DYNASTY:

DYNAMICS-AWARE THEORY OF DEEP LEARNING

Umut Şimşekli

Host institution: INRIA

ERC Starting Grant Interviews

October 8th, 2021

PRINCIPAL INVESTIGATOR: UMUT SIMSEKLI

Carrier Path:

- 2020 – Present: *Research Faculty* *INRIA – Ecole Normale Supérieure*, France
- 2019 – 2020: *Visiting Faculty Member* *University of Oxford*, UK
- 2016 – 2020: *Associate Professor* *Telecom ParisTech*, France
- 2010 – 2015: *PhD. in Computer Engineering* *Bogaziçi University*, Turkey

Updates: 8 new papers (since the proposal submission)

- **ICML 2021**:

  3 new papers (1 long oral presentation) – **2 preliminary studies** to this project

- **NeurIPS 2021**:

  5 new papers (1 spotlight presentation) – **4 preliminary studies** to this project

CONTEXT: DEEP LEARNING

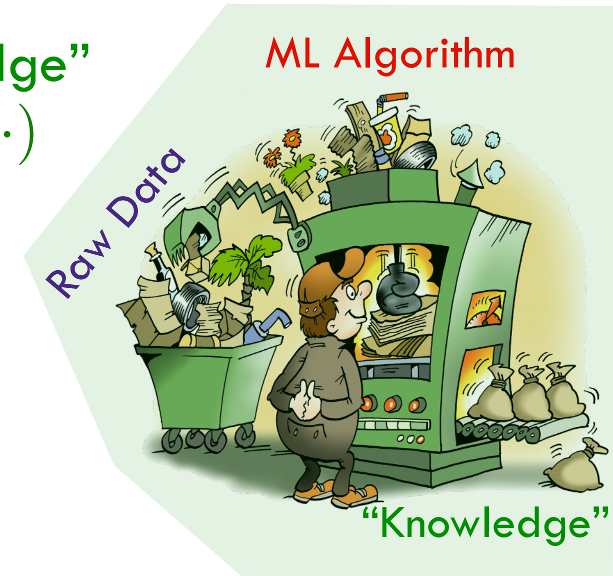
- **Machine Learning**: transformed many domains: industrial & academic

Raw data → ML algorithm → "Knowledge"

(x, y) $f(w, x) \approx y$ $f(w^\star, \cdot)$

(Image, Label)   (f: linear, w: parameter)

"≈" → optimization



- Last decade has witnessed a big increase in:

$\underbrace{\text{(Number of Data Points} \quad + \quad \text{Computation Power)}}_{\text{More and More Complicated Models}}$

- **Deep Learning (Neural Networks):** very complicated $f(w, x) \approx y$

Optimization Problem

$$\min_{w \in \mathbb{R}^d} \left\{ L(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(f(w, x_i), y_i) \right\}$$

$L(w)$: non-convex cost function

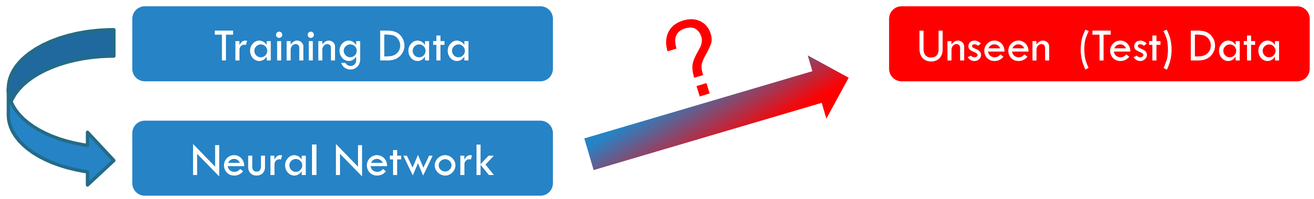
Optimization Algorithm (Training)

$$w_{k+1} = w_k - \eta \nabla \tilde{L}_k(w_k)$$

η: step-size (learning rate) $\nabla \tilde{L}_k(w_k)$: stochastic gradient

MOTIVATION

- **Deep Learning Theory** → Understand the "**Error on Unseen Test Data**"



- State of the art **upper bounds on "test error"**:

Shortcoming 1

Becomes **vacuous** with increasing **number of parameters**

(Neyshabur et al., NeurIPS 2017)

Shortcoming 2

Cannot capture the **interaction** between

- **data**
- model **architecture**
- optimization **algorithm**
- algorithm **hyperparameters**

(Zhang et al., NeurIPS 2020; Zhou et al., NeurIPS 2020)

Large Gap Between Theory and Practice

Current Deep Learning Systems:

- Poorly understood / black box

Designing New Methods:

- Trial&Error, ad-hoc, heuristic
- Time/energy consuming

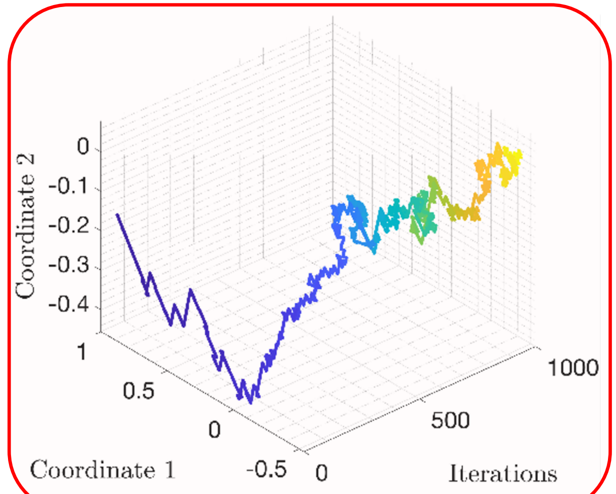
DYNASTY – GOALS & VISION

- **Ultimate Goals**

★ Mathematically **sound** & practically **relevant DL theory**
★ **Software library/practical tools** for DL practitioners

- **New Perspective:** ***"Dynamical Systems Theory"*** **(Pesin, 2008)**

Iterative Optimization → Training ***Trajectories*** → ***Stochastic Dynamical System***



  - Choice of the **optimization algorithm**
  - Algorithm **hyperparameters**
  - Training **data**
  - Model **architecture**

- **Four Main Challenges**

CHALLENGE 1: COMPLEXITY METRICS

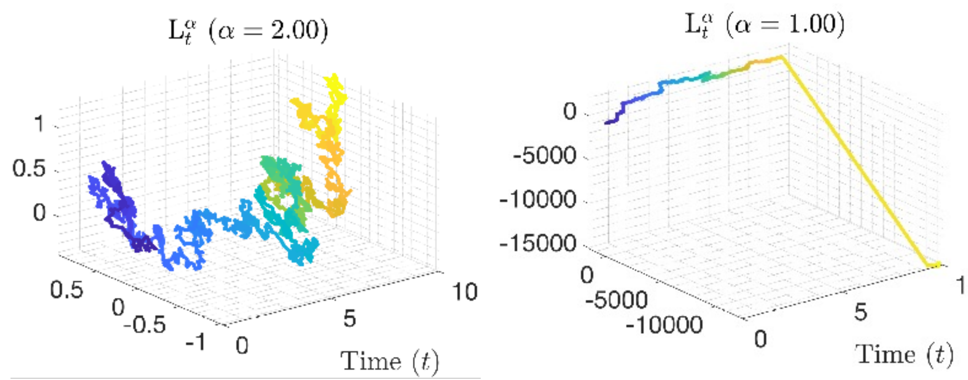
- Which **mathematical properties** of the dynamics ⇒ **Performance** ?

Hypothesis:

The performance is linked to the "**complexity**" **of the dynamics**

e.g., **Fractal Dimension**

(Falconer, 2014)



- **Expected Result**: novel **notions of complexity** → **error bounds**

→ reflects **practice**

- **Preliminary Studies:** [NeurIPS2020], [NeurIPS2021a], [NeurIPS2021b], [arXiv:2108.00781]

CHALLENGE 2: INTERACTION

The choices of

- Network **architecture**
- Training **data**
- Optimization **algorithm**
- Algorithm **hyperparameters**

Interact in a nontrivial way

? → **Complexity of Dynamics** → **Performance**

Hypothesis:

Affect the performance through a **common complexity metric**

- **Expected Result**: **rigorously link** these elements to the complexity metrics
- **Preliminary Studies: [ICML2021a], [ICML2021b], [NeurIPS2021a]**

CHALLENGE 3: NOVEL ALGORITHMS

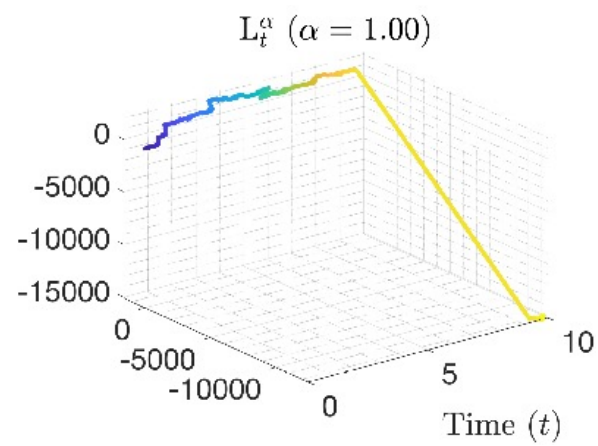
- **Task 1:** New **optimization algorithms** → exploit developed theory

  Improve the performance → **explicitly incorporate** the complexity metrics

- **Task 2:** New **compression algorithms**

  Hypothesis:

  The complexity metrics will be precisely linked to **compressibility**



- **Expected Result: improved performance** &reduced **storage complexity**
- **Preliminary Studies: [ICML2020], [NeurIPS2021c]**

CHALLENGE 4: DISSEMINATION

Proactive **dissemination** strategy

- **Practical** & **Open-Source** software libraries

  Evaluation → **predictive performance** and **complexity**

  Domains: Computer Vision, Audio/Music/Natural Language Processing

- **Expected Result**: software library → exploit **all previous outcomes**

  → **automatic model selection**

  → **adaptive optimization**

  will help liberate the trial/error design process

DYNASTY AT A GLANCE

- Fluency in **stochastic dynamical systems, non-convex optimization, high dimensional statistics, applications**

  My background lies at the **intersection**

- **Scientific impact** on disciplines using Deep Learning

- **Industrial impact** on e.g., automotive, marketing, entertainment

- **Team & Resources:**

  – PI, 3 PhD students, 2 postdocs, 1 engineer

  – Local support: learning theory/optimization/applications

  – Network of **international academic** (Oxford, Stanford, Berkeley) and **industrial** (Google, Facebook, Intel) **collaborators**

BUDGET

- Total requested grant: **€1.5M**
  - Principal Investigator (70%) **€330K**
  - 3 PhD Students **€360K**
  - 2 Postdocs (2 years each) **€230K**
  - 1 Research Engineer **€103K**

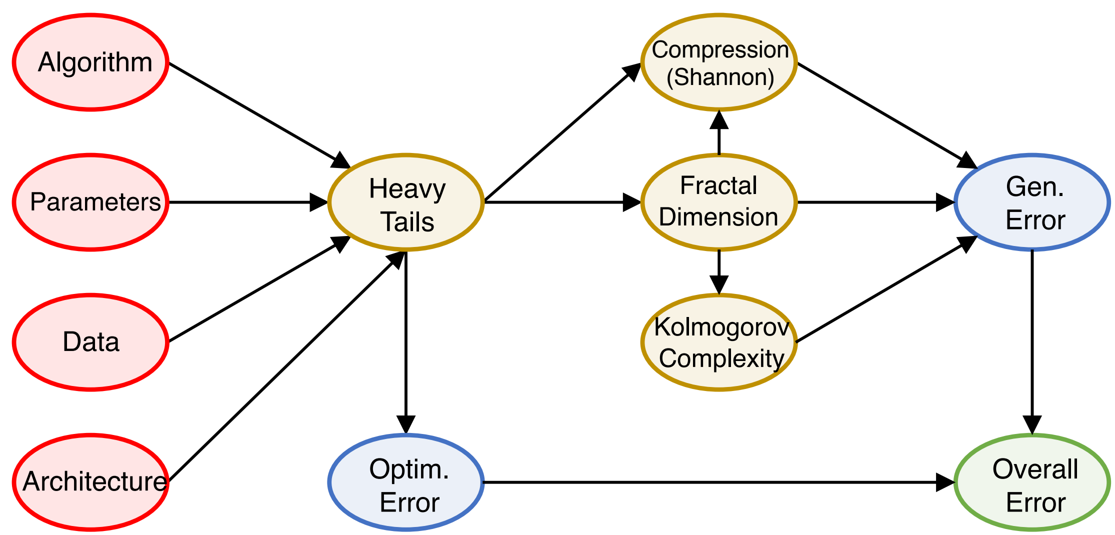
  - Travel (including invited researchers) **€88K**
  - Scientific Meetings **€50K**
  - Equipment **€30K**

WORK PACKAGES

- Overall organization

	C1 Complexity & generalization	C2 Quantification of interaction	C3 Improved algorithms	C4 Deployment & dissemination
WP1 - Empirical investigation				
WP2 - Error bounds				
WP3 - Algorithm development				
WP4 - Benchmarks				

- High-level roadmap



ORGANIZATION

- Initial fast pace → emphasis on theory
- Followed by the methodological developments

	Year 1	Year 2	Year 3	Year 4	Year 5
PhD Student 1 Fractal Dim. ⇆ Heavy Tails ⇆ Kolmogorov Cpx. ⇆ Generalization	Tasks 1.1, 2.1, 2.2, WP4				
PhD Student 2 Data, Algorithm, Parameters ⇆ Heavy Tails ⇆ Fractal Dim	Tasks 1.2, 2.3, WP4				
PhD Student 3 Novel Optimization Algorithms			Task 3.2, WP4		
Postdoc 1 Shannon Compression ⇆ Heavy Tails ⇆ Generalization	Tasks 2.1, 2.2, 3.3, WP4				
Postdoc 2 Optimization Bounds ⇆ Heavy Tails			Task 2.4, WP4		
Research Engineer Model Selection Algorithm, Open Source Dissemination					Task 3.1, WP4

REFERENCES

- **K. Falconer**. Fractal Geometry: *"Mathematical Foundations and Applications"*. John Wiley & Sons, **2004**.

- **B. Neyshabur**, S. Bhojanapalli, D. McAllester, and N. Srebro. *"Exploring generalization in deep learning"*. In: Advances in Neural Information Processing Systems (**NIPS**). **2017**

- **Y. B. Pesin.** *"Dimension Theory in Dynamical Systems: Contemporary Views and Applications"*. University of Chicago Press, **2008**.

- **J. Zhang**, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. *"Why are Adaptive Methods Good for Attention Models?"* In: Advances in Neural Information Processing Systems (**NeurIPS**). **2020**.

- **P. Zhou**, J. Feng, C. Ma, C. Xiong, S. C. H. Hoi, et al. *"Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning"*. In: Advances in Neural Information Processing Systems (**NeurIPS**). (**2020**).

PERSONAL REFERENCES

- **[NeurIPS2021a]** A. Camuto, G. Deligiannidis, M. A. Erdogdu, M. Gürbüzbalaban, U. Şimşekli & L. Zhu (**2021**). *"Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms"*. In: Advances in Neural Information Processing Systems (**NeurIPS**). **2021**.
- **[NeurIPS2021b]** H. Wang, M. Gürbüzbalaban, L. Zhu, U. Şimşekli, & M. A. Erdogdu (**2021**). *"Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance"*. In: Advances in Neural Information Processing Systems (**NeurIPS**). **2021**.
- **[NeurIPS2021c]** M. Barsbey, M. Sefidgaran, M. A. Erdogdu, G. Richard & U. Şimşekli (**2021**). *"Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks"*. In: Advances in Neural Information Processing Systems (**NeurIPS**). **2021**.
- **[ICML2021a]** M. Gurbuzbalaban, U. Simsekli, and L. Zhu. *"The Heavy-Tail Phenomenon in SGD"*, In: International Conference on Machine Learning (**ICML**) (**2021**)
- **[ICML2021b]** A. Camuto, X. Wang, L. Zhu, C. Holmes, M. Gurbuzbalaban, and U. Simsekli. *"Asymmetric Heavy Tails and Implicit Bias in Gaussian Noise Injections"*, In: International Conference on Machine Learning (**ICML**) (**2021**)
- **[NeurIPS2020]** U. Simsekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu. *"Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks"*. In: Advances in Neural Information Processing Systems (**NeurIPS**). **2020**.
- **[ICML2020]** U. Simsekli, L. Zhu, Y. W. Teh, and M. Gurbuzbalaban. *"Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise"*. In: International Conference on Machine Learning (**ICML**) (**2020**)
- **[arXiv:2108.00781]** L. Hodgkinson, U. Şimşekli, R. Khanna, & M. W. Mahoney. (**2021**). *"Generalization Properties of Stochastic Optimizers via Trajectory Analysis"*. **arXiv** preprint.