RESEARCH STATEMENT

Pratik Chaudhari

✉ pratikac@ucla.edu pratikac.info

It has been eight decades since Norbert Wiener first articulated the "cybernetic dream." Yet, despite recent excitement around Artificial Intelligence, we are still far from realizing robots that can interact intelligently with humans and the physical environment. My ambition is to bring this dream closer to reality, towards **Embodied Intelligence**. With this target in mind:

As an **Academic**, my first goal is to understand. I explore the interplay of optimization and generalization in deep networks[1,2], with some recent breakthroughs[3]. This is a cross-disciplinary effort spanning physics, non-convex optimization, and probability, in addition to classical statistical learning.

As an **Engineering Academic**, I believe the understanding should improve practice. I develop algorithms to train deep networks[2,4], that are faster and generalize better. My work on distributed algorithms[5] is being considered for deployment by some of the major technology companies.

As a **modern Engineering Academic**, I aim for broad societal impact. My work on robotics[6,7,8] has been incorporated into the world's first autonomous taxi service, launched in 2016 in Singapore.

As a **modern Engineering Academic Educator**, I aim to make powerful analytical and computational tools available to students in Computer Science, and to foster a collaborative, inter-disciplinary culture.

I next summarize my scientific accomplishments and articulate how they fit into my long-term research plan.

1 UNDERSTANDING DEEP LEARNING

Deep networks are machine learning models with a very large number, up to a hundred million, of parameters ("weights"). My work focuses on **algorithms to train deep networks**. This involves finding a weight vector x^* that optimizes some loss function, say the classification error on the training data. This is difficult because deep networks are highly non-convex functions of their weights. Further, it is challenging to ensure generalization, i.e., x^* should correctly classify unseen samples not part of the training data.

1.1 Generalization and optimization go hand-in-hand

Training a deep network in a way that ensures generalization is, at present, an art form. The gold standard of training algorithms is stochastic gradient descent (SGD). It updates weights using the gradient computed on a subset of samples ("mini-batch"), instead of the entire data; this makes it stochastic. When SGD is successful, deep networks generalize beyond what statistical learning theory suggests for such highly over-parametrized models. Understanding this conundrum can help improve learning algorithms.

Enter statistical physics: My first result in this area[2] shows that SGD, when it generalizes well, converges to "flat minima". I devised a modified loss, called **Local Entropy**, that biases optimization towards such regions. It is a local smoothing of the exponentiated original loss, motivated from statistical physics. The algorithm I designed to optimize this, named Entropy-SGD, converges up to $2\times$ faster than SGD while obtaining the state-of-the-art in generalization on benchmark datasets. I also proved accelerated convergence

over SGD[4] and generalization bounds[2] for this algorithm. This work **is influencing recent investigations** on generalization[9,10], non-convex optimization[11] and representation learning[12].

Connect with partial differential equations (PDEs): In a collaborative effort to connect Entropy-SGD to the standard literature on optimization, I showed that Local Entropy is the solution of a viscous Hamilton-Jacobi PDE[4] with the original loss as the initial condition. This connection brings the rich PDE literature to the table and is an interpretable way to exploit new PDEs to obtain even better performance. I proved that the non-viscous limit of Local Entropy leads to an established algorithm called **proximal point iteration**. The consequent insensitivity to hyper-parameters is very important in the practice of deep learning.

Obtain distributed algorithms: By exploiting connections of Local Entropy with statistical physics and PDEs, I, along with collaborators, devised an algorithm called Parle[5] for distributed training of deep networks that obtains **state-of-the-art** generalization and also converges 2-5$\times$ faster than SGD. We also proved[4] that Elastic-SGD[13], a popular distributed training algorithm for deep networks, actually minimizes Local Entropy. We are currently exploring extensions, as well as analysis, of this method.

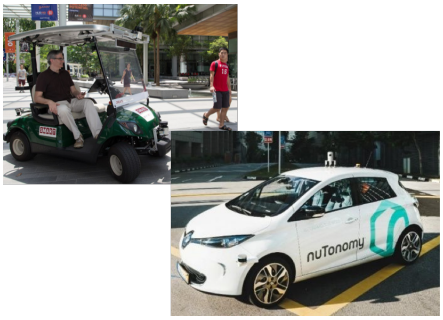
1.2 Mysteries of SGD on deep networks

It is widely believed that SGD performs implicit regularization. I proved this to be the case in a precise sense: **SGD performs variational inference** with a uniform prior[3]. However, SGD **minimizes a different loss**, which is, in general, not the original loss used to compute the gradients. I showed that the two are different in deep networks due to highly **non-isotropic noise** introduced by mini-batches[3]. This result recovers the celebrated Jordan-Kinderleher-Otto functional of optimal transportation and the two loss functions become equal, but only if gradient noise is isotropic.

Enter non-equilibrium thermodynamics: With non-vanishing step-sizes, iterates of SGD after convergence are expected to perform Brownian motion in the neighborhood of a critical point. I showed that non-isotropic noise leads to closed-loop trajectories in the weight space called **limit cycles**[3]. These loops can be far away from local minima; SGD may even converge to limit cycles around saddle points. This result is surprising as it challenges accepted wisdom in deep learning and calls for additional exploration.

2 Robotics

My goal is to develop **perception algorithms** *not* for classifying objects in images, but for robots to inhabit the physical world. Actions in the real world have consequences: an imperfect machine learning algorithm can be lethal when deployed on an autonomous vehicle. Being able to provide **provable guarantees for real-time algorithms** is the key to the continued development and societal acceptance of autonomy.



Societal impact: A prototype campus mobility system in 2011 to a commercial taxi-service in 2016.

I devised algorithms for **state-estimation and control**[14,15] based on novel discretizations of general stochastic dynamical systems using techniques from stochastic geometry. These algorithms are well-suited to real-time implementations, they provide a rough "guess" very quickly and improve upon it if additional computational resources become available. They are also optimal, i.e., the guess converges to the optimal output, asymptotically, with probability one, with very little work per iteration.

Building upon these techniques, I constructed algorithms for **provably-safe urban navigation**[6,8]. They use novel variants of Linear Temporal Logic to model road-safety rules and synthesize control actions to satisfy such high-level constraints. I further worked on game-theoretic approaches[7,16] using these ideas to devise algorithms for **multiple-agents** that share the same urban infrastructure while provably satisfying road rules.

This work has been incorporated into the **world's first autonomous taxi service launched by nuTonomy** in Singapore in 2016.

3 Future Work

3.1 Near-term: Deep Learning

Analysis in the continuum: The continuous-time limit provides powerful tools for the analysis and design of optimization algorithms[17] and I intend to continue working on it, towards the design of **algorithms for deep learning**. I aim to exploit ideas such as under-damped Langevin dynamics[18] from statistical physics for non-convex optimization and mean-field games[19], optimal transportation[20] from PDEs for the design of large-scale distributed training algorithms for deep networks.

Non-equilibrium phenomena: These are only now beginning to be explored, and little has been done to exploit them yet[21,22]. For instance, Markov Chain Monte Carlo samplers on deep networks automatically benefit from these effects. Connections of these ideas to **generalization in deep learning** may lead to improved design and training of deep networks, which I intend to explore.

This dual theme, analysis and informed design of algorithms for deep learning, will constitute the core of my first NSF proposal submission.

3.2 Long-term: Embodied Intelligence

Reinforcement learning: Scaling robot learning to real-world data and creating models that work for diverse robot sensors and actuators are present challenges that stand in the way of incorporating learning into robots. To mitigate these, I intend to leverage upon rich sources of data such as videos for **imitation learning** and devise algorithms for **transfer learning** policies that can be implemented on diverse robots and rapidly adapted to new operating environments. My ongoing work in this area, supervising a junior graduate student, composes pre-learned controllers, e.g., for hands and legs, to synthesize complex motions such as running, leveraging upon simulation and data with weak or no supervision.

Joint perception, learning and control: It is challenging to ascertain the performance of an autonomous system which performs actions on the basis of imperfect real-world data fed into machine learning models learned on curated datasets; uncertainty in sensing the environment further exacerbates this. Modern machine learning methods are agnostic to data modality. I intend to leverage upon this to **quantify the uncertainty** of the perception-learning-control loop and design robotic systems with **provable guarantees**, of safety and of optimality.

Intelligent beings are embodied: in a battle for survival, the best chess-playing program will lose to any dog. I see the physical form with its capacity to manipulate the surroundings as the hallmark of *intelligence*. Crucial to achieving this are *perception and control*. My explorations on learning and robotics are an effort to understand these two aspects of Embodied Intelligence and integrating them will be my focus for the foreseeable future.

REFERENCES

[1] Pratik Chaudhari and Stefano Soatto. On the energy landscape of deep networks. *Workshop on Advances in non-convex analysis and optimization, International Conference on Machine Learning (ICML)*, 2015. *arXiv:1511.06485*.

[2] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Proc. of International Conference of Learning and Representations (ICLR)*, 2017. *arXiv:1611.01838*.

[3] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *submitted to the International Conference of Learning and Representations (ICLR)*, 2017. *arXiv:1710.11029*.

[4] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier. Deep Relaxation: partial differential equations for optimizing deep neural networks. *Communications of Pure and Applied Mathematics (in review)*, 2017. *arXiv:1704.04932*.

[5] Pratik Chaudhari, Carlo Baldassi, Riccardo Zecchina, Stefano Soatto, Ameet Talwalkar, and Adam Oberman. Parle: parallelizing stochastic gradient descent. 2017. *arXiv:1707.00424*.

[6] Luis I. Reyes Castro, Pratik Chaudhari, Jana Tumova, Sertac Karaman, Emilio Frazzoli, and Daniela Rus. Incremental sampling-based algorithm for minimum-violation motion planning. In *Proc. of Conference on Decision and Control (CDC)*, 2013. *arXiv:1305.1102*.

[7] Pratik Chaudhari, Tichakorn Wongpiromsarn, and Emilio Frazzoli. Incremental minimum-violation control synthesis for robots interacting with external agents. In *Proc. of American Control Conference (ACC)*, 2014. *pdf*.

[8] Valerio Varricchio, Pratik Chaudhari, and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning using process algebra specifications. In *Proc. of International Conference on Robotics and Automation (ICRA)*, 2014. *pdf*.

[9] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv:1706.08947*, 2017.

[10] Anonymous. Entropy-SG(L)D optimizes the prior of a (valid) PAC-Bayes bound. *submitted to ICLR*, 2017.

[11] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *arXiv:1702.03849*, 2017.

[12] Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv:1706.01350*, 2017.

[13] Sixin Zhang, Anna Choromanska, and Yann LeCun. Deep learning with elastic averaging SGD. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[14] Pratik Chaudhari, Sertac Karaman, and Emilio Frazzoli. Sampling-based algorithm for filtering using Markov chain approximations. In *Proc. of Conference on Decision and Control (CDC)*, 2012. *pdf*.

[15] Pratik Chaudhari, Sertac Karaman, David Hsu, and Emilio Frazzoli. Sampling-based algorithms for continuous-time POMDPs. In *Proc. of American Control Conference (ACC)*, 2013. *pdf*.

[16] Minghui Zhu, Michael Otte, Pratik Chaudhari, and Emilio Frazzoli. Game theoretic controller synthesis for multi-robot motion planning Part I: Trajectory based algorithms. In *Proc. of International Conference on Robotics and Automation (ICRA)*, 2014. *arXiv:1402.2708*.

[17] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *PNAS*, page 201614734, 2016.

[18] Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv:1710.00095*, 2017.

[19] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

[20] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkauser, NY*, 2015.

[21] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Lucibello Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical review letters*, 115(12):128101, 2015.

[22] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.