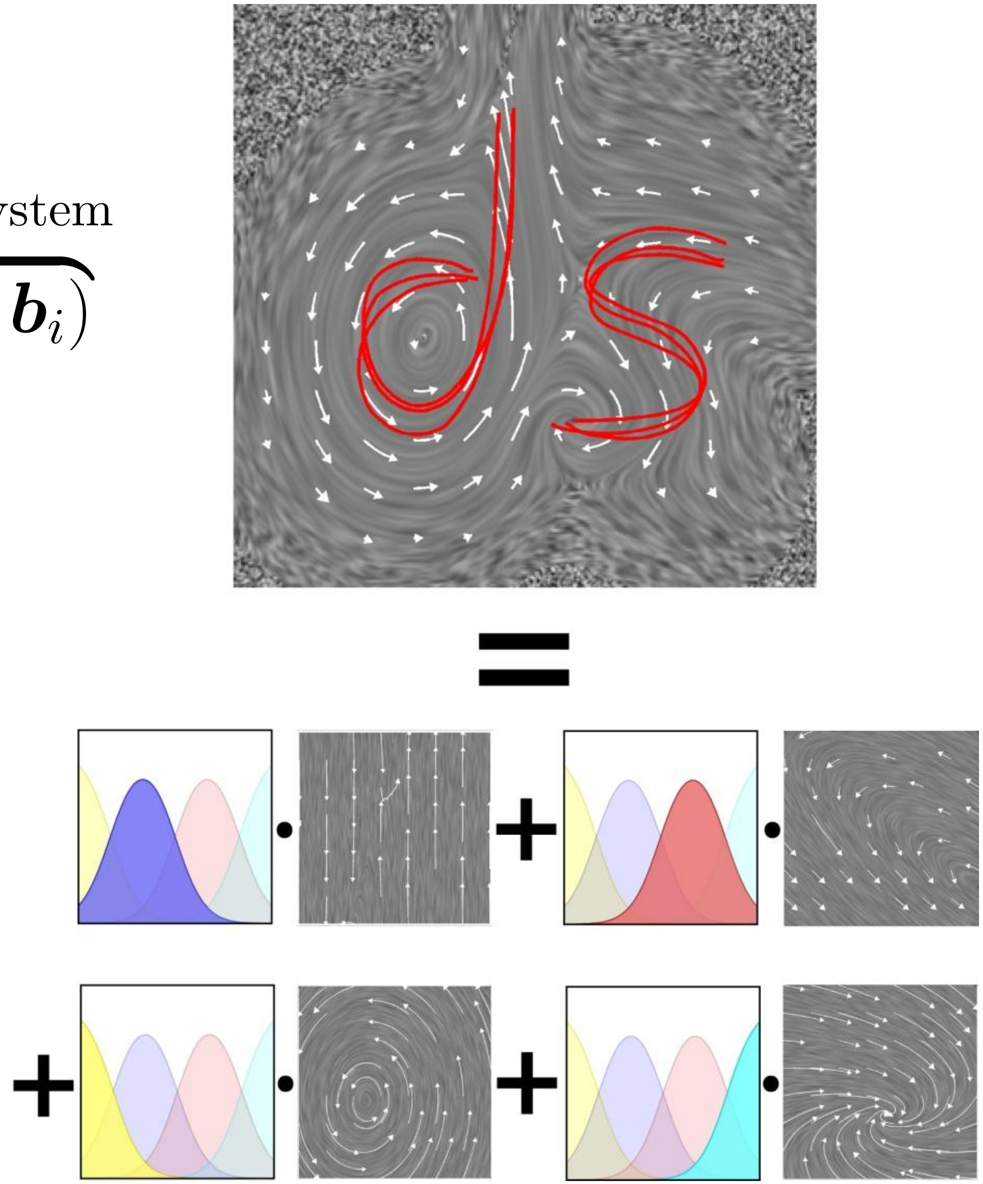
Consideration of time and space constraints in the weighting mechanism

$$\dot{\boldsymbol{x}} = \sum_i \overbrace{h_i(\boldsymbol{x})}^{\text{scalar weight}} \; \overbrace{(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i)}^{\text{linear subsystem}}$$

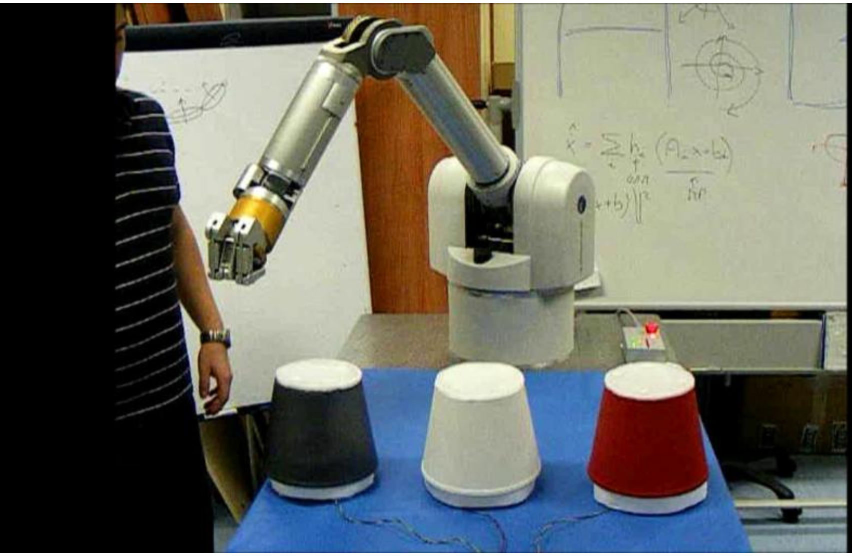


Consideration of time and space constraints in the weighting mechanism

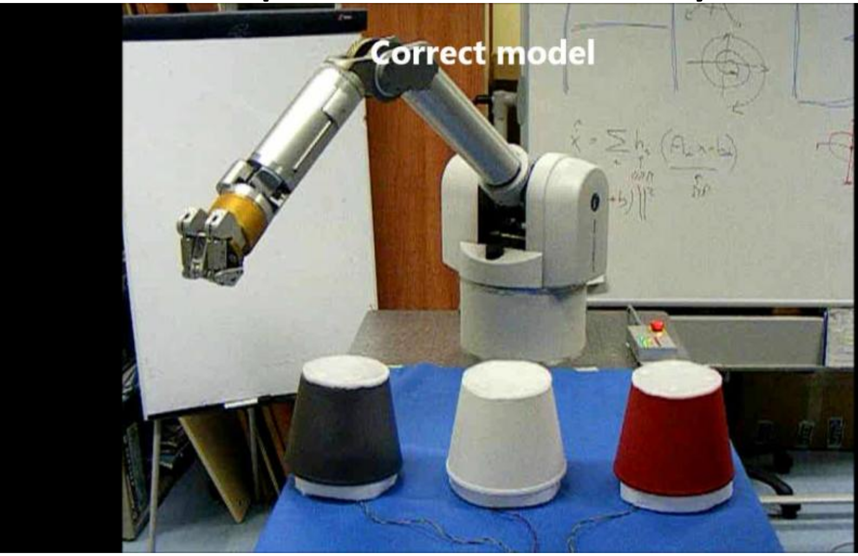
$$\dot{\boldsymbol{x}} = \sum_i \overbrace{h_i(t)}^{\text{scalar weight}} \overbrace{(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i)}^{\text{linear subsystem}}$$

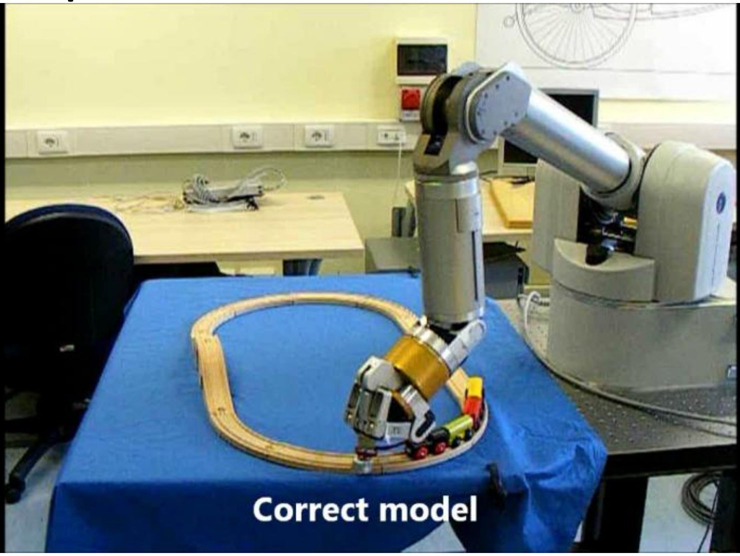


Which weighting mechanism to use?



Task-dependent recovery strategies after perturbation:



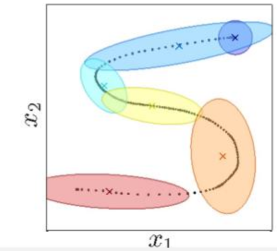


[Sylvain Calinon, Antonio Pistillo and Darwin Caldwell, IROS'2011]

Which weighting mechanism to use?

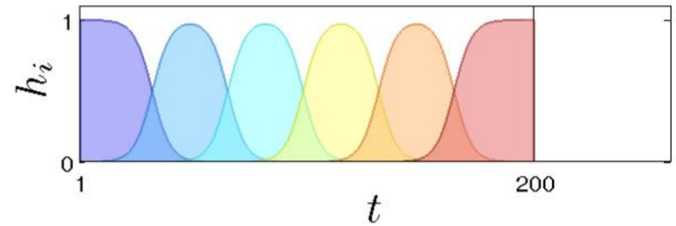
Gaussian Mixture Model (GMM)

$$\alpha_i^{\text{GMM}} = \mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu}_i^{\mathcal{X}}, \boldsymbol{\Sigma}_i^{\mathcal{X}})$$



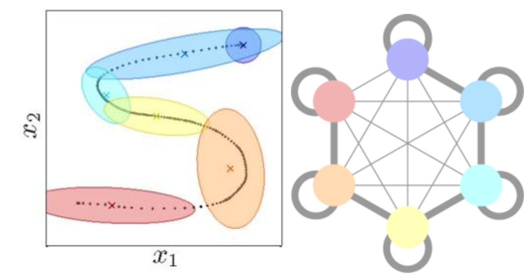
Time-based weighting mechanism

$$\alpha_i^{\text{TIME}} = \mathcal{N}(t;\ \mu_i^{\mathcal{T}}, \Sigma_i^{\mathcal{T}})$$



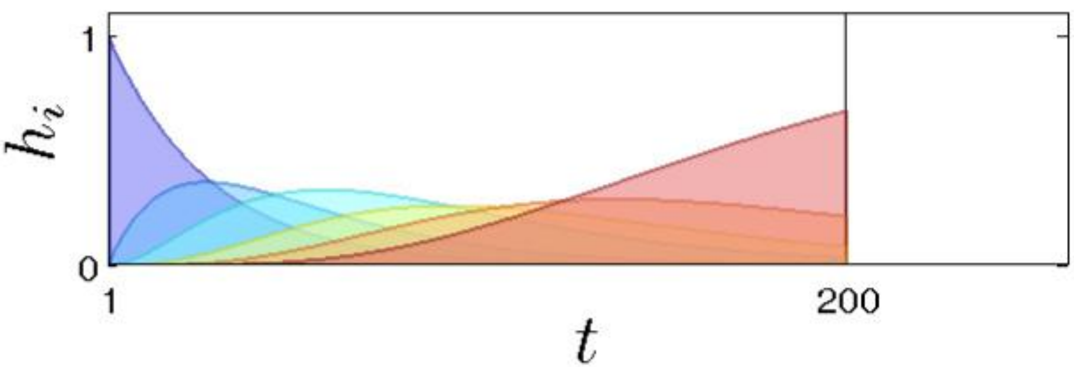
Hidden Markov Model (HMM)

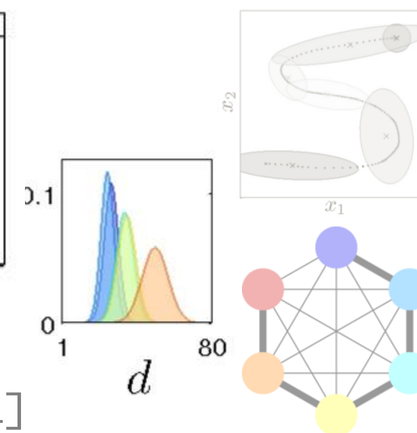
$$\alpha_{i,n}^{\text{HMM}} = \Big(\sum_{j=1}^{K} \alpha_{j,n-1}^{\text{HMM}}\ a_{j,i}\Big)\mathcal{N}(\boldsymbol{x}_n;\ \boldsymbol{\mu}_i^{\mathcal{X}}, \boldsymbol{\Sigma}_i^{\mathcal{X}})$$



Hidden Semi

$$\alpha_{i,n}^{\text{HSMM}} = \sum_{j=1}^{K} \sum \ldots$$

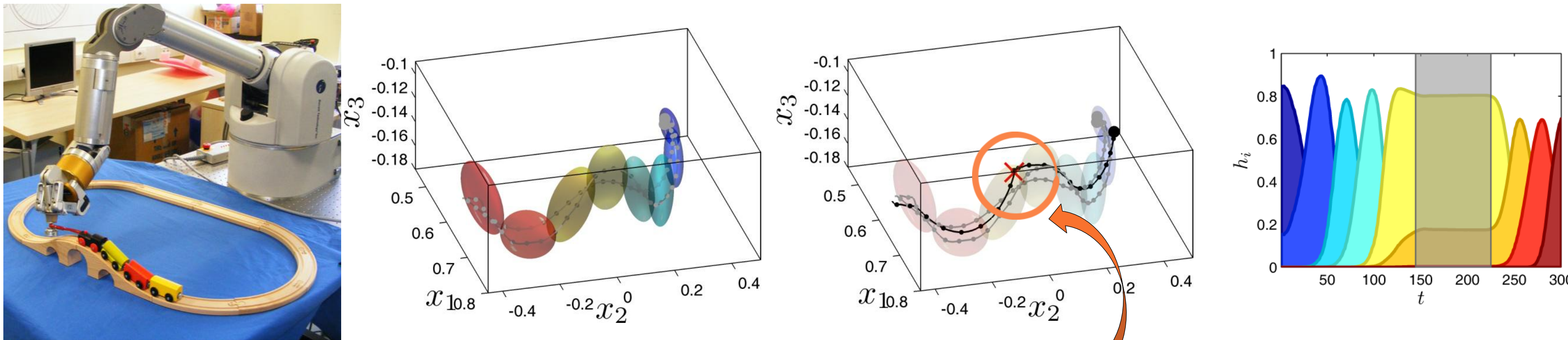




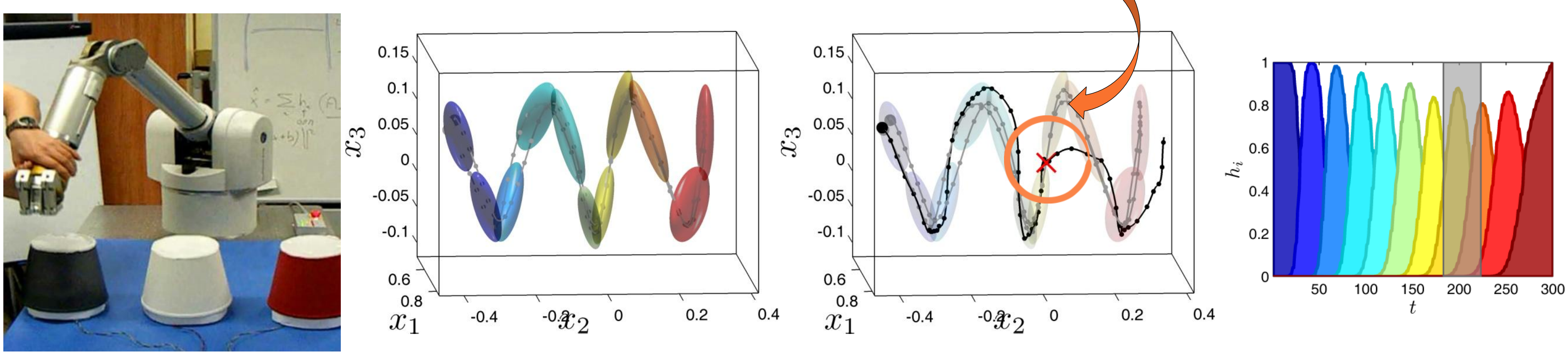
[Yu and Kobayashi, ...]

[Sylvain Calinon, Antonio Pistillo and Darwin Caldwell, IROS’2011]

Generic weighting mechanism based on Hidden Semi-Markov Model (HSMM)

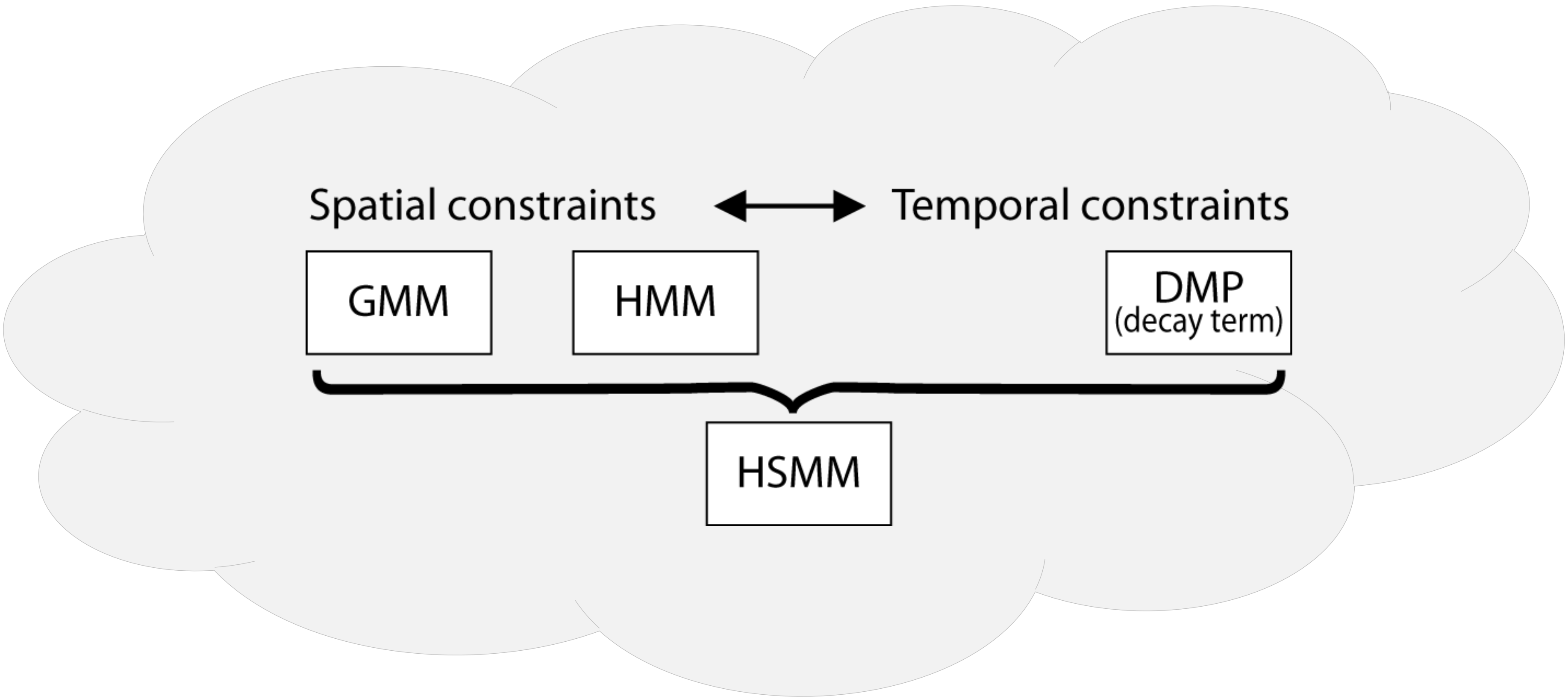


Perturbation from the user holding the robot



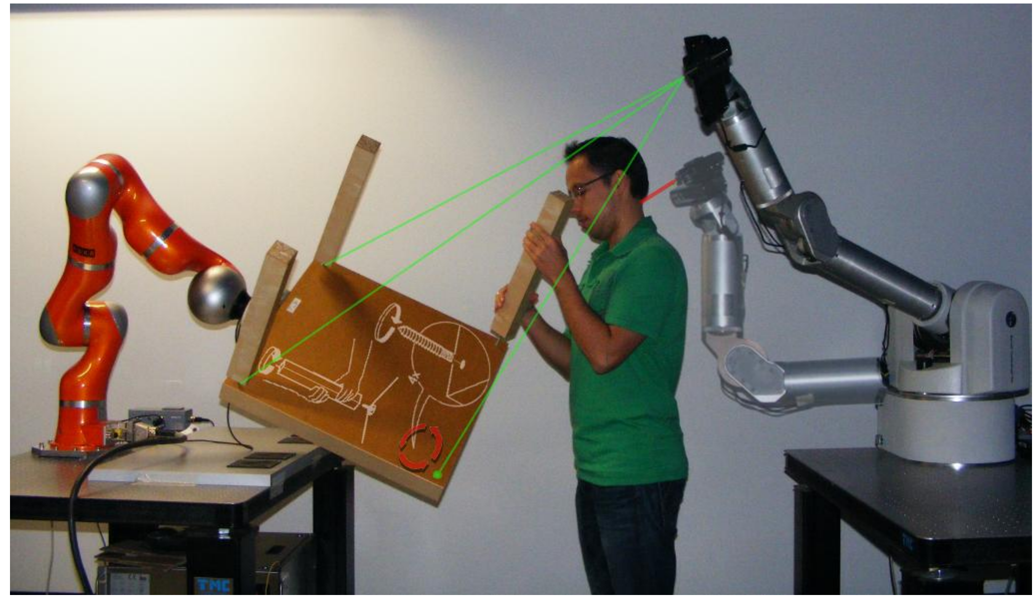
[Sylvain Calinon, Antonio Pistillo and Darwin Caldwell, IROS'2011]

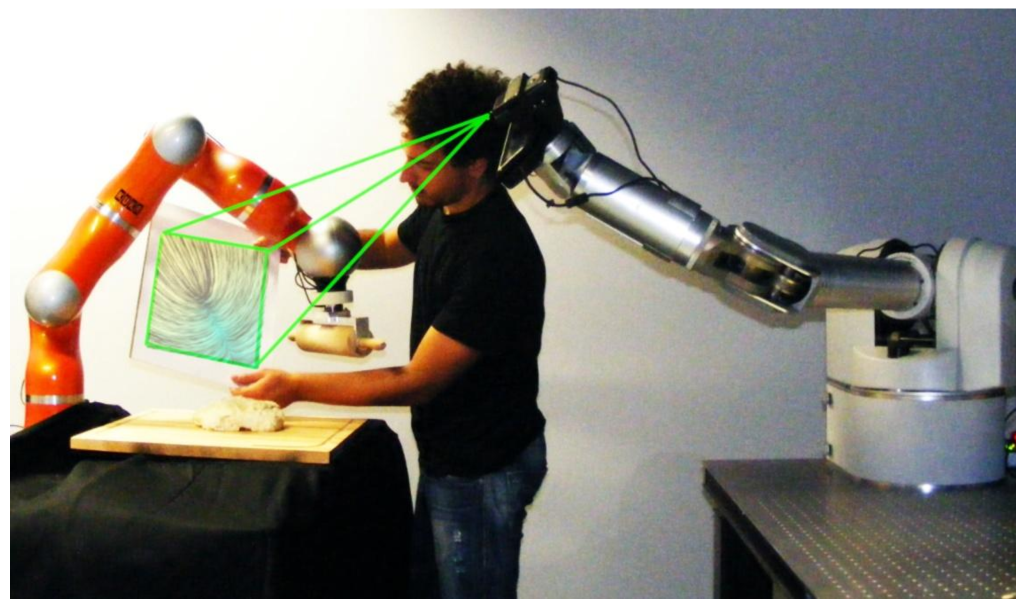
Generic weighting mechanism based on Hidden Semi-Markov Model (HSMM)



[Sylvain Calinon, Antonio Pistillo and Darwin Caldwell, IROS’2011]

Active visualization and assessment of skills





[De Tommaso, Calinon and Caldwell, Intl Journal of Social Robotics (in press)]

Conclusion

The development of new actuators and control architectures is bringing a new focus on passive and active compliance, energy optimization, human-robot collaboration and safety.

Existing machine learning tools need to be re-thought and adapted to these new developments, with systems that can:

- simultaneously learn **motion and impedance behaviors**.
- exploit the **statistical information** contained in multiple demonstrations of the same task.
- be modulated with respect to **task input parameters**.
- be used in **imitation and reinforcement learning** settings.
- reproduce natural movements and reactive behaviors in a **smooth and continuous** way.
- be analyzed and visualized during the training process.