FY2025/2025年度

Adaptive Bayesian Intelligence Team

Mohammad Emtiyaz Khan
適応ベイズ知能チーム

RIKEN Center for Advanced Intelligence Project

Overview and Goals

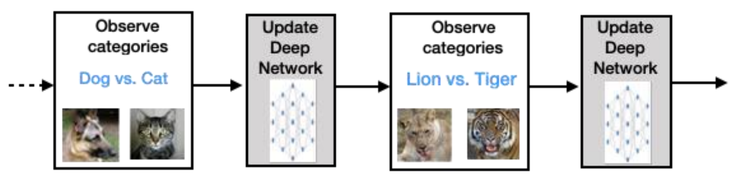
Goal: to design low-cost AI systems that can learn and improve continually throughout their lives, just like humans and animals. Currently, deep learning requires a large amount of data which is costly, rigid, and cannot quickly adapt. We aim to fix this with a new principle called Bayes-duality.

Summary for the year 2025 (paper number shown in green boxes)

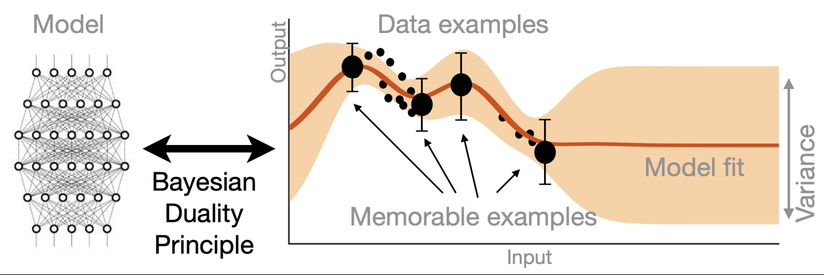
1. We provide new connections between Federated ADMM and Variational Bayes and propose new algorithms using this connections.
2. We show, both in theory and practice, that variational Bayesian learning finds flatter solutions than standard training algorithms for neural nets.
3. We propose a compact memory method to drastically reduce the storage for continual learning without forgetting.
4. We propose a method that allows LLMs to leverage parameter uncertainty, and show it improves the quality of generated text.
5. We highlight a fundamental connection between information geometry and variational Bayes and discuss its consequences.

Standard Deep Learning

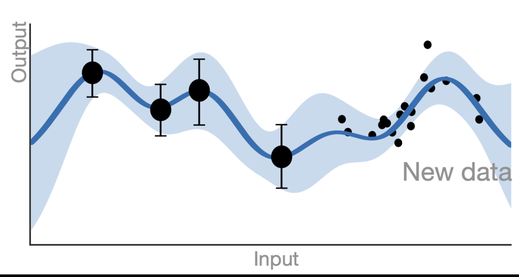
Continual Lifelong Learning



Bayes-duality relates *models parameters (left)* to the *data examples* (small dots at the right). The principle enables us to identify a few *memorable examples* (big black circle).



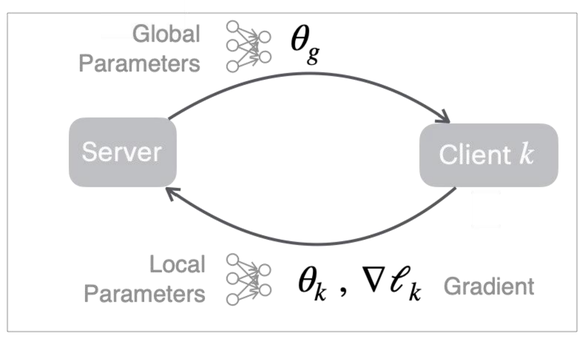
The memorable examples can be reused later during training with the new data. This avoids forgetting of the past.



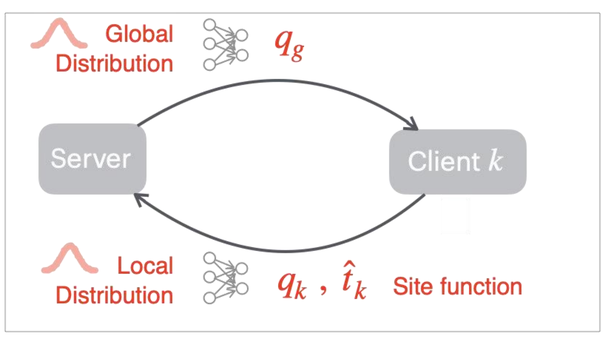
Connecting Federated ADMM to Bayes

Problem & Contribution [1]: We provide new connections between Federated ADMM and Variational Bayes, showing that the dual variables in ADMM naturally emerge through the "site" parameters in VB for isotropic Gaussian posteriors. Based on this, we derive two new extensions of ADMM by using full-covariance Gaussians.

ADMM



Partition Variational Inference



A line-by-line correspondence between Federated ADMM and Partition Variational Inference

$$\boldsymbol{w}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{w}} \ \bar{\ell}_k(\boldsymbol{w}) + \boldsymbol{v}_k^\top \boldsymbol{w} + \tfrac{1}{2}\alpha\|\boldsymbol{w} - \boldsymbol{w}_g\|^2, \qquad q_k \leftarrow \arg\min_{q \in \mathcal{Q}} \ \mathbb{E}_q[\ell_k(\boldsymbol{w})] + \mathbb{E}_q[\log \hat{t}_k] + \mathbb{D}_{KL}[q \,\|\, q_g]$$

$$\boldsymbol{v}_k \leftarrow \boldsymbol{v}_k + \alpha(\boldsymbol{w}_k - \boldsymbol{w}_g), \quad \text{for all } k, \qquad \log \hat{t}_k \leftarrow \log \hat{t}_k + \rho\,(\log q_k - \log q_g)$$

$$\boldsymbol{w}_g \leftarrow \frac{1}{K}\sum_{k=1}^{K}\left[\boldsymbol{w}_k + \frac{1}{\alpha}\boldsymbol{v}_k\right]. \qquad \log q_g \leftarrow \log p_0 + \sum_{k=1}^{K} \log \hat{t}_k + \text{const},$$

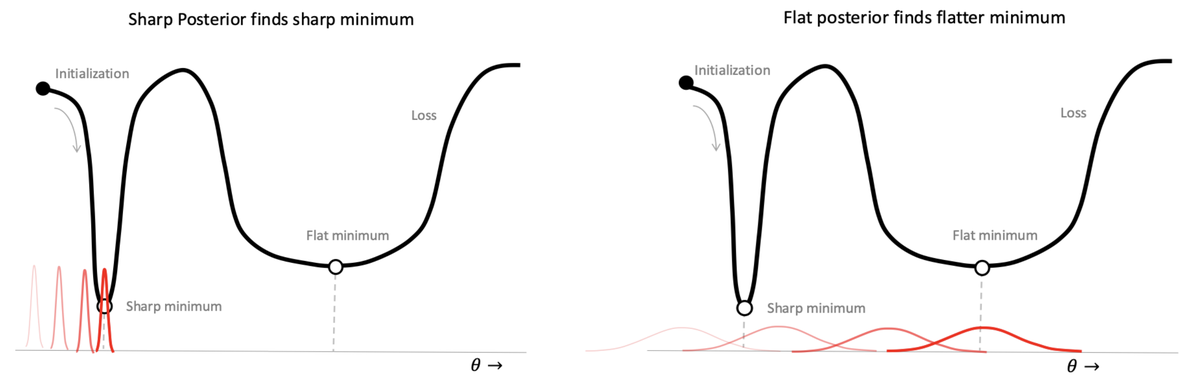
New ADMM extensions

Dataset	Comm Round	FedAvg	FedProx	FedDyn	FedLap	FedLap -Cov	FedLap -Func
FMNIST (homog)	10	72.3(0.4)	72.2(0.3)	**75.3(0.8)**	72.1(0.2)	**75.0(0.6)** (↓0.3)	73.7(0.7)
	25	77.7(0.3)	77.4(0.1)	77.5(0.8)	77.1(0.1)	**79.8(0.4)** (↑2.3)	**77.9(0.3)**
	50	80.0(0.2)	**80.3(0.1)**	78.2(0.5)	80.2(0.1)	**81.8(0.1)** (↑2.6)	80.0(0.2)
FMNIST (heterog)	10	70.4(0.9)	69.9(0.4)	**73.0(0.6)**	71.3(0.9)	**74.6(0.7)** (↑1.6)	72.2(0.9)
	25	74.3(0.5)	74.7(0.6)	74.6(0.4)	74.3(0.4)	**78.3(1.0)** (↑3.7)	**75.4(0.8)**
	50	76.0(0.7)	76.9(0.9)	74.6(0.5)	77.6(0.7)	**80.5(0.6)** (↑5.9)	**78.1(0.7)**

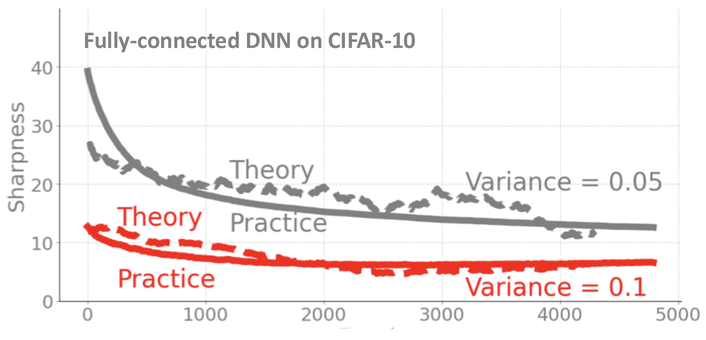
1. S. Swaroop, M. E. Khan, F. Doshi-Velez, Connecting Federated ADMM to Bayes, *ICLR 2025*.

Variational Learning at the Edge of Stability

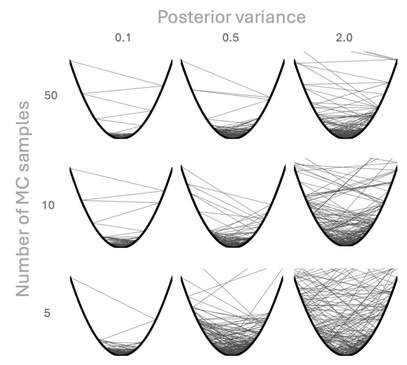
Problem & Contribution [2]: Variational learning has recently been shown to work well for deep neural networks, but the reasons for success are not fully understood. We provide a theory explaining why variational learning works well for nonconvex problems. In the presence of multiple minima, it prefers flatter minima. We characterize flatness of found minima via learning rate, posterior shape & samples.



Our theory closely predicts the sharpness of solutions found during deep neural network training



Intuition: Posterior noise helps to escape from sharp minima

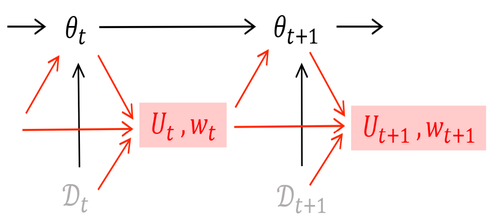


2. A. Ghosh, B. Cong, R. Yokota, S. Ravishankar, R. Wang, M. Tao, M. E. Khan, T. Möllenhoff. Variational Learning Finds Flatter Solutions at the Edge of Stability, NeurIPS 2025 (spotlight).

Compact Memory for Continual Logistic Regression

Problem & Contribution [3]: Continual learning methods need to store a large number of past example to avoid forgetting. Here, we drastically reduce the size of the memory without any significant change in the performance.

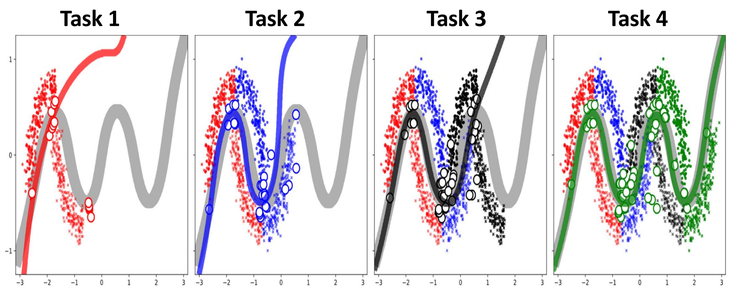
We use Hessian matching to find memory and weight in two stage procedure:



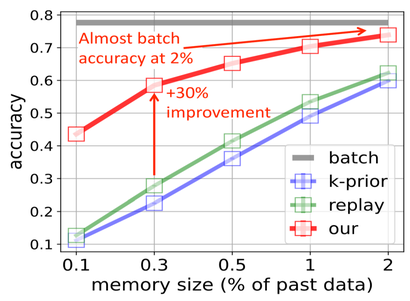
Step 1:
Learn parameter θ_{t+1} with given $(\mathrm{U_t}, \mathrm{w}_t)$

Step 2:
Learn memory $(\mathrm{U_{t+1}}, \mathrm{w}_{t+1})$ by Hessian matching

Our method achieves near-batch performance with significantly less memory.



Continual learning on Four moon dataset



Split-ImageNet-1000

3. Y. Jung, H. Lee, W. Chen, T. Möllenhoff, Y. Li, J. Lee, M. E. Khan, Compact Memory for Continual Logistic Regression, *NeurIPS 2025*.

Other Works

Uncertainty-Aware Decoding: LLMs are known to generate undesirable outputs, for example, hallucinated text. In this work, we use parameter uncertainty to improve text generation. By using our previously developed variational training methods which obtain parameter uncertainty for free, we improve LLMs on several tasks such as translation and document summarization without incurring any overhead.

4. N. Daheim, C. Meister, T. Möllenhoff, I. Gurevych. Uncertainty-Aware Decoding with Minimum Bayes Risk. ICLR 2025.

Information geometry of VB: We highlight a fundamental connection between information geometry and variational Bayes and discuss its consequences.

5. Khan. Information Geometry of Variational Bayes, *Information geometry, Springer Nature 2025*

Theoretical Guarantees for Natural-Gradient Methods: Natural gradient methods are among the best performing approaches for variational inference, but their convergence properties are not understood. This work proves novel guarantees in the Gaussian case by considering a square-root parametrization of the covariance.

6. N. Kumar, T. Möllenhoff, M. E. Khan, A. Lucchi. Optimization Guarantees for Square-Root Natural-Gradient Variational Inference. TMLR 2025.

Label Smoothing: We show that variational learning naturally induces an *adaptive* label smoothing where label noise is specialized for each example. We show that the form of the adaptive noise is similar to an existing proposal by Zhang et al. (2021).

7. S. H. Yang, Z. Liu, G. M. Marconi, M. E. Khan, Variational Learning Induces Adaptive Label Smoothing. *AABI 2025*