Towards Reliable and Efficient AI for Communications via Bayesian Meta-Learning

Osvaldo Simeone
Joint work with Kfir Cohen, Sangwoo Park, and Shlomo Shamai (Shitz)

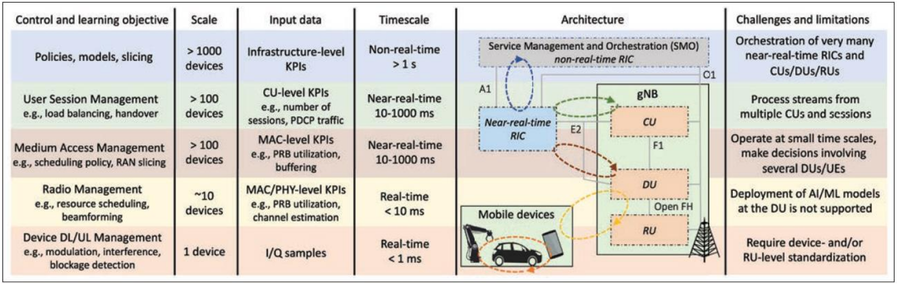
King's College London

MLCOM 2022, 24/3/2022

Motivation

The Role of AI in 6G & Beyond

- AI is playing an increasingly significant role in engineering.
- As a case in point, next-generation communication systems will leverage AI at all layers of the protocol stack.

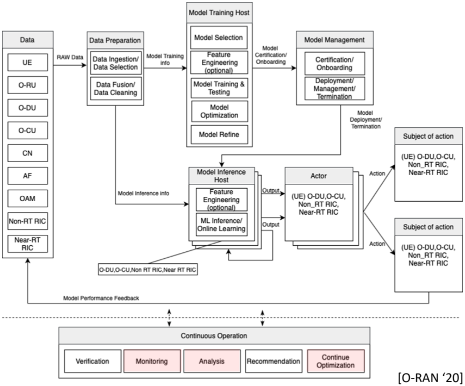


Control and learning objective	Scale	Input data	Timescale	Architecture	Challenges and limitations
Policies, models, slicing	> 1000 devices	Infrastructure-level KPIs	Non-real-time > 1 s		Orchestration of very many near-real-time RICs and CUs/DUs/RUs
User Session Management e.g., load balancing, handover	> 100 devices	CU-level KPIs e.g., number of sessions, PDCP traffic	Near-real-time 10-1000 ms		Process streams from multiple CUs and sessions
Medium Access Management e.g., scheduling policy, RAN slicing	> 100 devices	MAC-level KPIs e.g., PRB utilization, buffering	Near-real-time 10-1000 ms		Operate at small time scales, make decisions involving several DUs/UEs
Radio Management e.g., resource scheduling, beamforming	~10 devices	MAC/PHY-level KPIs e.g., PRB utilization, channel estimation	Real-time < 10 ms		Deployment of AI/ML models at the DU is not supported
Device DL/UL Management e.g., modulation, interference, blockage detection	1 device	I/Q samples	Real-time < 1 ms		Require device- and/or RU-level standardization

[Bonati et al '21]

The Life Cycle of an AI Model

- In many engineering problems (e.g., in digital twin platforms), AI modules should be:
- 1) **well calibrated**, providing a faithful quantification of the uncertainty of their decisions, e.g., for **monitoring**
- 2) **sample efficient**, enabling fast **adaptation**



[O-RAN '20]

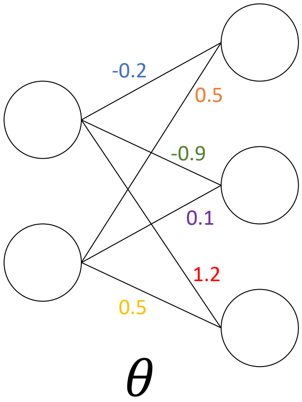
This Talk

- **Reliable** AI, enabling monitoring and analysis:
  - Bayesian learning
- **Sample-efficient** AI, enabling fast adaptation:
  - Meta-learning
- **Reliable and sample-efficient** AI:
  - Bayesian meta-learning

Reliable AI: Bayesian Learning

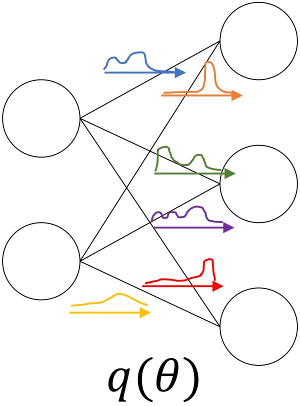
Frequentist Learning vs. Bayesian Learning

- **Frequentist** learning (e.g., standard deep learning):
  - Optimization of a single model parameter vector θ
  - Decision based on a single model $p(x|\theta)$

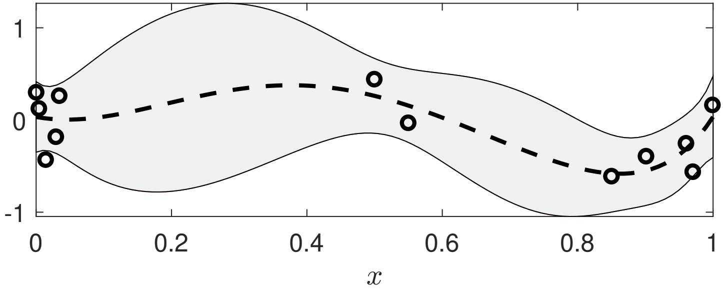


Frequentist Learning vs. Bayesian Learning

- **Bayesian** learning:
  - Optimization of a distribution $q(\theta)$ in the model parameter space
  - Decision obtained via ensembling, i.e., via $\mathrm{E}_{\theta\sim q(\theta)}[p(x|\theta)]$



Frequentist Learning vs. Bayesian Learning



- Bayesian learning leverages the disagreement of models outside the training data to quantify epistemic uncertainty.
- Other advantages: improved generalization, active learning, online learning, efficient distributed/federated learning[1],...

---

[1] R. Kassab and O. Simeone, "Federated generalized Bayesian learning via distributed Stein variational gradient descent," arXiv:2009.06419, 2020.

Frequentist Learning vs. Bayesian Learning

- Given a training set $\mathcal{D}$, conventional frequentist learning minimizes the training loss $L_{\mathcal{D}}(\theta)$ over θ.
- Bayesian learning minimizes the variational **free energy**

$$F_{\mathcal{D}}(q(\theta)) = \underbrace{\mathrm{E}_{\theta\sim q(\theta)}[L_{\mathcal{D}}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}},$$

  over distribution $q(\theta)$, where $p_0(\theta)$ is a prior distribution.
- Without regularization, the problem reduces to standard frequentist learning.
- The regularization term provides a bound on the **generalization error** via PAC Bayes theory, and it underlies the free energy principle in neuroscience.[2]

[2] S. T. Jose and O. Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization," IEEE Signal Processing Magazine, 2021.

Frequentist Learning vs. Bayesian Learning

- Given a training set $\mathcal{D}$, conventional frequentist learning minimizes the training loss $L_{\mathcal{D}}(\theta)$ over θ.
- Bayesian learning minimizes the variational **free energy**

$$F_{\mathcal{D}}(q(\theta)) = \underbrace{\mathrm{E}_{\theta \sim q(\theta)}[L_{\mathcal{D}}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}},$$

over distribution $q(\theta)$, where $p_0(\theta)$ is a prior distribution.

- Without regularization, the problem reduces to standard frequentist learning.
- The regularization term provides a bound on the **generalization error** via PAC Bayes theory, and it underlies the free energy principle in neuroscience.[2]

[2] S. T. Jose and O. Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization," IEEE Signal Processing Magazine, 2021.

Frequentist Learning vs. Bayesian Learning

- Given a training set $\mathcal{D}$, conventional frequentist learning minimizes the training loss $L_{\mathcal{D}}(\theta)$ over θ.
- Bayesian learning minimizes the variational **free energy**

$$F_{\mathcal{D}}(q(\theta)) = \underbrace{\mathrm{E}_{\theta\sim q(\theta)}[L_{\mathcal{D}}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}},$$

over distribution $q(\theta)$, where $p_0(\theta)$ is a prior distribution.

- Without regularization, the problem reduces to standard frequentist learning.
- The regularization term provides a bound on the **generalization error** via PAC Bayes theory, and it underlies the free energy principle in neuroscience.[2]

[2] S. T. Jose and O. Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization," IEEE Signal Processing Magazine, 2021.

Frequentist Learning vs. Bayesian Learning

- Given a training set $\mathcal{D}$, conventional frequentist learning minimizes the training loss $L_{\mathcal{D}}(\theta)$ over θ.
- Bayesian learning minimizes the variational **free energy**

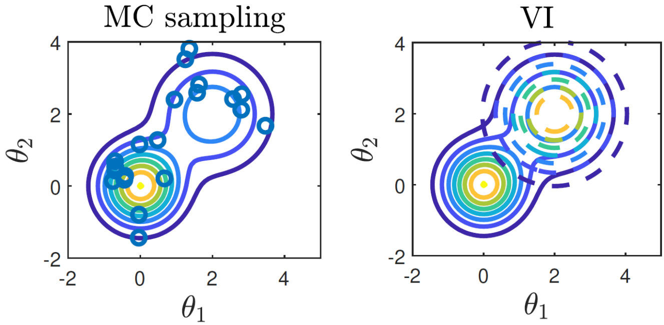
$$F_{\mathcal{D}}(q(\theta)) = \underbrace{\mathrm{E}_{\theta\sim q(\theta)}[L_{\mathcal{D}}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}},$$

over distribution $q(\theta)$, where $p_0(\theta)$ is a prior distribution.
- Without regularization, the problem reduces to standard frequentist learning.
- The regularization term provides a bound on the **generalization error** via PAC Bayes theory, and it underlies the free energy principle in neuroscience.[2]

[2] S. T. Jose and O. Simeone, "Free energy minimization: A unified framework for modeling, inference, learning, and optimization," IEEE Signal Processing Magazine, 2021.

Approximate Bayesian Learning

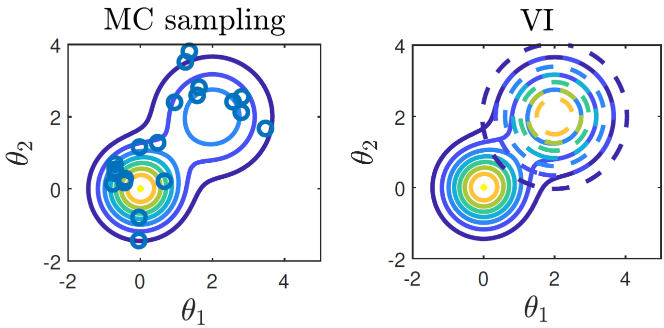
- Exact Bayesian learning is generally intractable (it requires computing the posterior distribution over θ).
- Approximate solutions can be obtained via variational inference (VI) or Monte Carlo (MC) sampling[3].



[3] O. Simeone, *Machine learning for Engineers*, Cambridge University Press, 2022.

Approximate Bayesian Learning

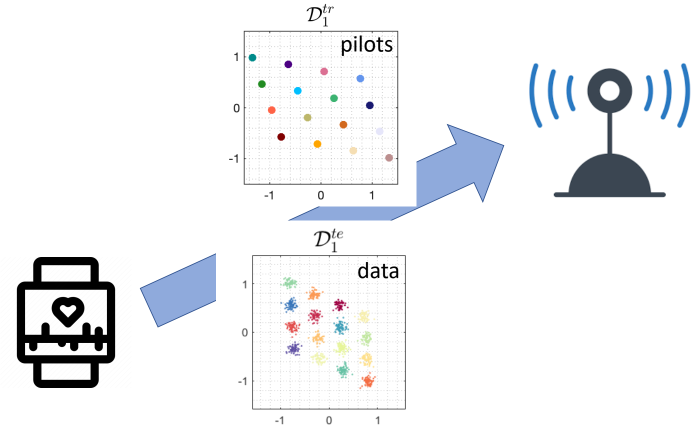
- Exact Bayesian learning is generally intractable (it requires computing the posterior distribution over θ).
- Approximate solutions can be obtained via variational inference (VI) or Monte Carlo (MC) sampling[3].



3 O. Simeone, *Machine learning for Engineers*, Cambridge University Press, 2022.

Application to Demodulation

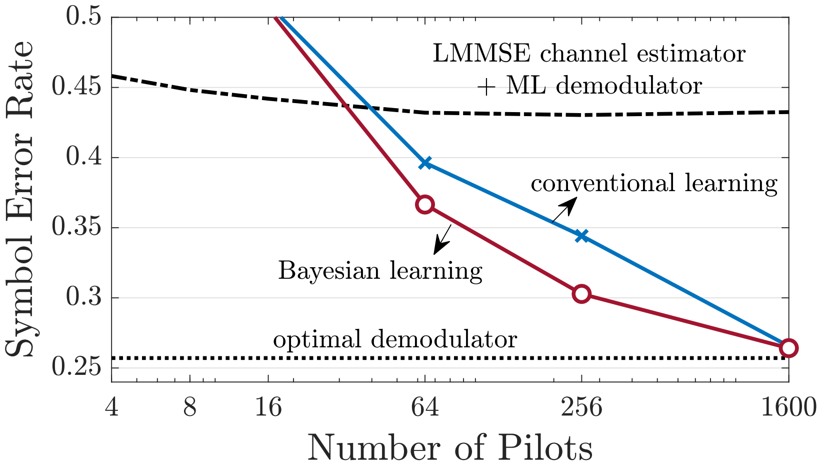
- Short-packet transmission with I/Q imbalance[4]



[4] K. Cohen, S. Park, and O. Simeone, "Learning to Learn to Demodulate with Uncertainty Quantification via Bayesian Meta-Learning," in Proc. WSA, 2021.

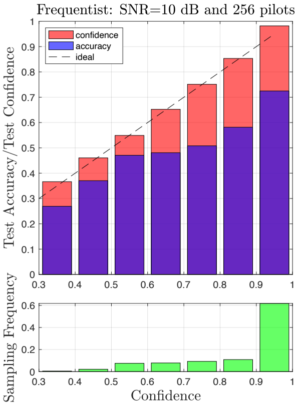
Application to Demodulation

- Frequentist and Bayesian learning yields similar accuracy levels.



Application to Demodulation

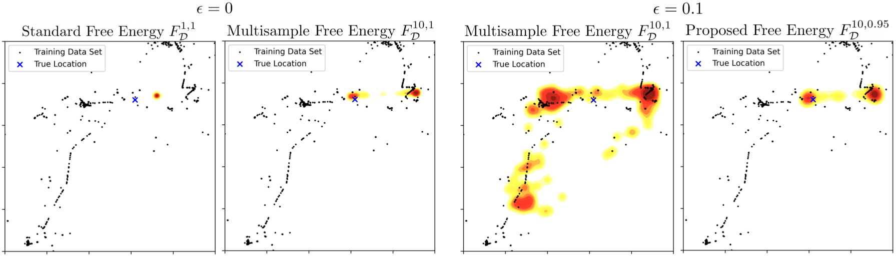
- Reliability plots: accuracy vs. confidence[5]
- Frequentist learning yields **overconfident** decisions, while Bayesian learning produces **well-calibrated** outputs.



5 C. Guo, et al, "On calibration of modern neural networks," ICML 2017.

Extension for Robustness

- Theoretically principled modifications of the free energy to account for model misspecification and outliers[6]

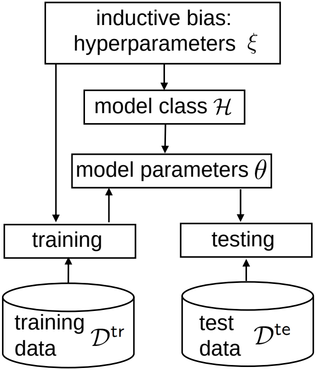


6 M. Zecchin, S. Park, O. Simeone, M. Kountouris, and D. Gesbert, "Robust PAC^m: Training Ensemble Models Under Model Misspecification and Outliers" arXiv:2203.01859, 2022.

Sample-Efficient AI: Meta-Learning

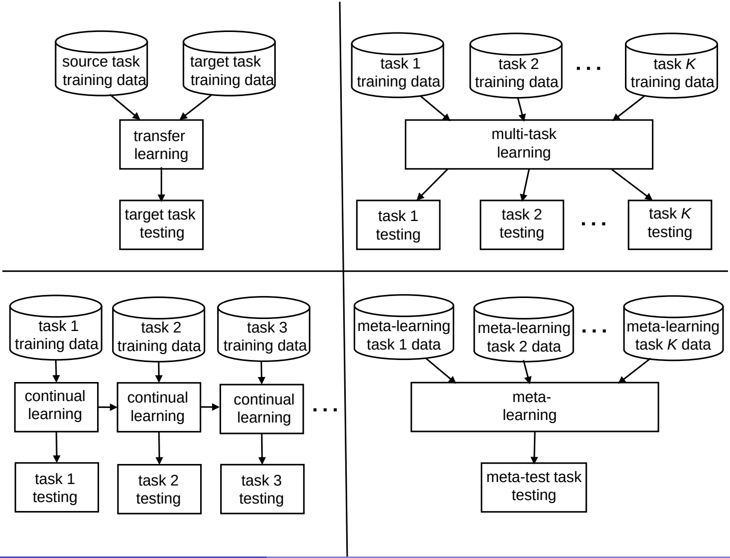
Conventional Learning

- Conventional machine learning may require excessive training data, particularly in settings with time-varying conditions
- Meta-learning provide tools to reduce sample complexity by transferring knowledge from other learning tasks



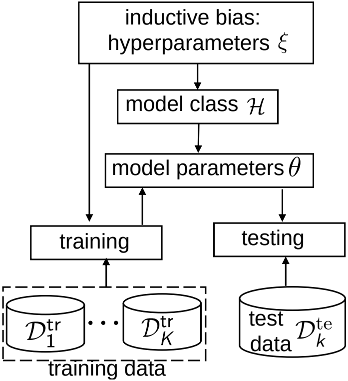
Transferring Knowledge Across Tasks

- There are several ways to formalize the problem of knowledge transfer across tasks.



Joint Learning

- For reference, let us first consider **joint learning** as a simplified form of multi-task learning.
- Joint learning trains a shared model across K tasks, and tests the model on any one of the K tasks.



Joint Learning

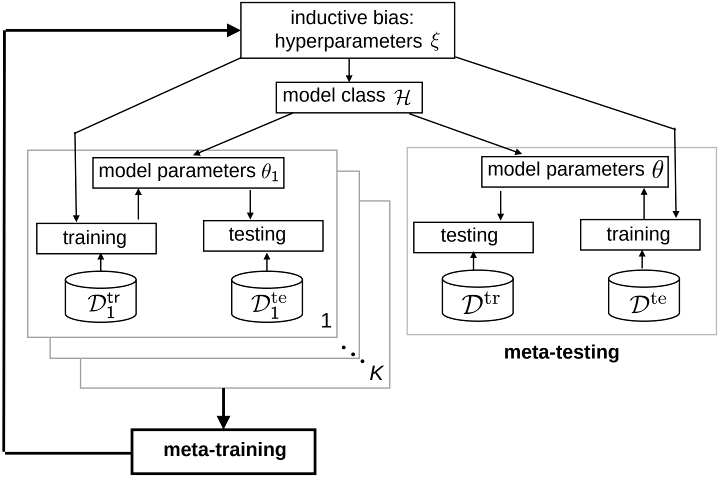
- Joint learning can effectively increase the amount of data by pooling data sets from multiple tasks.
- Joint learning has two potentially critical shortcomings:
  - The jointly trained model only works if there is a single model parameter θ that "works well" for **all** tasks.
  - There is no guarantee that the jointly trained model would be able to **adapt** (even with **fine-tuning**) to a **new** task.

Joint Learning

- Joint learning can effectively increase the amount of data by pooling data sets from multiple tasks.
- Joint learning has two potentially critical shortcomings:
  - The jointly trained model only works if there is a single model parameter θ that "works well" for **all** tasks.
  - There is no guarantee that the jointly trained model would be able to **adapt** (even with **fine-tuning**) to a **new** task.

Meta-Learning

- Meta-learning optimizes **shared hyperparameters**, while enabling **adaptation of the model parameters** for each task (**"learning to learn"**).



Meta-Learning

- Fix a given training algorithm $\theta^{\mathrm{tr}}(\mathcal{D}_k^{\mathrm{tr}}|\xi)$ dependent on hyperparameters ξ (e.g., initialization).
- Meta-learning addresses the aggregate training loss

$$\mathcal{L}_{\{\mathcal{D}_k\}_{k=1}^K}(\xi) = \frac{1}{K}\sum_{k=1}^{K} L_{\mathcal{D}_k^{\mathrm{te}}}(\theta^{\mathrm{tr}}(\mathcal{D}_k^{\mathrm{tr}}|\xi)).$$

- The resulting minimization problem
  - only assumes **common hyperparameters** ξ;
  - inherently prepares the training algorithm $\theta^{\mathrm{tr}}(\mathcal{D}_k^{\mathrm{tr}}|\xi)$ to **adapt** to new tasks.

Meta-Learning

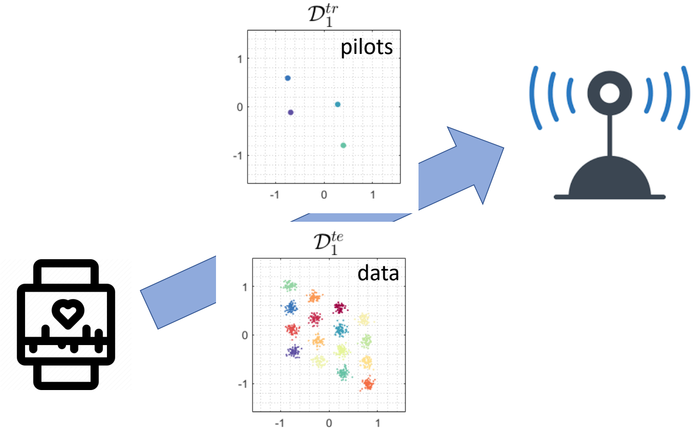
- Fix a given training algorithm $\theta^{\rm tr}(\mathcal{D}_k^{\sf tr}|\xi)$ dependent on hyperparameters ξ (e.g., initialization).
- Meta-learning addresses the aggregate training loss

$$\mathcal{L}_{\{\mathcal{D}_k\}_{k=1}^K}(\xi) = \frac{1}{K}\sum_{k=1}^{K} L_{\mathcal{D}_k^{\sf te}}(\theta^{\rm tr}(\mathcal{D}_k^{\sf tr}|\xi)).$$

- The resulting minimization problem
  - only assumes **common hyperparameters** ξ;
  - inherently prepares the training algorithm $\theta^{\rm tr}(\mathcal{D}_k^{\sf tr}|\xi)$ to **adapt** to new tasks.

Application to Demodulation

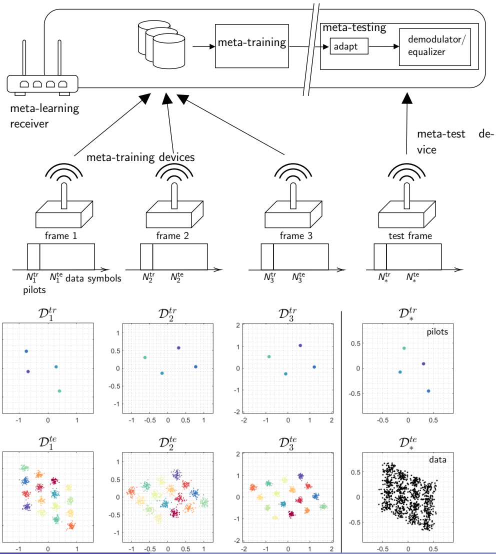
- Short-packet transmission with I/Q imbalance[7]



[7] S. Park, H. Hang, and O. Simeone, "Learning to demodulate from few pilots via offline and online meta-learning," IEEE Transactions on Signal Processing, 2020.

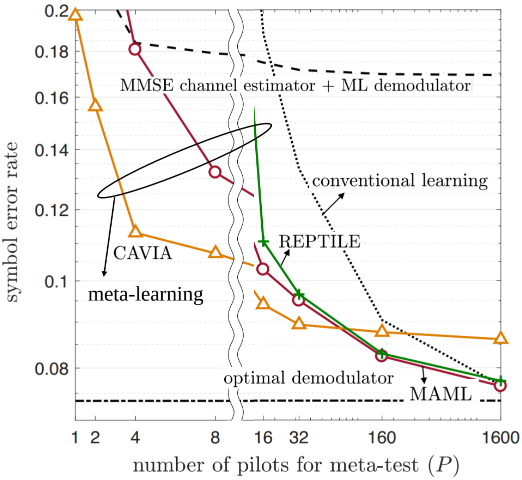
Application to Demodulation

- Meta-learning uses pilots received in previous frames from other devices.



Application to Demodulation

- Meta-learning allows for a much faster adaptation than joint and conventional learning.



Other Applications of Meta-Learning

- Channel prediction[8,9]
- Equalization of multi-path channels[10]
- End-to-end design of encoder and decoder[11]
- Channel acquisition and precoding in FDD massive MIMO[12]
- Power control in time-varying topologies (via graph neural networks)[13]
- Radar processing[14]

---

8 A. Kalor, O. Simeone, and P. Popovski, "Prediction of mmWave/THz Blockages through Meta-Learning and Recurrent Neural Networks," IEEE Comm. Letters, 2022.

9 S. Park and O. Simeone, "Predicting Flat-Fading Channels via Meta-Learned Closed-Form Linear Filters and Equilibrium Propagation", arXiv:2110.00414, 2021.

10 T. Raviv, S. Park, N. Shlezinger, O. Simeone, Y. Eldar, and J. Kang, "Meta-ViterbiNet: Online Meta-Learned Viterbi Equalization for Non-Stationary Channels," in Proc. ICC, 2021.

11 S. Park, O. Simeone, and J. Kang, "End-to-End Fast Training of Communication Links Without a Channel Model via Online Meta-Learning," in Proc. SPAWC, 2020.

12 Y. Liu and O. Simeone, "HyperRNN: Deep learning-aided downlink CSI acquisition via partial channel reciprocity in FDD massive MIMO," in Proc. IEEE SPAWC, 2021.

13 I. Nikoloska and O. Simeone, "Fast power control adaptation via meta-learning for random edge graph neural networks," in Proc. IEEE SPAWC, 2021.

14 W. Jiang, A. Haimovich, M. Govoni, T. Garner, and O. Simeone, "Fast Data-Driven Adaptation of Radard Detection via Meta-Learning," in Proc. Asilomar, 2021.

Reliable and Sample-Efficient AI: Bayesian Meta-Learning

Integrating Bayesian Learning and Meta-Learning

- Recall that, given a prior $p_0(\theta)$, for each learning task k, Bayesian learning aims at minimizing the free energy

$$F_{\mathcal{D}_k}(q(\theta)) = \underbrace{\mathrm{E}_{\theta \sim q(\theta)}[L_{\mathcal{D}_k}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}}$$

- With variational inference, minimization is done over the parameters φ of a variational distribution $q(\theta|\varphi)$ (e.g., Gaussian).
- Hyperparameters ξ may determine
  - the prior $p_0(\theta|\xi)$ (empirical Bayes)
  - the optimizer over φ (e.g., initialization)

Integrating Bayesian Learning and Meta-Learning

- Recall that, given a prior $p_0(\theta)$, for each learning task k, Bayesian learning aims at minimizing the free energy

$$F_{\mathcal{D}_k}(q(\theta)) = \underbrace{\mathrm{E}_{\theta\sim q(\theta)}[L_{\mathcal{D}_k}(\theta)]}_{\text{average training loss}} + \underbrace{\mathrm{KL}\left(q(\theta)||p_0(\theta)\right)}_{\text{information-theoretic regularization}}$$

- With variational inference, minimization is done over the parameters φ of a variational distribution $q(\theta|\varphi)$ (e.g., Gaussian).
- Hyperparameters ξ may determine
  - the prior $p_0(\theta|\xi)$ (empirical Bayes)
  - the optimizer over φ (e.g., initialization)

Integrating Bayesian Learning and Meta-Learning

- Accordingly, we obtain a (variational) posterior distribution $q^{\mathrm{tr}}(\theta|\mathcal{D}_k^{\mathrm{tr}}, \xi)$ for task k given data $\mathcal{D}_k^{\mathrm{tr}}$ and hyperparameter vector ξ.
- Given data from K tasks, meta-learning can be defined as the minimization of the aggregate average training loss

$$\mathcal{F}_{\{\mathcal{D}_k\}_{k=1}^K}(\xi) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}_{\theta\sim q^{\mathrm{tr}}(\theta|\mathcal{D}_k^{\mathrm{tr}},\xi)}[L_{\mathcal{D}_k^{\mathrm{te}}}(\theta)].$$

- This criterion can be derived (and extended) via PAC Bayes theory[15,16,17]

15 S. T. Jose, O. Simeone, and G. Durisi, "Transfer Meta-Learning: Information-Theoretic Bounds and Information Meta-Risk Minimization," IEEE Trans. Inf. Theory, to appear.

16 S. T. Jose and O. Simeone, "An information-theoretic analysis of the impact of task similarity on meta-learning," in Proc. IEEE ISIT 2021.

17 S. Jose, S. Park, and O. Simeone, "Information-Theoretic Analysis of Epistemic Uncertainty in Bayesian Meta-learning," in Proc. AISTATS 2022.

Integrating Bayesian Learning and Meta-Learning

- Accordingly, we obtain a (variational) posterior distribution $q^{\mathrm{tr}}(\theta|\mathcal{D}_k^{\mathsf{tr}}, \xi)$ for task k given data $\mathcal{D}_k^{\mathsf{tr}}$ and hyperparameter vector ξ.
- Given data from K tasks, meta-learning can be defined as the minimization of the aggregate average training loss

$$\mathcal{F}_{\{\mathcal{D}_k\}_{k=1}^K}(\xi) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}_{\theta\sim q^{\mathrm{tr}}(\theta|\mathcal{D}_k^{\mathsf{tr}},\xi)}[L_{\mathcal{D}_k^{\mathsf{te}}}(\theta)].$$

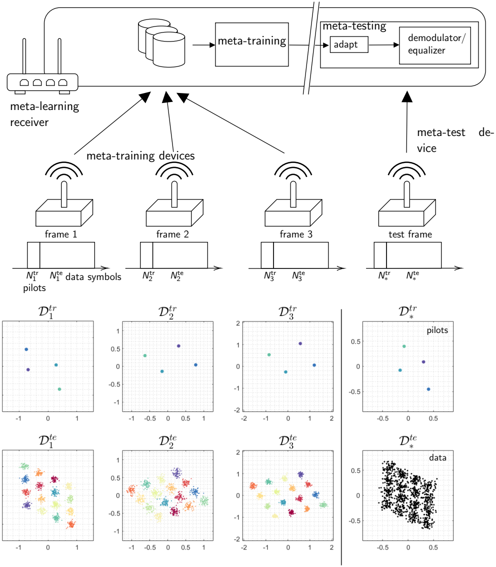
- This criterion can be derived (and extended) via PAC Bayes theory[15,16,17]

15 S. T. Jose, O. Simeone, and G. Durisi, "Transfer Meta-Learning: Information-Theoretic Bounds and Information Meta-Risk Minimization," IEEE Trans. Inf. Theory, to appear.

16 S. T. Jose and O. Simeone, "An information-theoretic analysis of the impact of task similarity on meta-learning," in Proc. IEEE ISIT 2021.

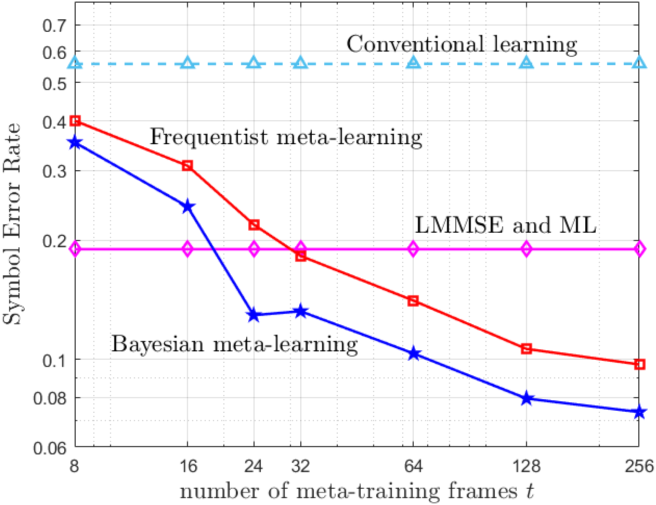
17 S. Jose, S. Park, and O. Simeone, "Information-Theoretic Analysis of Epistemic Uncertainty in Bayesian Meta-learning," in Proc. AISTATS 2022.

Application to Demodulation



Application to Demodulation

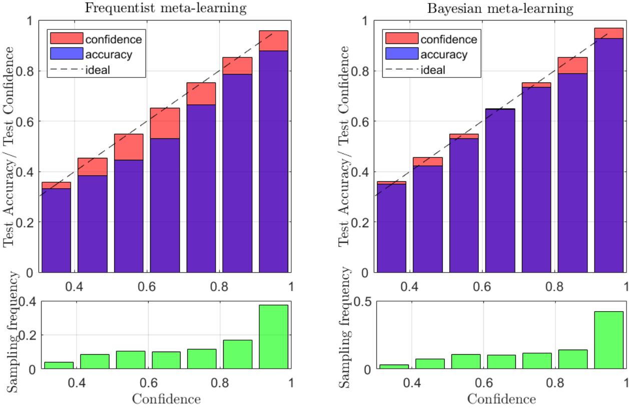
- Symbol error rate vs. SNR (with 8 pilots)[18]



18 K. Cohen, et al, "Learning to Learn to Demodulate with Uncertainty Quantification via Bayesian Meta-Learning," in Proc. WSA, 2021.

Application to Demodulation

- Reliability plots (with 8 pilots)



Other Applications of Bayesian Meta-Learning

- Active Bayesian meta-learning[19]
- Bayesian optimization for black-box optimization[20]

[19] K. Cohen, S. Park, and O. Simeone, "Towards Reliable and Efficient AI for 6G: Bayesian Active Meta-Learning for Few-Pilot Demodulation and Equalization," arXiv:2108.00785, 2022.

[20] I. Nikoloska and O. Simeone, "Bayesian Active Meta-Learning for Black-Box Optimization," submitted.

Conclusions

Conclusions

- Reliable AI via Bayesian learning
- Efficient AI via meta-learning
- Reliable and efficient AI via Bayesian meta-learning
- Directions for future research:
  - Robustness to model misspecification and outliers
  - Formal reliability guarantees
  - Active learning and meta-learning

Acknowledgements

This work has been supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 725731)