Presentation at University of Bath, Nov 24, 2022

SAM as an optimal approximation of Bayes

Mohammad Emtiyaz Khan
RIKEN Center for AI Project, Tokyo
http://emtiyaz.github.io

Summary of research at https://emtiyaz.github.io/papers/symposium_2021.pdf
Slides available at https://emtiyaz.github.io/papers/Nov24_2022_Ubath.pdf

AI that learn like humans

Quickly adapt to learn new skills, throughout their lives

Human Learning at the age of 6 months.

Converged at the age of 12 months

Transfer skills

at the age of 14 months

Fail because too slow to adapt

https://www.youtube.com/watch?v=TxobtWAFh8o The video is from 2017

AI that learn like humans

Quickly adapt to learn new skills, throughout their lives

Thomas Moellenhoff

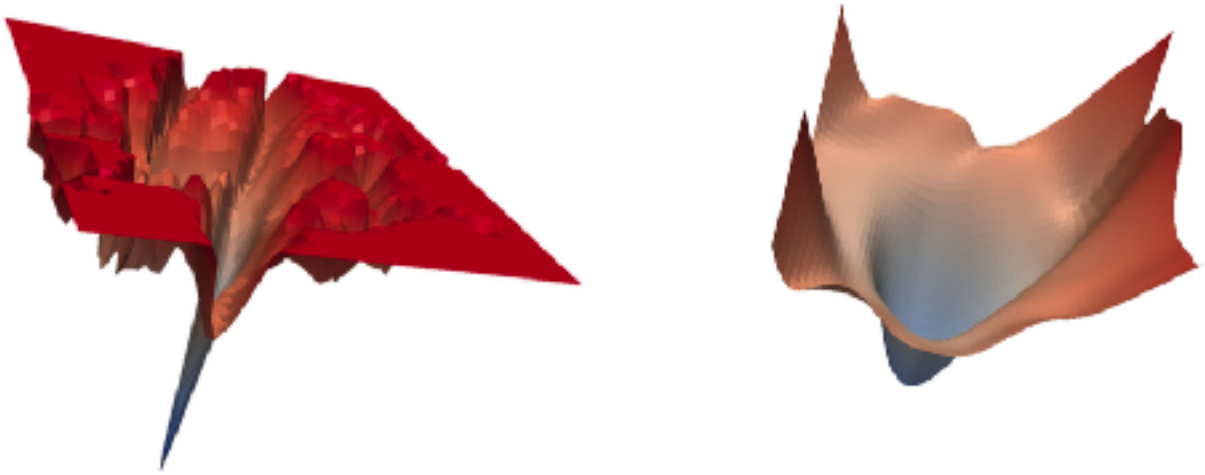
Robust Deep-Learning

- Sharpness-Aware Minimization (SAM)[1]
  - Huge improvements over SGD/Adam
  - Now used to train all sorts of models
  - Improves test accuracy for trained neural networks (ResNets [1,2,3], Vision-Transformers [4], Language Models [5], …)
  - Also improves robustness [3], calibration [2], interpretability [4], transfer-learning [4], compressibility [5], federated learning [6]
- Why does it work, and how to improve it?
- SAM as an “optimal” relaxation of Bayes [7]

1. Foret et al. Sharpness-aware minimization for efficiently improving generalization. *ICLR 2021.*
2. Wu et al. Adversarial Weight Perturbation Helps Robust Generalization . *NeurIPS 2020.*
3. Zheng et al. Regularizing Neural Networks via Adversarial Model Perturbation . *CVPR 2021.*
4. Chen et al. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. *ICLR 2022.*
5. Na et al. Train Flat, Then Compress: SAM Learns More Compressible Models. *EMNLP 2022.*
6. Qu et al. Generalized Federated Learning via Sharpness Aware Minimization. *ICML 2022.*
7. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, arXiv, 2022.

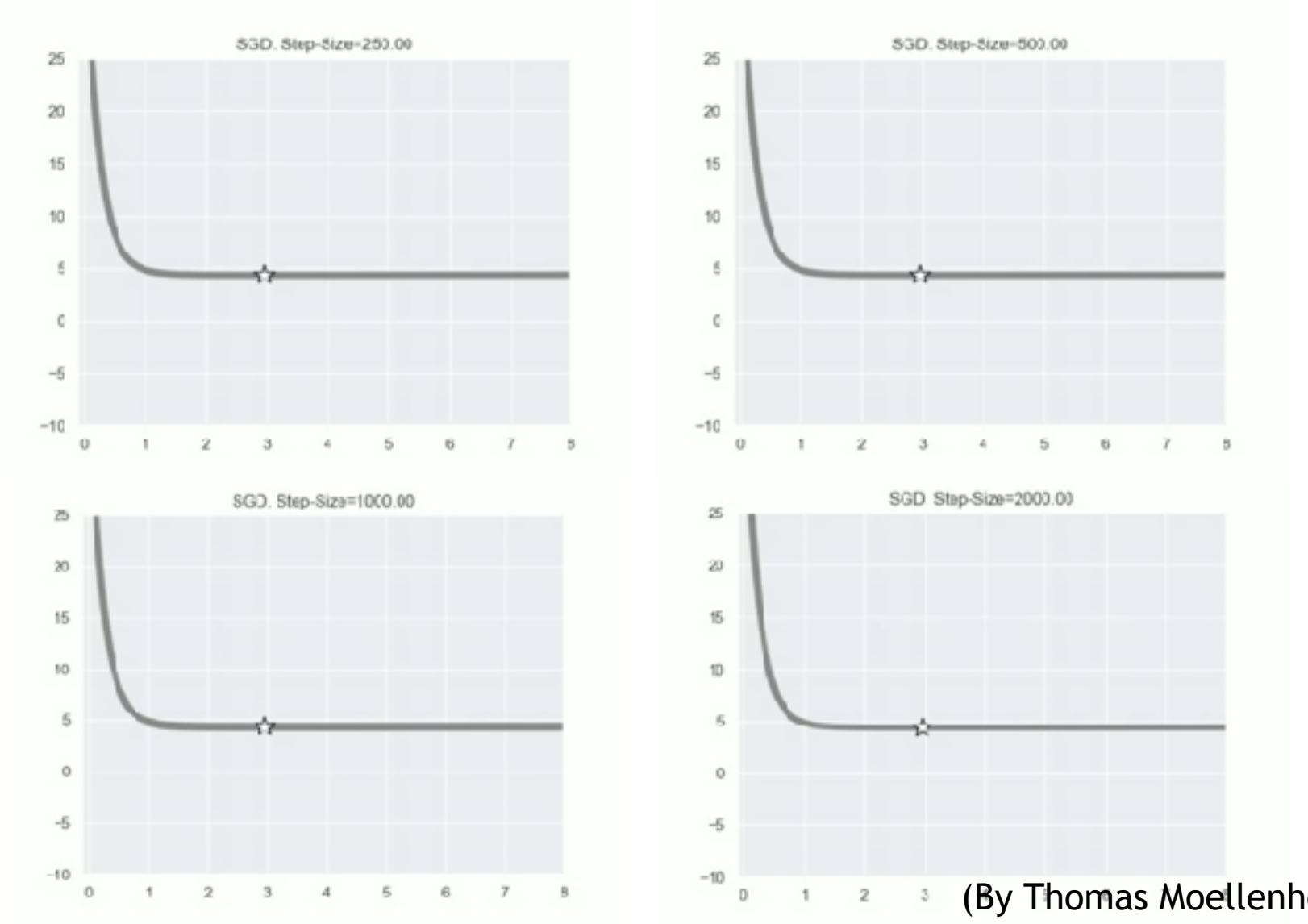
Flat Minima in DL

- Empirical [1] and theoretical [2] evidence suggest that finding “flat minima” is desirable
  - Several ways to bias the learning towards flatter minima, e.g., in SGD via batch-size or learning rate



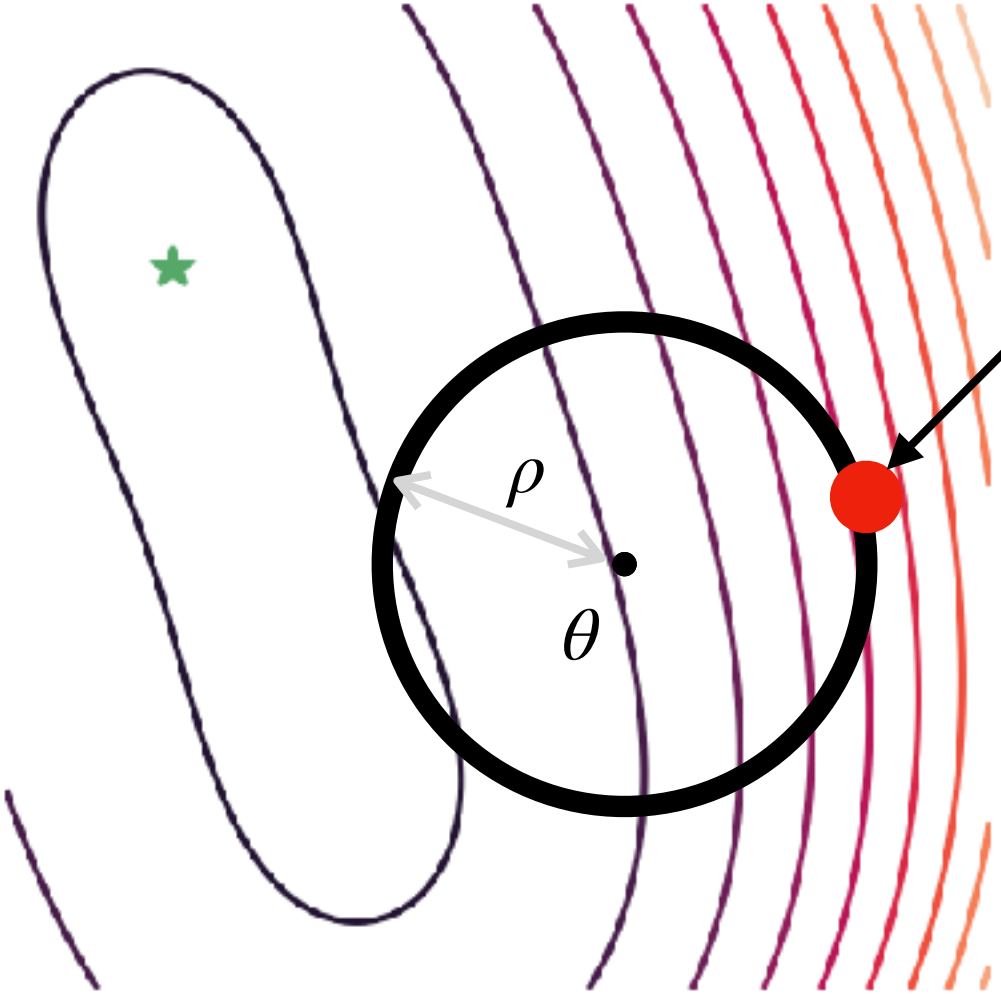
1. Jiang et al., Fantastic generalization measures and where to find them, *ICLR 2020.*
2. Dzuigate and Roy, Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, *UAI 2017.*.

Flat Minima via Learning Rate in SGD



(By Thomas Moellenhoff)

Flat Minima via Adversarial Weight-Perturbation



SAM: $\sup_{|\epsilon|<\rho} \ell(\theta + \epsilon)$

See Section 1.2, Eq 2 in Khan and Rue, 2021

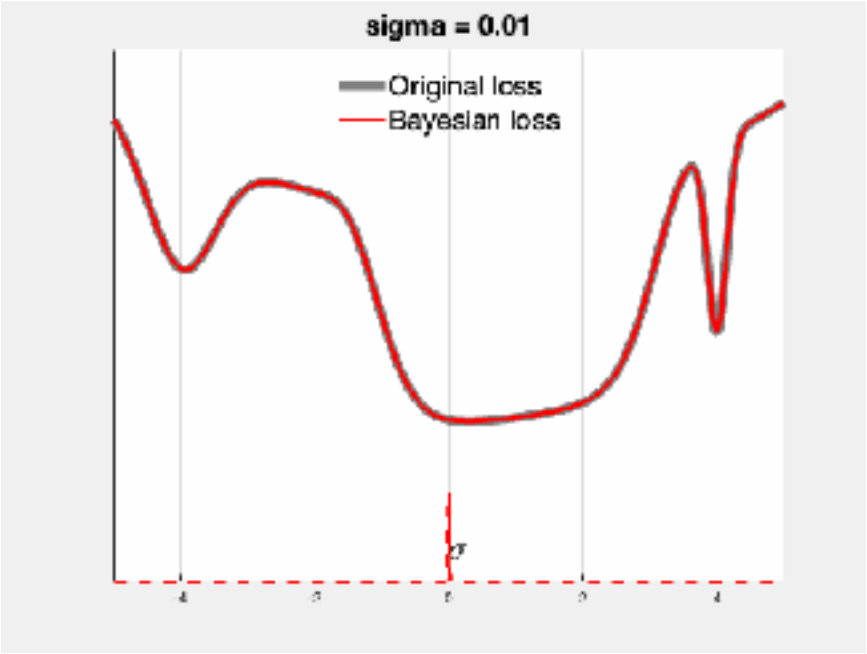
Flat minima via Bayesian Learning

$$\min_{\theta} \ \ell(\theta) \qquad \text{vs} \qquad \min_{q \in \mathcal{Q}} \ \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

Entropy

Gaussian approximation

$$\mathbb{E}_{\mathcal{N}(\theta|m,\sigma^2)}[\ell(\theta)]$$



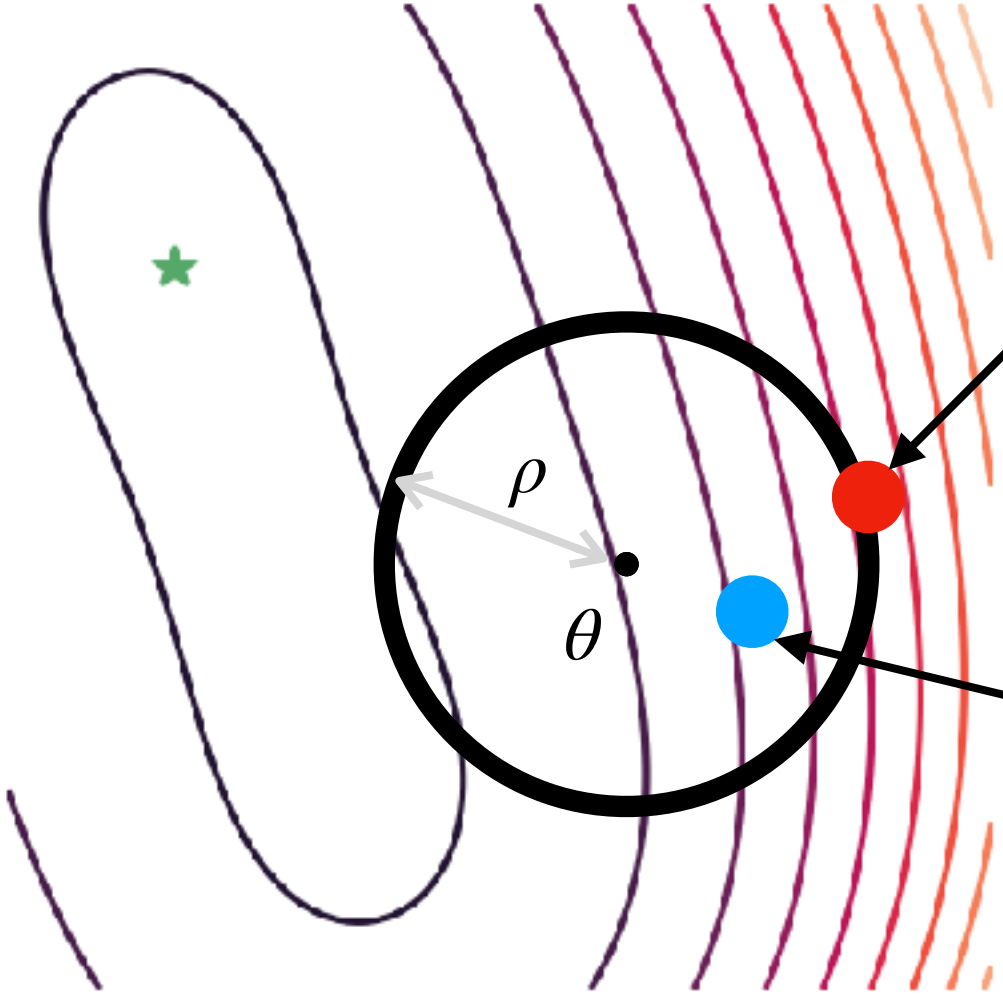
- First term"smooths" the loss by weight perturbation

$$\mathbb{E}_{\mathcal{N}(\epsilon|0,\sigma^2)}[\ell(m+\epsilon)]$$

- Similar mechanism to SAM by it is the "expected-loss" instead of "max-loss"
- An advantage of Bayes is that the second term can be used to estimate σ

1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Smith et al., On the Origin of Implicit Regularization in Stochastic Gradient Descent, ICLR, 2021

SAM as an Optimal relaxation of Bayes



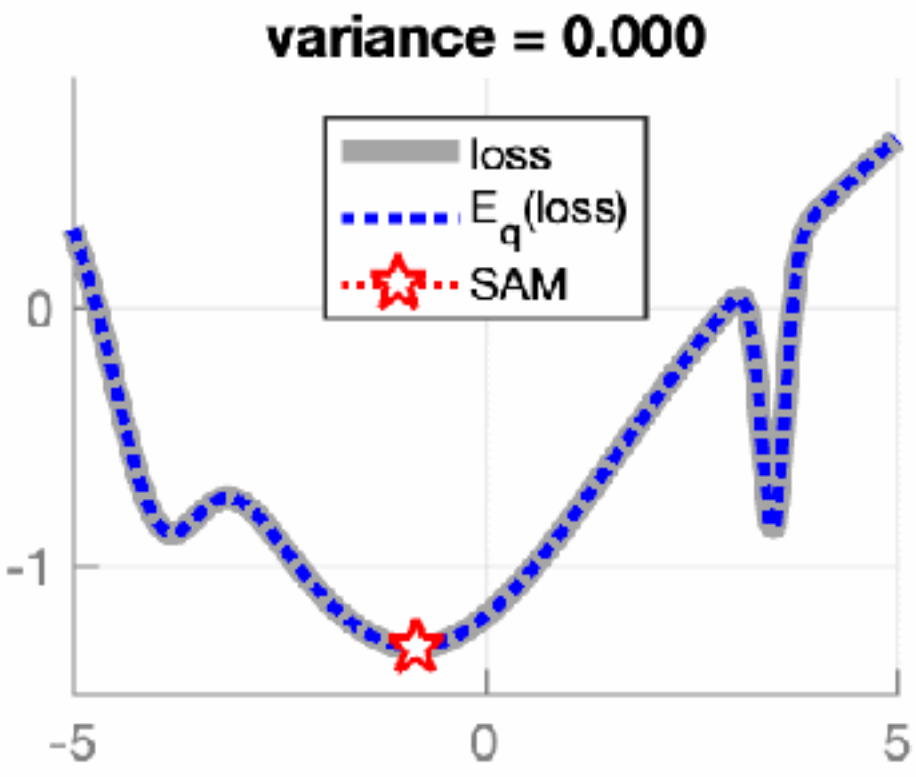
SAM: $\sup_{|\epsilon|<\rho} \ell(\theta+\epsilon)$

Our work: Fenchel Biconjugate

Bayes: $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[\ell(\theta+\epsilon)]$

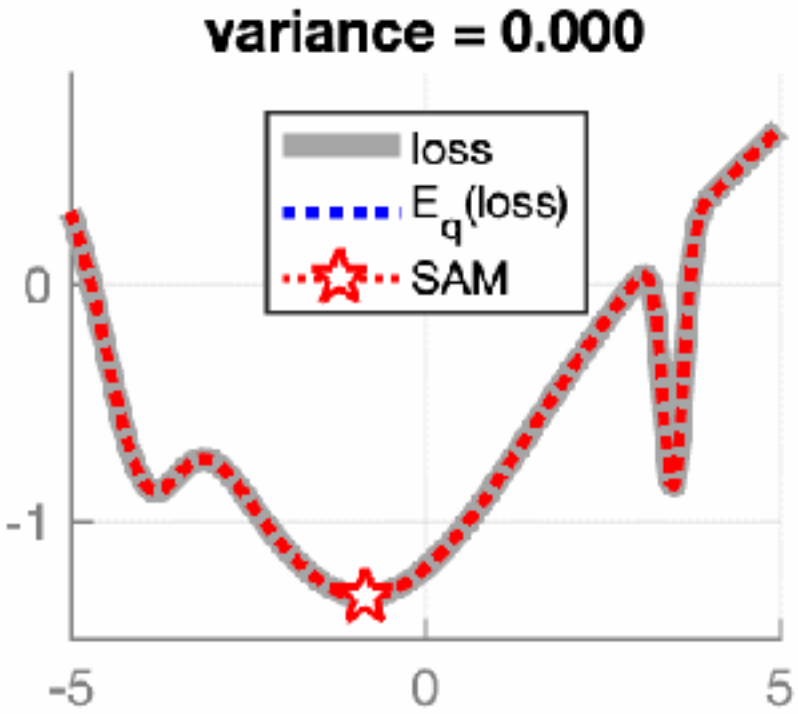
SAM as a relaxation of Bayes

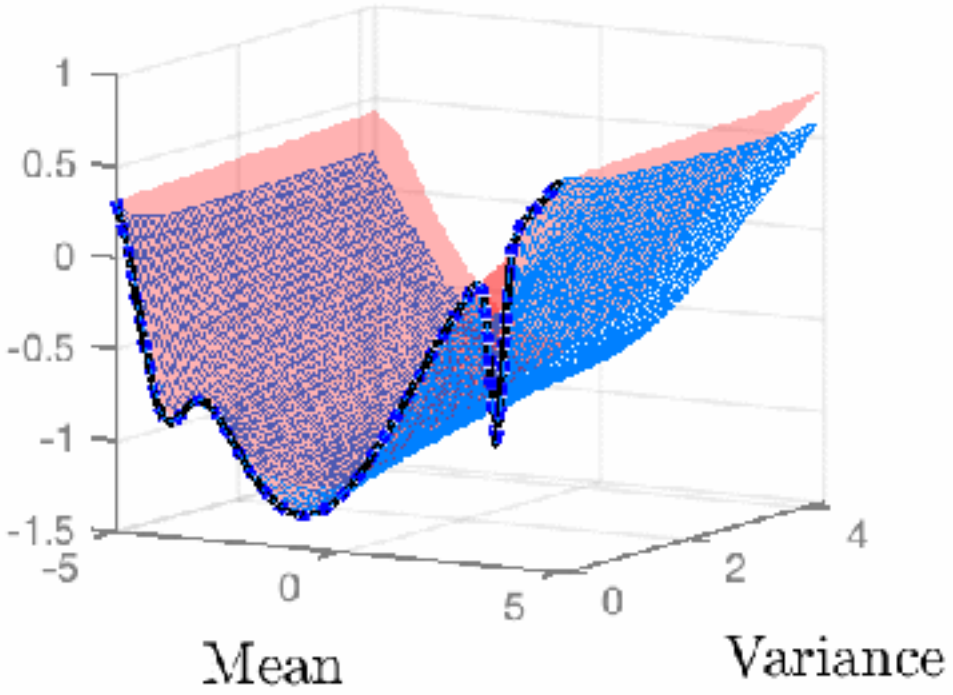
SAM (red star) upper bounds the Bayesian $\mathbb{E}_q[\ell]$



Optimal relaxation: Fenchel Conjugate

SAM minimizes the best-concave upper bound to $\mathbb{E}_q[\ell]$ wrt the mean, while keeping variance fixed.





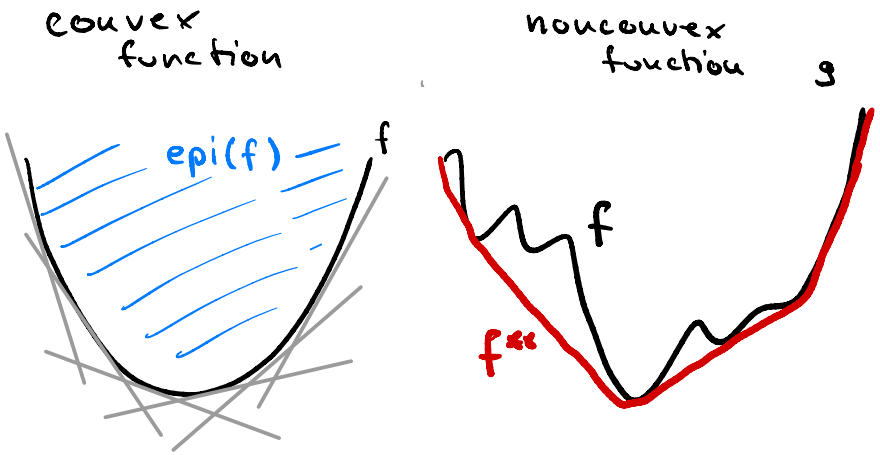
Legendre-Fenchel Transform

Fenchel conjugate

$$f^*(v) = \sup_{u \in U} \langle v, u \rangle - f(u)$$

Fenchel biconjugate

$$f^{**}(u) = \sup_{v \in V} \langle u, v \rangle - f^*(v)$$

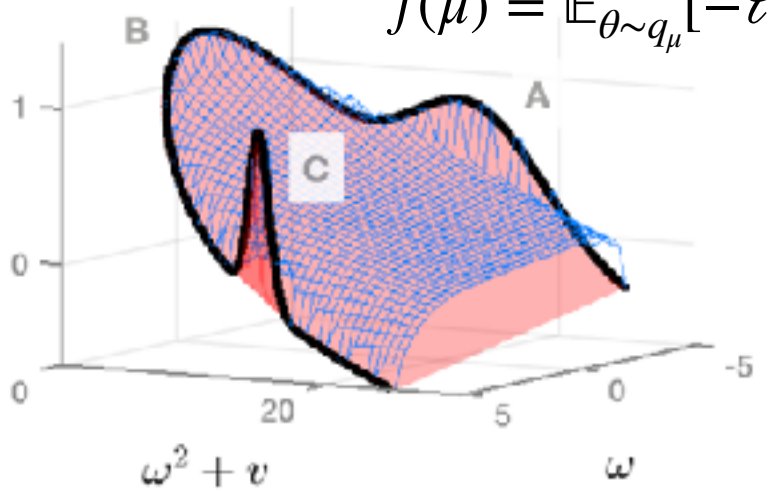


Our bound is obtained in the 2D space of "expectation parameters" of Gaussians win mean ω and variance v

Expectation parameter $\mu = (\omega, \omega^2 + v)$

Mean Variance

$$f(\mu) = \mathbb{E}_{\theta \sim q_\mu}[-\ell(\theta)]$$



The bound in the previous slide is in (ω, v) parameterization

Relaxed Bayesian Objective

- The relaxed objective

$$\mathcal{L}(q_\mu) \geq \mathcal{L}_{\text{relaxed}}(q_\mu) = f^{**}(\boldsymbol{\mu}) - \mathbb{D}_{\text{KL}}\left[q_\mu(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})\right].$$

- This is a difference of convex objective, for which there is a **dual problem** [1]

Natural parameter

$$\max_{\mu} \; f^{**}(\mu) - g(\mu) = \max_{\lambda} \left[-f^{*}(\lambda) + g^{*}(\lambda)\right]$$

- In our case, the dual problem can be rewritten in the following form:

$$\mathcal{E}_{\text{relaxed}}(\mathbf{m}, \sigma; \delta') = \left[\sup_{\boldsymbol{\epsilon}} \ell(\mathbf{m}+\boldsymbol{\epsilon}) - \frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|^2\right] + \underbrace{\frac{\delta'}{2}\|\mathbf{m}\|^2 - \frac{P}{2}\log(\sigma^2\delta')}_{=-\log Z},$$

[1] John F Toland. **Duality in Nonconvex Optimization** . Journal of Mathematical Analysis and Applications, 1978.

Our Result: Relaxed Bayes is SAM!

$$\mathcal{E}_{\text{relaxed}}(\mathbf{m}, \sigma; \delta') = \left[\sup_{\boldsymbol{\epsilon}} \ell(\mathbf{m}+\boldsymbol{\epsilon}) - \frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|^2\right] + \underbrace{\frac{\delta'}{2}\|\mathbf{m}\|^2 - \frac{P}{2}\log(\sigma^2\delta')}_{=-\log Z},$$

$$\mathscr{E}_{SAM}(\theta; \rho, \delta) = \left[\sup_{\|\epsilon\| \le \rho} \ell(\theta+\epsilon)\right] + \frac{\delta}{2}\|\theta\|^2$$

Theorem 2. *For every* (ρ, δ)*, there exist* (σ, δ') *such that*

$$\arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{E}_{SAM}(\boldsymbol{\theta}; \rho, \delta) = \arg\min_{\mathbf{m}\in\Theta} \mathcal{E}_{relaxed}(\mathbf{m}, \sigma; \delta').$$

Essentially, for fixed variance, we can always recover SAM's solution by minimizing the relaxed-bases objective with respect to the mean

Improving SAM

We can obtain the variance by optimizing the relaxed Bayes objective. Below is the Fenchel biconjugate where a covariance Σ is used

$$-f^{**}(\boldsymbol{\mu}) = \min_{\mathbf{m}\in\mathbb{R}^P, \mathbf{b}\in\mathbb{R}^P_+} \left[\sup_{\boldsymbol{\epsilon}\in\mathbb{R}^P} \ell(\mathbf{m}+\boldsymbol{\epsilon}) - \frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\epsilon}\|^2 \right] + \mathbb{E}_{\boldsymbol{\theta}\sim q_\mu}\left[\frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\theta}-\mathbf{m})\|^2\right]$$

Covariance

We use the Bayesian learning rule (BLR) [1] to do that, because it gives an Adam-Style objective

$$\boldsymbol{\lambda} \leftarrow (1-\alpha)\boldsymbol{\lambda} + \alpha\left[\nabla f^{**}(\boldsymbol{\mu}) + \boldsymbol{\lambda}_0\right]$$

Natural parameter — Natural gradient — Prior

[1] M. E. Khan, H. Rue. **The Bayesian Learning Rule** *arXiv:2107.04562, 2021.*

Bayesian-SAM

An Adam-style algorithm, derived using the BLR, where "perturbation-size" is adjusted by using σ^2 (or s)

SAM with RMSprop

$$
\begin{aligned}
g_1 &\leftarrow \hat{\nabla}\ell(\theta) \\
\epsilon &\leftarrow \rho \frac{g_1}{\|g_1\|} \\
g &\leftarrow \hat{\nabla}\ell(\theta+\epsilon) \\
s &\leftarrow (1-\rho)s + \rho g^2 \\
\theta &\leftarrow \theta - \alpha(\sqrt{s}+\delta)^{-1} g
\end{aligned}
$$

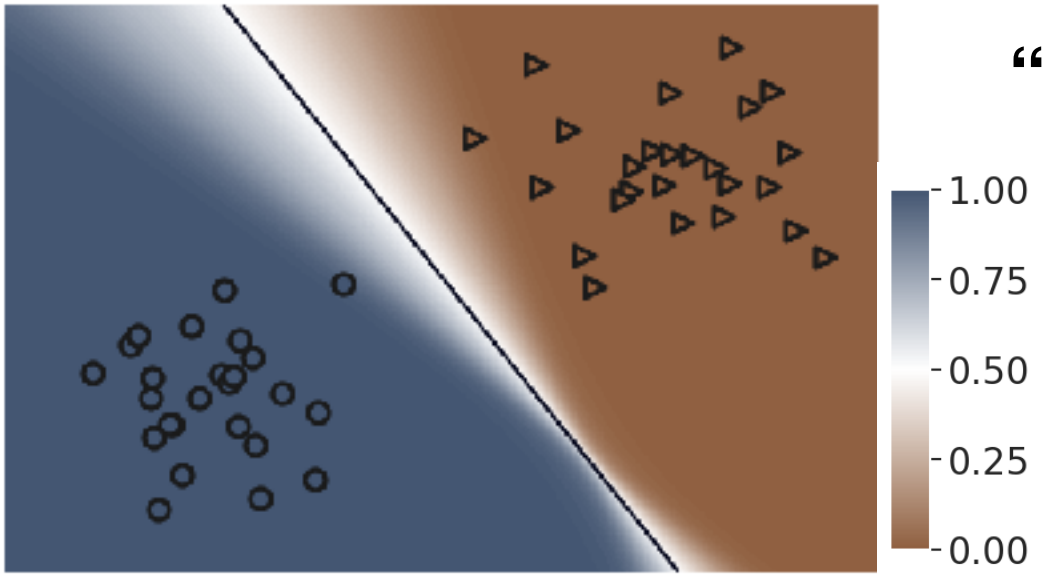
SAM with BLR

$$
\begin{aligned}
g_1 &\leftarrow \hat{\nabla}\ell(\theta), \text{ where } \theta \sim \mathcal{N}(m, \sigma^2) \\
\epsilon &\leftarrow \rho \frac{g_1}{s} \\
g &\leftarrow \hat{\nabla}\ell(\theta+\epsilon) \\
s &\leftarrow (1-\rho)s + \rho \sqrt{s}|g_1| \\
m &\leftarrow m - \alpha(s+\delta)^{-1}\nabla_\theta \ell(\theta) \\
\sigma^2 &\leftarrow (s+\delta)^{-1}
\end{aligned}
$$

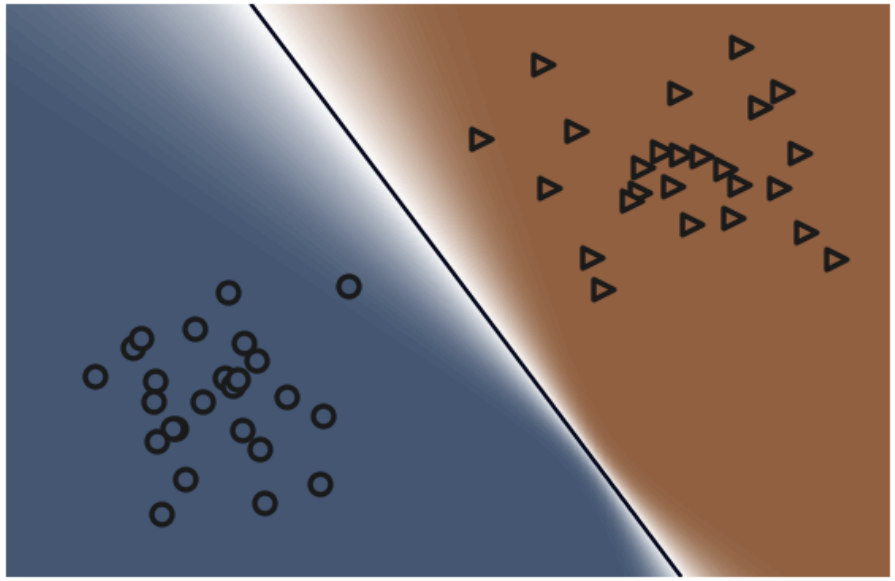
1. Foret et al. Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR, 2021
2. Moellenhoff and Khan, SAM as an optimal relaxation of Bayes, https://arxiv.org/abs/2210.01620, 2022

Improving "overconfident" SAM

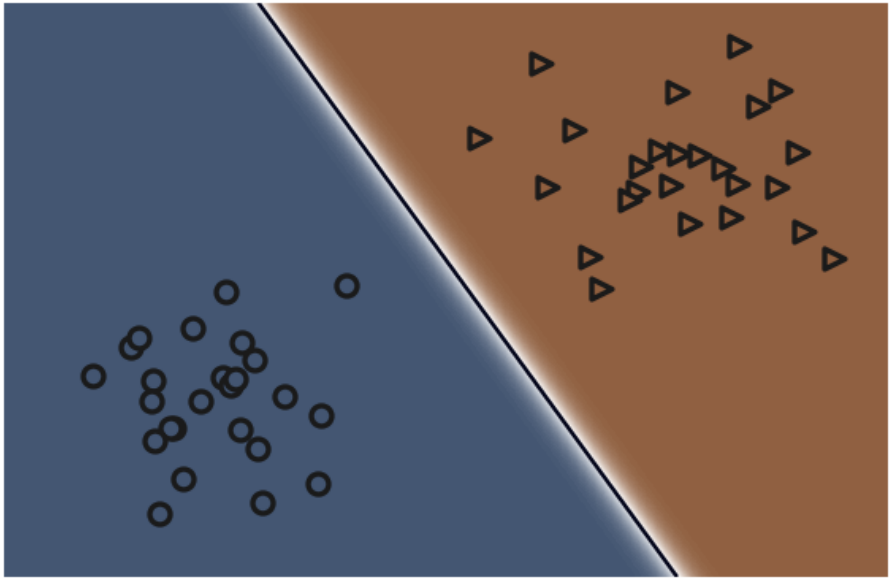
Bayes



Bayesian SAM



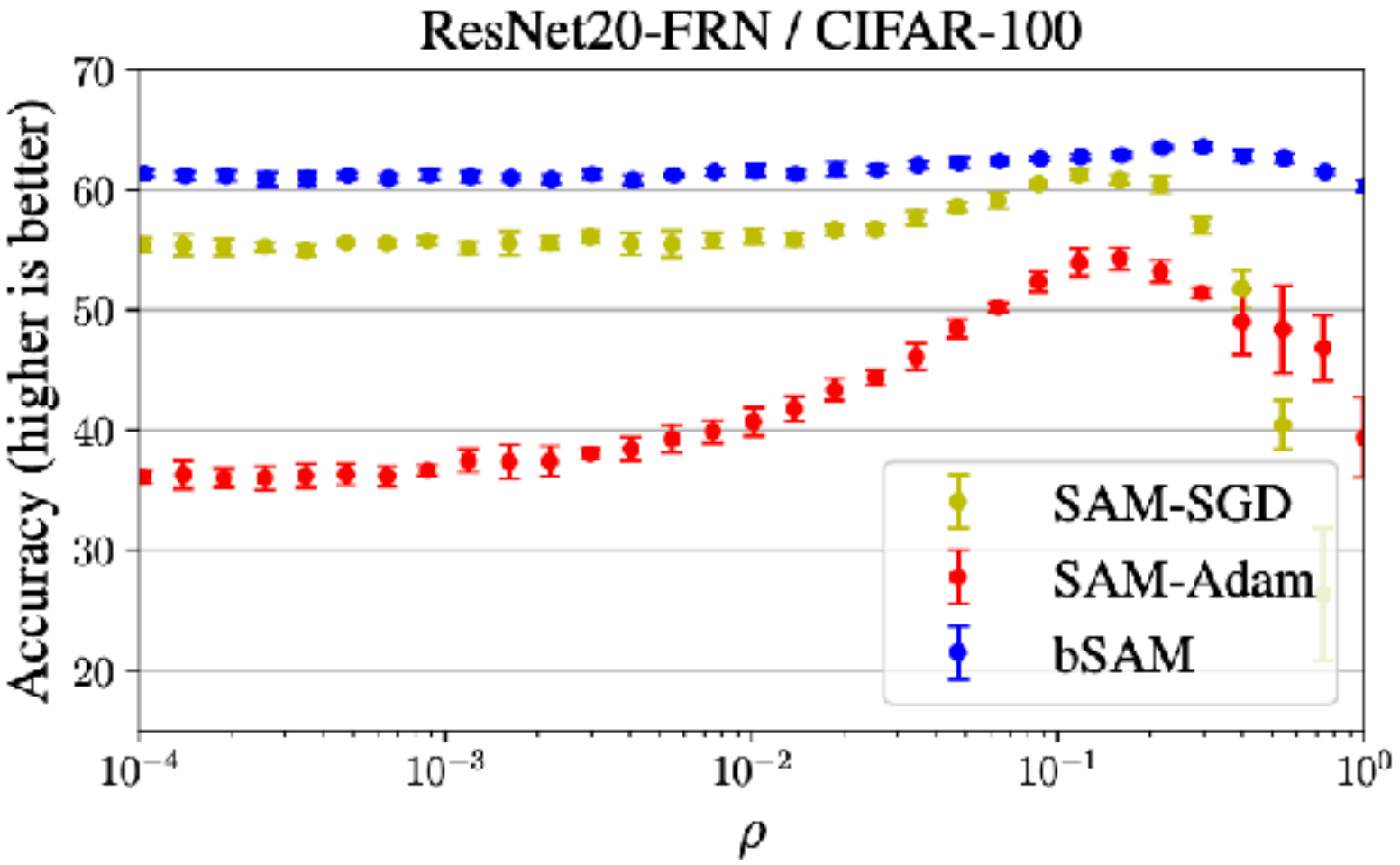
SAM



Results on Resnet-20 (200K params) and ResNet-18 (11M params)

Model / Dataset	Method	Accuracy ↑ (higher is better)	NLL ↓ (lower is better)	ECE ↓ (lower is better)	AUROC ↑ (higher is better)
ResNet-20-FRN / CIFAR-10	SGD	91.68 (0.26)	0.29 (0.008)	0.0397 (0.002)	0.915 (0.002)
	SAM-SGD	**92.29** (0.39)	0.25 (0.004)	0.0266 (0.003)	0.920 (0.003)
	Adam	89.97 (0.27)	0.41 (0.021)	0.0610 (0.003)	0.900 (0.003)
	SAM-Adam	91.57 (0.21)	0.26 (0.004)	0.0329 (0.002)	0.918 (0.001)
	bSAM (ours)	**92.16** (0.16)	**0.23** (0.003)	**0.0057** (0.002)	**0.925** (0.001)
ResNet-20-FRN / CIFAR-100	SGD	66.48 (0.10)	1.20 (0.007)	0.0524 (0.004)	0.846 (0.002)
	SAM-SGD	67.27 (0.22)	1.19 (0.011)	0.0481 (0.001)	0.848 (0.002)
	Adam	61.76 (0.67)	1.66 (0.049)	0.1582 (0.006)	0.826 (0.003)
	SAM-Adam	65.34 (0.32)	1.23 (0.012)	**0.0166** (0.003)	0.847 (0.002)
	bSAM (ours)	**68.22** (0.14)	**1.10** (0.013)	0.0258 (0.003)	**0.857** (0.004)
ResNet-20-FRN / TinyImageNet	SGD	52.01 (0.36)	1.98 (0.007)	0.0330 (0.002)	**0.832** (0.002)
	SAM-SGD	52.25 (0.26)	1.97 (0.013)	**0.0155** (0.002)	0.827 (0.005)
	Adam	49.04 (0.38)	2.14 (0.024)	0.0502 (0.004)	0.820 (0.004)
	SAM-Adam	51.17 (0.45)	2.02 (0.014)	0.0460 (0.004)	0.828 (0.004)
	bSAM (ours)	**52.90** (0.35)	**1.94** (0.009)	**0.0199** (0.003)	**0.831** (0.001)
ResNet-18 / CIFAR-10	SGD	94.76 (0.11)	0.21 (0.006)	0.0304 (0.001)	0.926 (0.006)
	SAM-SGD	95.72 (0.14)	0.14 (0.004)	0.0134 (0.001)	0.949 (0.003)
	bSAM (ours)	**96.15** (0.08)	**0.12** (0.002)	**0.0049** (0.001)	**0.954** (0.001)
ResNet-18 / CIFAR-100	SGD	76.54 (0.26)	0.98 (0.007)	0.0501 (0.002)	0.869 (0.003)
	SAM-SGD	78.74 (0.19)	0.79 (0.007)	0.0445 (0.002)	0.887 (0.003)
	bSAM (ours)	**80.22** (0.28)	**0.70** (0.008)	**0.0311** (0.003)	**0.892** (0.003)

Bayesian SAM is less sensitive to hyper-parameters



The webpage is available at https://bayesduality.github.io/, and Twitter account @BayesDuality

The Bayes-Duality Project

Toward AI that learns adaptively, robustly, and continuously, like humans

Emtiyaz Khan

Research director
(Japan side)

Approx-Bayes team at
RIKEN-AIP and OIST

Julyan Arbel

Research director
(France side)

Statify-team, Inria
Grenoble Rhône-Alpes

Kenichi Bannai

Co-PI (Japan side)

Math-Science Team at
RIKEN-AIP and Keio
University

Rio Yokota

Co-PI
(Japan side)

Tokyo Institute of
Technology

Received total funding of around USD 3 million through JST's CREST-ANR and Kakenhi Grants.

Approximate Bayesian Inference Team

https://team-approx-bayes.github.io/

Emtiyaz Khan
Team Leader

Pierre Alquier
Research Scientist

Hugo Monzón Maldonado
Postdoc

Happy Buzaaba
Postdoc

Erik Daxberger
Remote Collaborator
University of Cambridge

Paul Chang
Remote Collaborator
Aalto University

Gian Maria Marconi
Postdoc

Thomas Möllenhoff
Postdoc

Lu Xu
Postdoc

Jooyeon Kim
Postdoc

Alexandre Piché
Remote Collaborator
MILA

Geoffrey Wolfer
Postdoc

Wu Lin
PhD Student
University of British Columbia

Peter Nickl
Research Assistant

Dharmesh Tailor
Remote Collaborator
University of Amsterdam

Ang Mingliang
Remote Collaborator
National University of Singapore