Gaussian Process Reinforcement Learning for Dialog Management

Maryam Zare

Multi Domain Dialog Management

- Traditional Spoken Dialog Systems
  - Built over a limited domain described by a an underlying ***ontology***
- How to move to bigger ontologies?
  - Challenge: the size is big so some part of space will rarely get visited!
- How to Handle multiple Domains?
  - Is there any way to combine learned knowledge from different domains?

Proposed Solution: Use Gaussian Process

What is a Gaussian Process?

- Definiton
  - Bayesian method which specifies a prior distribution over the unknown function and then given some observations estimates the posterior
  - consists of
    1. mean function (a vector of size infinity)
    2. kernel function (correlation of function values , 2D of infinite size)

- How it can be used for Reinforcement Learning?
  - “For every input point, the kernel specifies the expected variation of where the function value will lie and once given some data, the kernel therefore defines the correlations between known and unknown function values. In that way, the known function values influence the regions where we do not have any data points. When used inside a reinforcement learning framework, the variance can be used to guide exploration, avoiding the need to explore parts of the space where the Gaussian process is very certain. All this leads to very data efficient learning ”

How to Map Q function to GP

- Q function: $$Q(\boldsymbol{b}, a) = E_{\pi}\left(\sum_{\tau=t+1}^{T} \gamma^{\tau-t-1} r_{\tau} | b_t = \boldsymbol{b}, a_t = a\right)$$

- Q function as a GP[1]: $$Q(\boldsymbol{b}, a) \sim \mathcal{GP}\left(\underset{\text{Prior Mean}}{\underset{\downarrow}{m(\boldsymbol{b}, a)}}, \underset{k_{\mathcal{B}}(\boldsymbol{b}, \boldsymbol{b}')k_{\mathcal{A}}(a, a')}{\underset{\downarrow}{k((\boldsymbol{b}, a), (\boldsymbol{b}, a))}}\right)$$

- Q function for any belief state-action pair is given by:

$$Q(\boldsymbol{b}, a) | \boldsymbol{r}, \boldsymbol{B} \sim \mathcal{N}(\overline{Q}(\boldsymbol{b}, a), cov((\boldsymbol{b}, a), (\boldsymbol{b}, a)))$$

[1] Y. Engel, S. Mannor, R. Meir, Reinforcement learning with Gaussian processes, in: Proceedings of ICML, 2005

How to Map Q function to GP (Continue)

- Posterior mean and Covariance where is K the Gram Matrix and H is a band Matrix with a with diagonal $1 - \gamma$

$$\overline{Q}(\boldsymbol{b}, a) = \boldsymbol{k}(\boldsymbol{b}, a)^\top \boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{H}\boldsymbol{H}^\top)^{-1}(\boldsymbol{r} - \boldsymbol{m}),$$

$$cov((\boldsymbol{b}, a), (\boldsymbol{b}, a)) = k((\boldsymbol{b}, a), (\boldsymbol{b}, a)) - \\ \boldsymbol{k}(\boldsymbol{b}, a)^\top \boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{H}\boldsymbol{H}^\top)^{-1} \boldsymbol{H}\boldsymbol{k}(\boldsymbol{b}, a)$$

- Next Action: $\pi(\mathbf{b}) = \arg\max_a \left\{ \hat{Q}(\mathbf{b}, a) : a \in \mathcal{A} \right\}$

- Kernel Definition:

$$k_{\mathcal{A}}(a, a') = \delta_a(a')$$

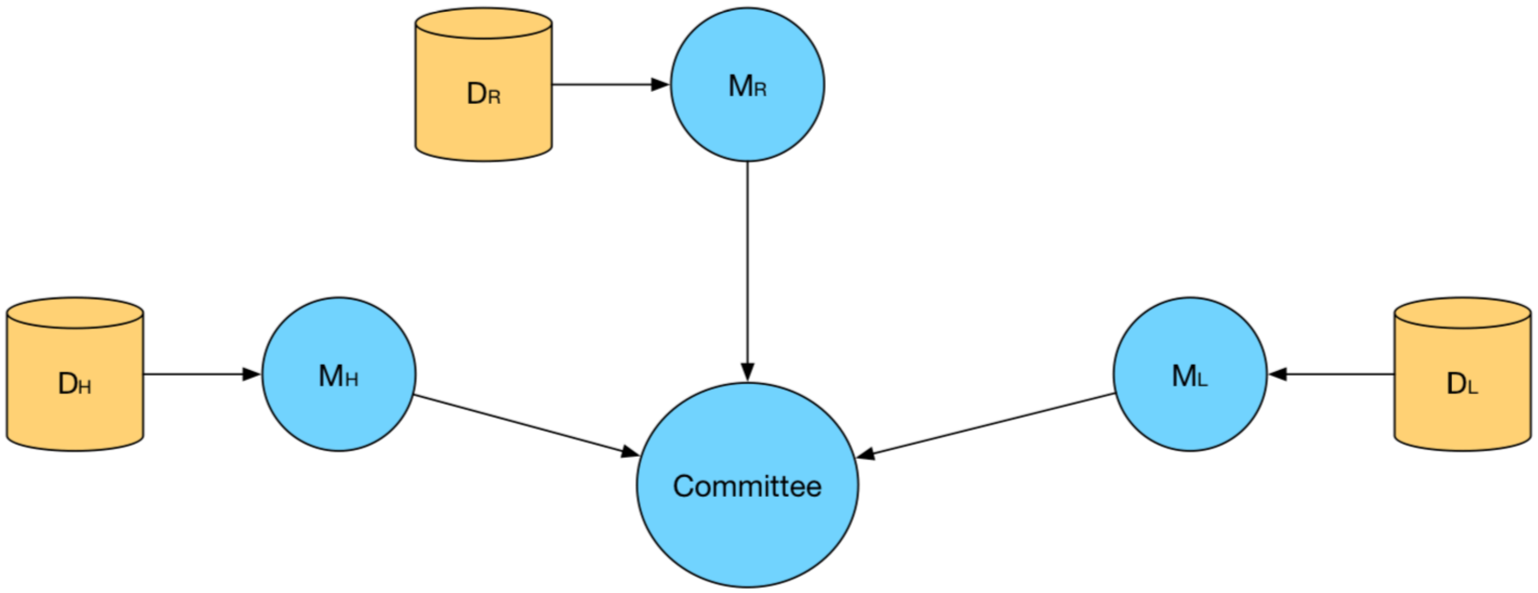
$$k_{\mathcal{B}}(\boldsymbol{b}, \boldsymbol{b}') = \sum_h \langle \boldsymbol{b}_h, \boldsymbol{b}'_h \rangle$$

Committee of Dialog Policies

- Combining multiple estimates of Q-value with a specific mean and covariance[2]

$$\overline{Q}(\boldsymbol{b},a) = \Sigma^{Q}(\boldsymbol{b},a)\sum_{i=1}^{M}\Sigma_i^{Q}(\boldsymbol{b},a)^{-1}\overline{Q}_i(\boldsymbol{b},a),$$

$$\Sigma^{Q}(\boldsymbol{b},a)^{-1} = -(M-1)*k((\boldsymbol{b},a),(\boldsymbol{b},a))^{-1}+\sum_{i=1}^{M}\Sigma_i^{Q}(\boldsymbol{b},a)^{-1}.$$



2. V. Tresp, A Bayesian Committee Machine, Neural Comput. 12 (11) (2000) 2719–2741. doi:10.1162/089976600300014908.

Multi Domain Dialog Manager

- How to operate on belief states and actions that come from different domains?
  - Define classes of topics and for each value in that class compute the entropy

$$\eta(s) = -\sum_{v \in \mathcal{V}_s} \frac{p(s=v)\log(p(s=v))}{|\mathcal{V}_s|},$$

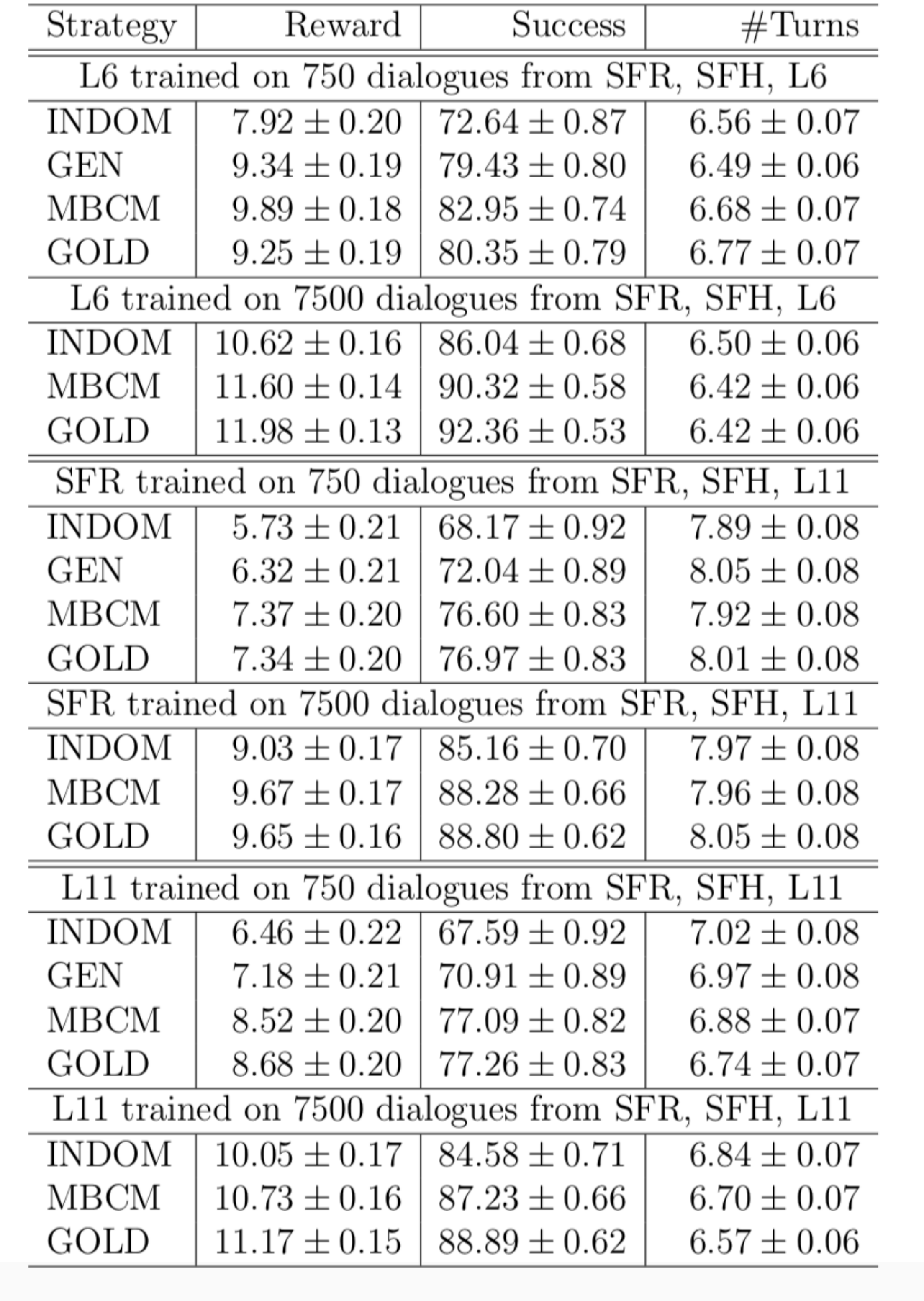
- Having slots sorted based on their entropy the kernel function is defined as below

- Iteratively, for each $slot_i^c$ where i ≤ min{$|M_c|$, $|N_c|$}, index of the ordered list, in semantic class c where $|M_c|$ denotes the number of slots in semantic class c in domain M:
  match the corresponding elements of belief space and actions, padding with zeros as necessary.
- Otherwise disregard the elements of the belief state relating to unpaired slots j and if one of the actions relates to $slot_j$, consider the action kernel to be 0.

Experiment Results

- Data
  - **SFR** consisting of restaurants in San Francisco
  - **SFH** consisting of hotels in San Francisco
  - **L6** consisting of laptops with 6 properties that the user can specify
  - **L11** same as L6 but with 11 user-specifiable properties.

- Strategy:
  - **INDOM:** In-domain policy trained only on in-domain data, other data is not taken into consideration, action-selection is based only on the in-domain policy.
  - **GEN** :Single generic policy one policy trained on all available data
  - **MBCM**: Multi-policy Bayesian committee machine
  - **GOLD**: Gold standard . this is the performance of the single policy where all training data comes from the same domain i.e. for N domains, GOLD has N times the number of in-domain dialogues for training as provided to INDOM.



Strategy	Reward	Success	#Turns
L6 trained on 750 dialogues from SFR, SFH, L6			
INDOM	7.92 ± 0.20	72.64 ± 0.87	6.56 ± 0.07
GEN	9.34 ± 0.19	79.43 ± 0.80	6.49 ± 0.06
MBCM	9.89 ± 0.18	82.95 ± 0.74	6.68 ± 0.07
GOLD	9.25 ± 0.19	80.35 ± 0.79	6.77 ± 0.07
L6 trained on 7500 dialogues from SFR, SFH, L6			
INDOM	10.62 ± 0.16	86.04 ± 0.68	6.50 ± 0.06
MBCM	11.60 ± 0.14	90.32 ± 0.58	6.42 ± 0.06
GOLD	11.98 ± 0.13	92.36 ± 0.53	6.42 ± 0.06
SFR trained on 750 dialogues from SFR, SFH, L11			
INDOM	5.73 ± 0.21	68.17 ± 0.92	7.89 ± 0.08
GEN	6.32 ± 0.21	72.04 ± 0.89	8.05 ± 0.08
MBCM	7.37 ± 0.20	76.60 ± 0.83	7.92 ± 0.08
GOLD	7.34 ± 0.20	76.97 ± 0.83	8.01 ± 0.08
SFR trained on 7500 dialogues from SFR, SFH, L11			
INDOM	9.03 ± 0.17	85.16 ± 0.70	7.97 ± 0.08
MBCM	9.67 ± 0.17	88.28 ± 0.66	7.96 ± 0.08
GOLD	9.65 ± 0.16	88.80 ± 0.62	8.05 ± 0.08
L11 trained on 750 dialogues from SFR, SFH, L11			
INDOM	6.46 ± 0.22	67.59 ± 0.92	7.02 ± 0.08
GEN	7.18 ± 0.21	70.91 ± 0.89	6.97 ± 0.08
MBCM	8.52 ± 0.20	77.09 ± 0.82	6.88 ± 0.07
GOLD	8.68 ± 0.20	77.26 ± 0.83	6.74 ± 0.07
L11 trained on 7500 dialogues from SFR, SFH, L11			
INDOM	10.05 ± 0.17	84.58 ± 0.71	6.84 ± 0.07
MBCM	10.73 ± 0.16	87.23 ± 0.66	6.70 ± 0.07
GOLD	11.17 ± 0.15	88.89 ± 0.62	6.57 ± 0.06

References

1. Y. Engel, S. Mannor, R. Meir, ***Reinforcement learning with Gaussian processes***, in: Proceedings of ICML, 2005
2. V. Tresp, ***A Bayesian Committee Machine, Neural Comput***. 12 (11) (2000) 2719–2741. doi: 10.1162/089976600300014908.
3. M. Gasic, N. Mrksic, L. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, T.-H. Wen and S. Young (2017). "***Dialogue manager domain adaptation using Gaussian process reinforcement learning***." Computer Speech and Language, 45(5):552-569
4. Helpful Video for understanding GP: ***https://www.youtube.com/watch?v=92-98SYOdlY***