UNIVERSITY OF
OXFORD

Towards Certifiable Machine Learning

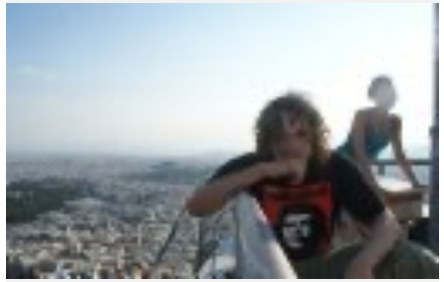
Francisco Girbal Eiras

Torr Vision Group · EPSRC CDT Autonomous Intelligent Machines and Systems (AIMS)

July 2023

UNIVERSITY OF OXFORD

Francisco Eiras
University of Oxford
eiras@robots.ox.ac.uk



Philip H.S. Torr
University of Oxford
philip.torr@eng.ox.ac.uk

Adel Bibi
University of Oxford
adel.bibi@eng.ox.ac.uk

PhD supervisors

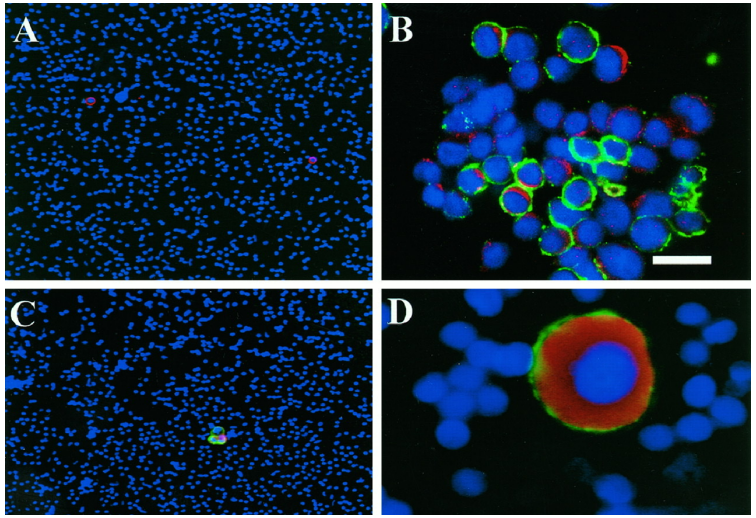
M. Pawan Kumar
Google DeepMind
mpawan@google.com

Industry supervisor

UNIVERSITY OF OXFORD

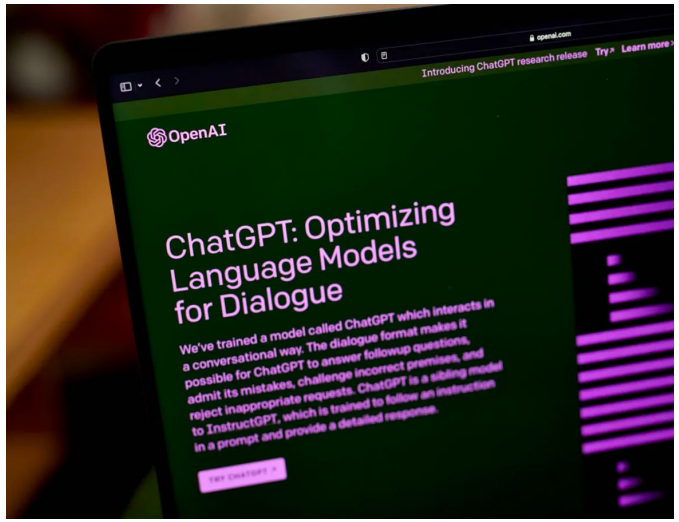
Why is Certified Machine Learning *important*?

Safety in Human interaction



Cancer cell detection in the bloodstream

Self-driving cars and autonomous vehicles



Large generative AI models

Safety in Human interaction

UNIVERSITY OF OXFORD

Interaction with humans → makes them **safety-critical** situations

Requires guarantees to be given about their performance!

Minimum guarantee is robustness, *i.e., the ability to tolerate small input perturbations*.

Safety in Human interaction – Adv. Robustness (?)

UNIVERSITY OF OXFORD

- The existence of adversarial examples is well known, particularly in image classification.
- *Is it a* **safety** *problem?*

$$\max_{\delta \in \mathbb{B}_\epsilon} \ell\left(f_\theta(x+\delta), y\right)$$

- They are an interesting phenomenon, potentially attributed to *non-robust/spurious* features.
- **Certification is still important**

Distill ABOUT PRIZE SUBM

A Discussion of *Adversarial Examples Are Not Bugs, They Are Features*

PUBLISHED Aug. 6, 2019

DOI 10.23915/distill.00019

On May 6th, Andrew Ilyas and colleagues published a paper [1] outlining two sets of experiments. Firstly, they showed that models trained on adversarial examples can transfer to real data, and secondly that models trained on a dataset derived from the representations

UNIVERSITY OF OXFORD

01

Safety in Human Interaction/ Understanding Decision-Making Process

Correctness of Machine Learning based-Systems

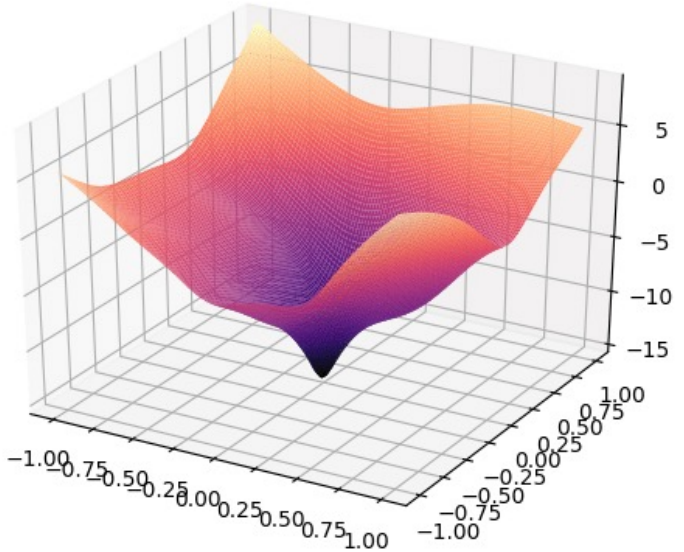
UNIVERSITY OF OXFORD

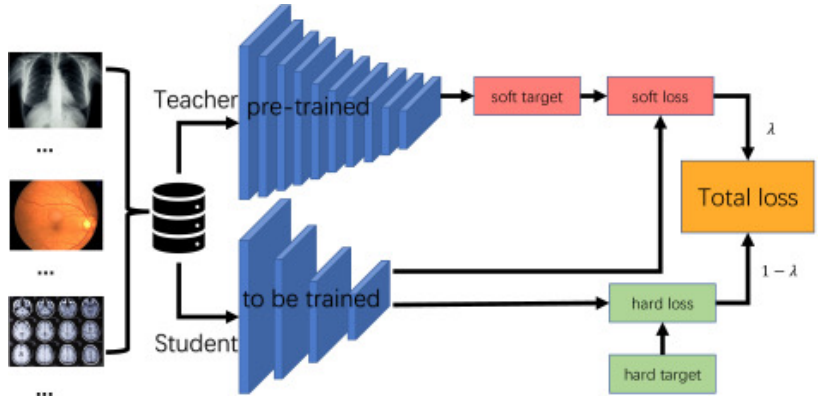
ML is used within systems to replace experts/speed-up processes.

E.g., to solve PDEs or knowledge distillation.

Deployed in real-world, but can we trust it to behave **correctly**?

Certified machine learning could help establish reliability/correctness of these components.





UNIVERSITY OF OXFORD

01

Safety in Human Interaction/
Understanding Decision-Making Process

02

Correctness of ML-based Systems

UNIVERSITY OF OXFORD

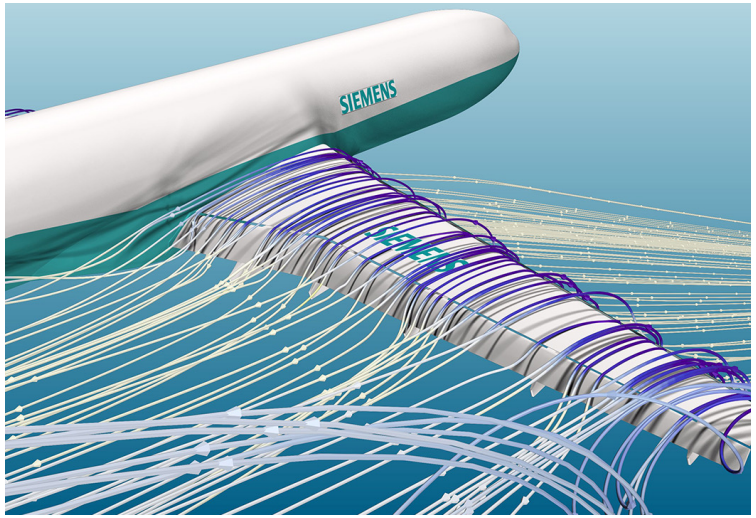
01

Safety in Human Interaction/
Understanding Decision-Making Process

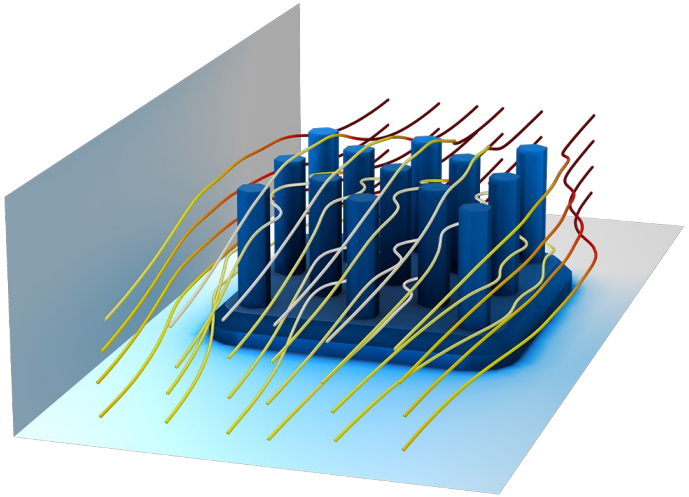
02

Correctness of ML-based Systems

Physical PDEs and Where to Find Them



Aerodynamics (e.g., Euler's equation)



Thermodynamics (e.g., Heat equation)

Earth and Space Sciences (e.g., turbulence in 2-D Navier-Stokes)

Among many others...

Nonlinear (Physical) Partial Differential Equations

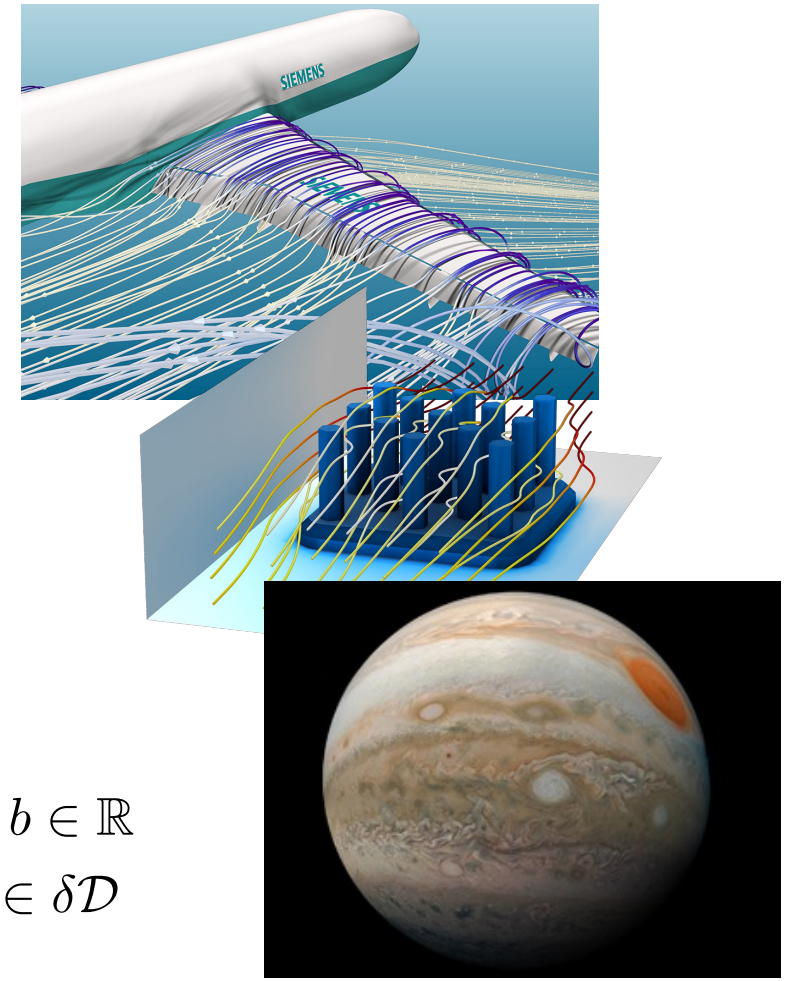
UNIVERSITY OF OXFORD

solution

nonlinear spatial differential operator

$$\underbrace{\partial_t u(t,x) + \mathcal{N}[u](t,x)}_{\text{residual}} = 0, \quad x \in \mathcal{D}, t \in [0,T]$$

domain



s.t.

1. **Initial condition**: $u(0,x) = u_0(x)$
2. **Robin boundary conditions**: $au(t,x) + b\partial_{\mathbf{n}} u(t,x) = u_b(t,x) \quad a,b \in \mathbb{R}$, $x \in \delta\mathcal{D}$

PDE example: Diffusion-Sorption Equation

- Applications in groundwater contaminant transportation
- **1D Diffusion-Sorption**:

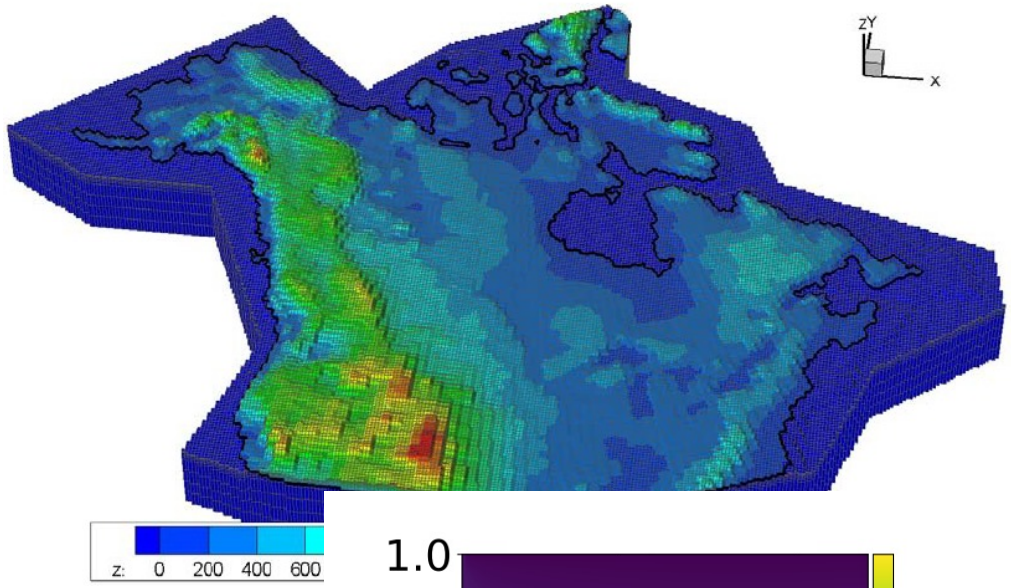
$$\partial_t u(t,x) - D/R(u(t,x))\partial_{x^2} u(t,x) = 0$$

where

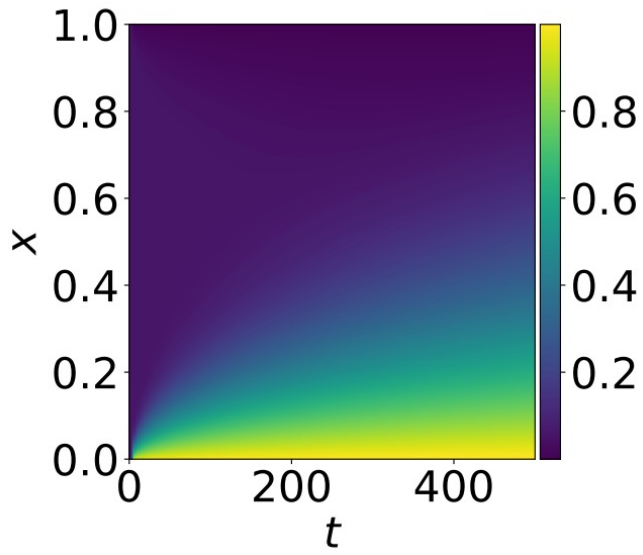
$$R(u(t,x)) = 1 + \frac{(1-\phi)}{(\phi)}\rho_s k n_f u^{n_f - 1}(t,x)$$

for physical constants D, ϕ, ρ_s, k, n_f, and:

$$u(0,x) = u(t,0) = 0, \quad u(t,1) = D\partial_x u(t,1)$$



1D Diffusion-Sorption



Takamoto, Makoto, et al. "PDEBench: An extensive benchmark for scientific machine learning." Advances in Neural Information Processing Systems 35 (2022): 1596-1611.

Lemieux, JM, et al. "Coupling continental glaciations with groundwater flow models–Surface/subsurface interactions over the Canadian landscape during the Wisconsinian glaciation."

UNIVERSITY OF OXFORD

Issue: Solving for $u(t, x)$ is computationally expensive

Solution: Use NNs to approximate it

Physics-Informed Neural Networks (PINNs)

UNIVERSITY OF OXFORD

- Approximate solution using a neural network, $u_\theta(t,x) \simeq u(t,x)$
- Take the residual evaluated for u_θ as the network $f_\theta(t,x) = \partial_t u_\theta(t,x) + \mathcal{N}[u_\theta](t,x)$

PINN

- Train *both networks* jointly using a loss evaluated at collocation points $\mathbb{P}$ (i.e. points in the domain):

$$\mathcal{L} = \underbrace{\sum_{x \in \mathbb{P}_0} |u(0,x) - u_\theta(0,x)|^2}_{\text{initial conditions}} + \underbrace{\sum_{(t,x) \in \mathbb{P}_b} |u(t,x) - u_\theta(t,x)|^2}_{\text{boundary conditions}} + \underbrace{\sum_{(t,x) \in \mathbb{P}_f} |f_\theta(t,x)|^2}_{\text{residual}}$$

- **Evaluate empirically by comparing $u_\theta(t,x)$ to the solution obtained by a numerical solver**

Raissi, Maziar, Paris Perdikaris, and George E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." Journal of Computational physics 378 (2019): 686-707.

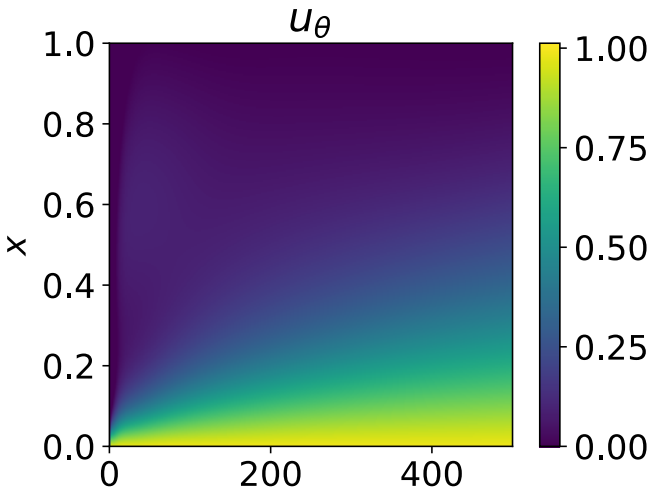
PINN: 1D Diffusion-Sorption Equation

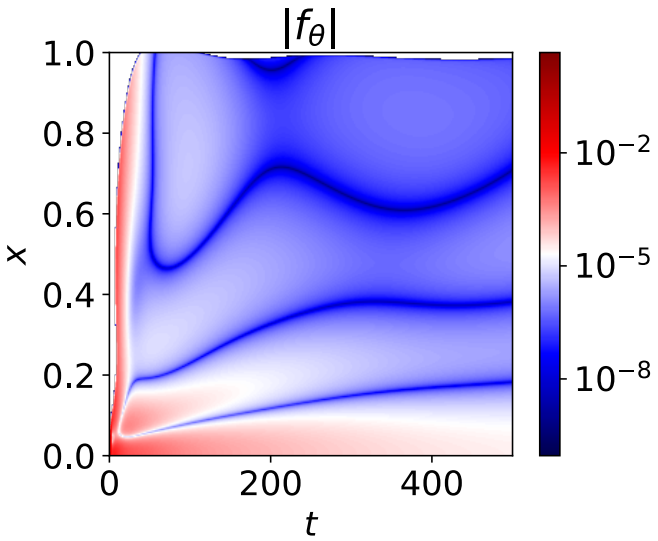
- Inference times on 2 core CPU + 1 GPU (NVIDIA V100):

Numerical Solver	PINN [Takamoto et al. 2022]
59.83 s	2.7 x 10^{-3} s

- PINN average ℓ_2 solution error is 9.9 x 10^{-2} compared to numerical solver
- Valid PDE solution **must** satisfy $f_\theta(t,x) = 0$
- *Is it satisfied across the domain?*

$\max \lvert f_\theta \rvert^2$	10^4 Uniform Samples	10^6 Uniform Samples
	1.1 x 10^{-3}	21.09





Takamoto, Makoto, et al. "PDEBench: An extensive benchmark for scientific machine learning." Advances in Neural Information Processing Systems 35 (2022): 1596-1611.

UNIVERSITY OF OXFORD

How can we be confident errors are small enough across the *entire* domain?

Defining Correctness Conditions for PINNs (PINN, 2023)

UNIVERSITY OF OXFORD

- Intuitively, a PINN is a correct approximation of the underlying PDE if:
  1. *The solution satisfies the initial conditions* to a reasonable degree
  2. *The solution satisfies the boundary conditions* to a reasonable degree
  3. The *norm of the PINN output is small* enough

- Formally, for a D dimensional spatial input $\hat{\mathbf{x}} \in \mathcal{D}$, and solution/PINN input $\mathbf{x} = (t, \hat{\mathbf{x}})$:

Definition 1 (Correctness Conditions for PINNs). *$u_\theta : [0,T] \times \mathcal{D} \to \mathbb{R}$ is a $\delta_0, \delta_b, \varepsilon$-globally correct approximation of the exact solution $u : [0,T] \times \mathcal{D} \to \mathbb{R}$ if:*

$$\textcircled{1} \quad \max_{\hat{\mathbf{x}} \in \mathcal{D}} |u_\theta(0, \hat{\mathbf{x}}) - u_0(\hat{\mathbf{x}})|^2 \leq \delta_0,$$

$$\textcircled{2} \quad \max_{t \in [0,T], \hat{\mathbf{x}} \in \delta\mathcal{D}} |a u_\theta(t, \hat{\mathbf{x}}) + b \partial_{\mathbf{n}} u_\theta(t, \hat{\mathbf{x}}) - u_b(t, \hat{\mathbf{x}})|^2 \leq \delta_b,$$

$$\textcircled{3} \quad \max_{\mathbf{x} \in \mathcal{C}} |f_\theta(\mathbf{x})|^2 \leq \varepsilon.$$

Scalable Adversarial Robustness Verification/Certification

UNIVERSITY OF OXFORD

- Complete vs. incomplete methods:
  - *Can it certify a ResNet50? If YES → incomplete.*

- "Sound, but incomplete" methods:
  - Bound propagation (e.g., DiffAI, IBP)
  - **Convex relaxation optimization-based methods** (e.g. CROWN)
  - Smoothing-based methods (e.g. randomized smoothing)

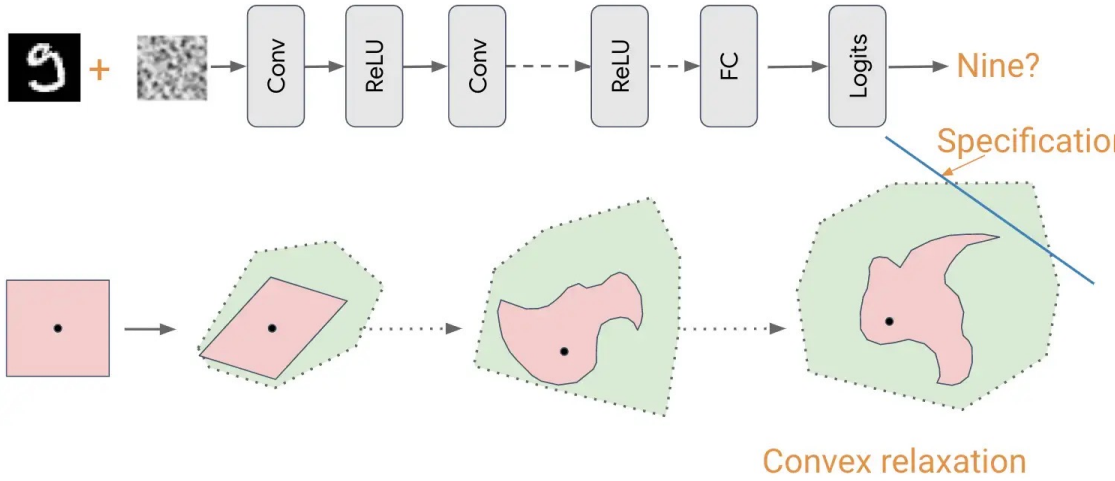
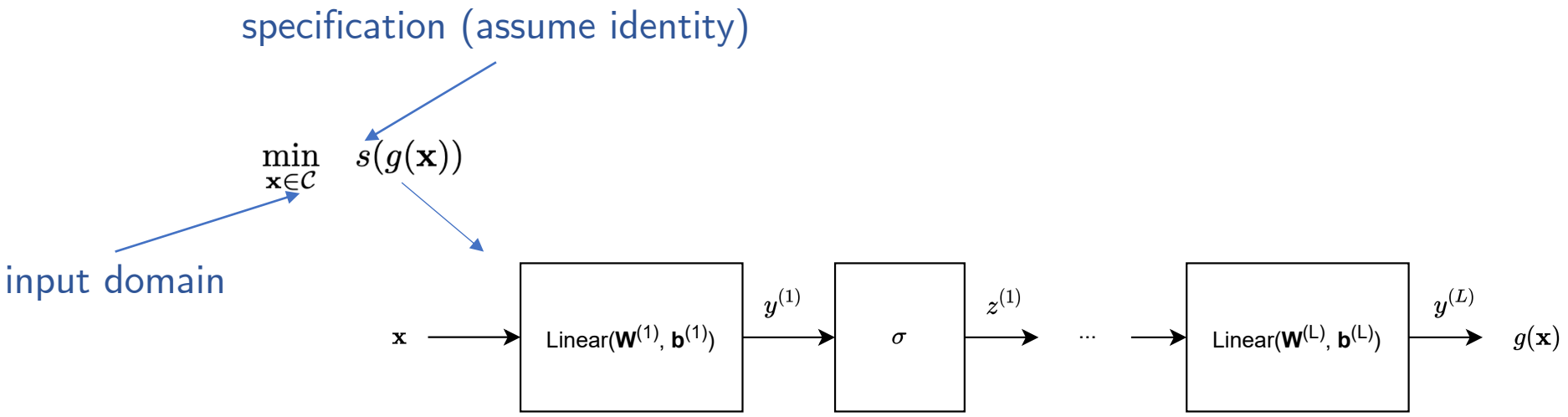


Image from "Efficient and tight neural network verification in JAX" (https://www.deepmind.com/open-source/efficient-and-tight-neural-network-verification-in-jax)

Verification of Neural Networks

- CROWN/α-CROWN [Zhang et. al 2018, Xu et. al 2020]



- Relax network to a linear program → solve it in closed form: $\min_{\mathbf{x}\in\mathcal{C}} \mathbf{A}^L\mathbf{x} + \mathbf{a}^L = l_b \leq \min_{\mathbf{x}\in\mathcal{C}} g(\mathbf{x})$

relaxed problem

- Apply it to u_θ and f_θ using the previous specifications

Zhang, Huan, et al. "Efficient neural network robustness certification with general activation functions." Advances in neural information processing systems 31 (2018).

Xu, Kaidi, et al. "Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers." arXiv preprint arXiv:2011.13824 (2020).

Correctness Certification of PINNs (PINN, 2023)

UNIVERSITY OF OXFORD

Definition 1 (Correctness Conditions for PINNs). *$u_\theta : [0,T] \times \mathcal{D} \to \mathbb{R}$ is a $\delta_0, \delta_b, \varepsilon$-globally correct approximation of the exact solution $u : [0,T] \times \mathcal{D} \to \mathbb{R}$ if:*

✅ $\max_{\hat{\mathbf{x}} \in \mathcal{D}} |u_\theta(0, \hat{\mathbf{x}}) - u_0(\hat{\mathbf{x}})|^2 \leq \delta_0,$

❌ $\max_{t \in [0,T], \hat{\mathbf{x}} \in \delta\mathcal{D}} |a u_\theta(t, \hat{\mathbf{x}}) + b \partial_{\mathbf{n}} u_\theta(t, \hat{\mathbf{x}}) - u_b(t, \hat{\mathbf{x}})|^2 \leq \delta_b,$

❌ $\max_{\mathbf{x} \in \mathcal{C}} |f_\theta(\mathbf{x})|^2 \leq \varepsilon.$

- *Applying CROWN/α-CROWN to boundary/residual conditions*:
  - ❌ Architecture is completely different - f_θ is a nonlinear function of partial derivatives of u_θ
  - ❌ Regression problem – bounds might be too loose to be informative

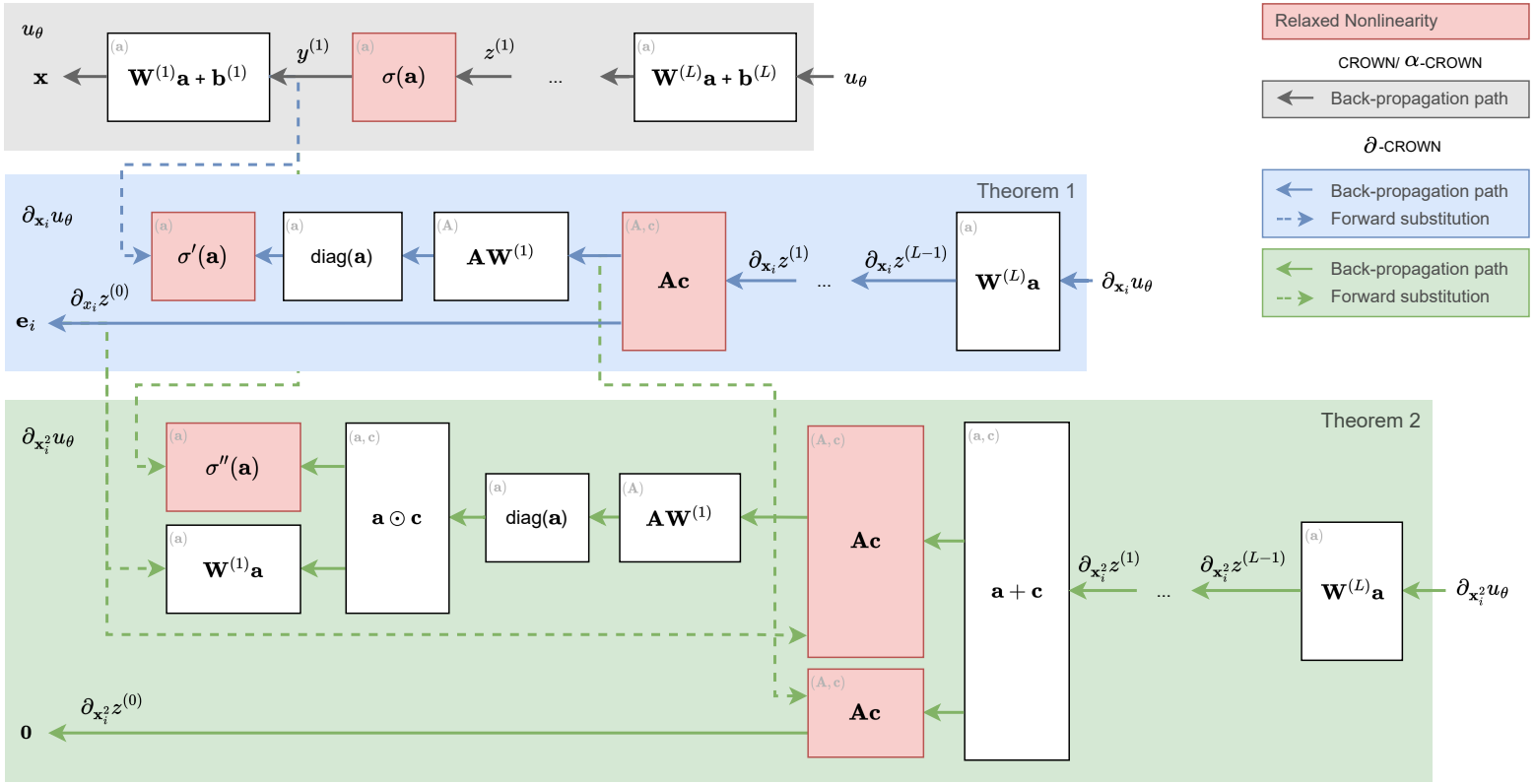
∂-CROWN: Bounding Derivatives of u_θ and f_θ (PINN, 2023)

- u_θ is bound using CROWN [Zhang et al. 2018]; partial derivatives require purpose-built efficient solution

1st (Theorem 1) and 2nd (Theorem 2) linear bounding of partial derivatives

Hybrid scheme with complexity $\mathcal{O}(L)$ instead of $\mathcal{O}(L^2)$ from [Xu et al. 2020]

Global bounds computed in close-form (similarly to [Zhang et al. 2018])



- f_θ is linearly bounded using McCormick envelopes; global bounds computed in close-form

Zhang, Huan, et al. "Efficient neural network robustness certification with general activation functions." Advances in neural information processing systems 31 (2018).

Xu, Kaidi, et al. "Automatic perturbation analysis for scalable certified robustness and beyond." Advances in Neural Information Processing Systems 33 (2020): 1129-1141.

Correctness Certification of PINNs (PINN, 2023)

UNIVERSITY OF OXFORD

Definition 1 (Correctness Conditions for PINNs). *$u_\theta : [0,T] \times \mathcal{D} \to \mathbb{R}$ is a $\delta_0, \delta_b, \varepsilon$-globally correct approximation of the exact solution $u : [0,T] \times \mathcal{D} \to \mathbb{R}$ if:*

✅ $\max_{\hat{\mathbf{x}} \in \mathcal{D}} |u_\theta(0, \hat{\mathbf{x}}) - u_0(\hat{\mathbf{x}})|^2 \leq \delta_0,$

❌ $\max_{t \in [0,T], \hat{\mathbf{x}} \in \delta\mathcal{D}} |a u_\theta(t, \hat{\mathbf{x}}) + b \partial_{\mathbf{n}} u_\theta(t, \hat{\mathbf{x}}) - u_b(t, \hat{\mathbf{x}})|^2 \leq \delta_b,$

❌ $\max_{\mathbf{x} \in \mathcal{C}} |f_\theta(\mathbf{x})|^2 \leq \varepsilon.$

- *Applying CROWN/α-CROWN to boundary/residual conditions*:
  - ✅ Architecture is completely different - f_θ is a nonlinear function of partial derivatives of u_θ
  - ❌ Regression problem – bounds might be too loose to be informative

Greedy Input Branching

Correctness Certification of PINNs (PINN, 2023)

UNIVERSITY OF OXFORD

Definition 1 (Correctness Conditions for PINNs). *$u_\theta : [0,T] \times \mathcal{D} \to \mathbb{R}$ is a $\delta_0, \delta_b, \varepsilon$-globally correct approximation of the exact solution $u : [0,T] \times \mathcal{D} \to \mathbb{R}$ if:*

✅ $\max_{\hat{\mathbf{x}} \in \mathcal{D}} |u_\theta(0, \hat{\mathbf{x}}) - u_0(\hat{\mathbf{x}})|^2 \leq \delta_0,$

✅ $\max_{t \in [0,T], \hat{\mathbf{x}} \in \delta\mathcal{D}} |a u_\theta(t, \hat{\mathbf{x}}) + b \partial_{\mathbf{n}} u_\theta(t, \hat{\mathbf{x}}) - u_b(t, \hat{\mathbf{x}})|^2 \leq \delta_b,$

✅ $\max_{\mathbf{x} \in \mathcal{C}} |f_\theta(\mathbf{x})|^2 \leq \varepsilon.$

- *Applying CROWN/α-CROWN to boundary/residual conditions*:
  - ✅ Architecture is completely different - f_θ is a nonlinear function of partial derivatives of u_θ
  - ✅ Regression problem – bounds might be too loose to be informative

Experiments: Certifying with ∂-CROWN (PINN, 2023)

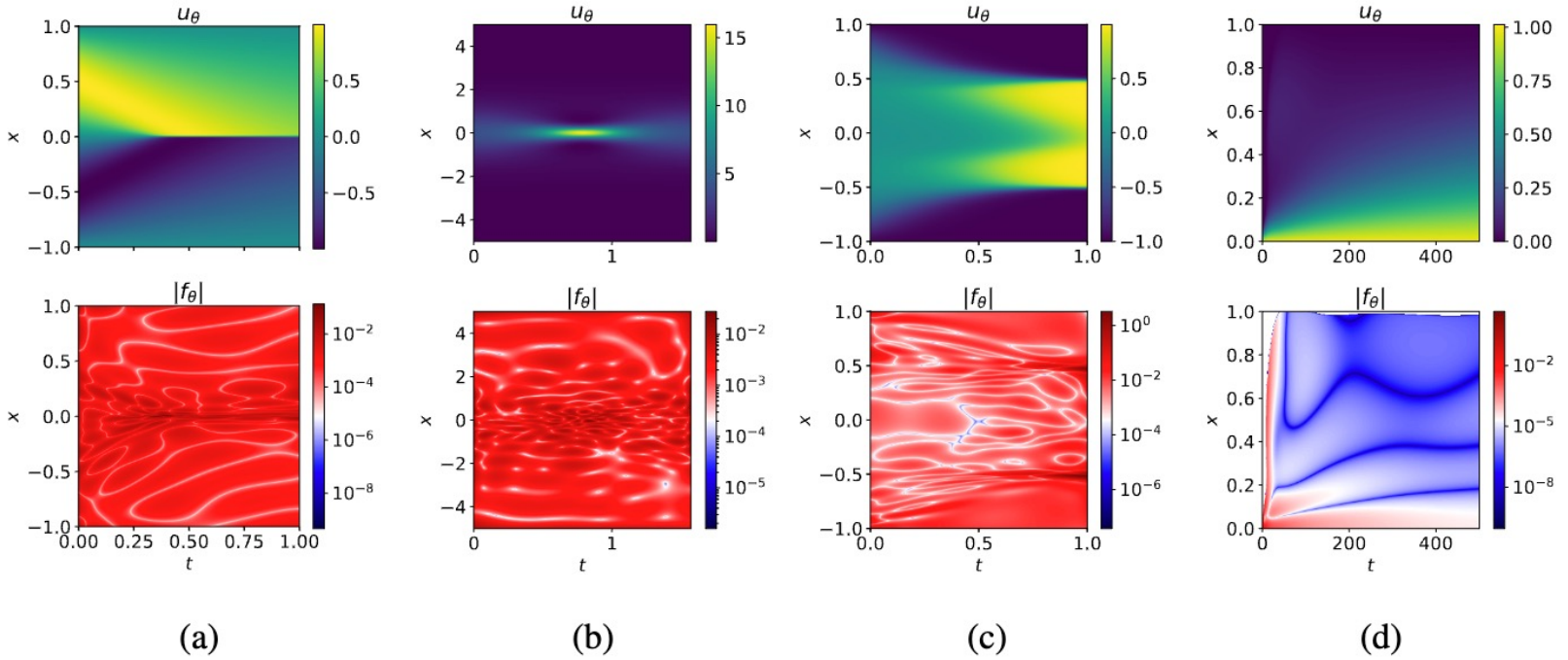


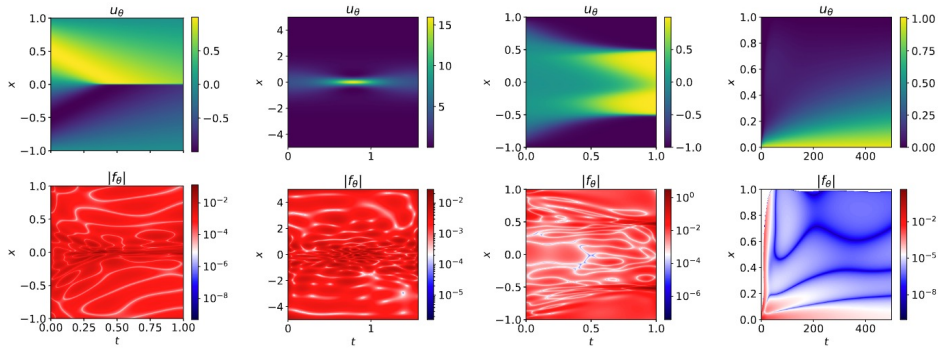
Figure 2: **Certifying with ∂-CROWN**: visualization of the time evolution of u_θ, and the residual errors as a function of the spatial temporal domain (log-scale), $|f_\theta|$, for **(a)** Burgers' equation [Raissi et al., 2019b], **(b)** Schrödinger's equation [Raissi et al., 2019b], **(c)** Allan-Cahn's equation [Monaco and Apiletti, 2023], and **(d)** the Diffusion-Sorption equation [Takamoto et al., 2022].

Raissi, Maziar, Paris Perdikaris, and George E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." Journal of Computational physics 378 (2019): 686-707.

Monaco, Simone, and Daniele Apiletti. "Training physics-informed neural networks: One learning to rule them all?." Results in Engineering 18 (2023): 101023.

Takamoto, Makoto, et al. "PDEBench: An extensive benchmark for scientific machine learning." Advances in Neural Information Processing Systems 35 (2022): 1596-1611.

Experiments: Certifying with ∂-CROWN (PINN, 2023)



	MC max (10^4)	MC max (10^6)	∂-CROWN u_b (time [s])
(a) **Burgers** [Raissi et al., 2019b]			
① $\lvert u_\theta(0,x) - u_0(x)\rvert^2$	1.59×10^{-6}	1.59×10^{-6}	2.63×10^{-6} (116.5)
② $\lvert u_\theta(t,-1)\rvert^2$	8.08×10^{-8}	8.08×10^{-8}	6.63×10^{-7} (86.7)
② $\lvert u_\theta(t,1)\rvert^2$	6.54×10^{-8}	6.54×10^{-8}	9.39×10^{-7} (89.8)
③ $\lvert f_\theta(x,t)\rvert^2$	1.23×10^{-2}	1.80×10^{-2}	1.03×10^{-1} (2.8×10^5)
(b) **Schrödinger** [Raissi et al., 2019b]			
① $\lvert u_\theta(0,x) - u_0(x)\rvert^2$	7.06×10^{-5}	7.06×10^{-5}	8.35×10^{-5} (305.2)
② $\lvert u_\theta(t,5) - u_\theta(t,-5)\rvert^2$	7.38×10^{-7}	7.38×10^{-7}	5.73×10^{-6} (545.4)
② $\lvert \partial_x u_\theta(t,5) - \partial_x u_\theta(t,-5)\rvert^2$	1.14×10^{-5}	1.14×10^{-5}	5.31×10^{-5} (2.4×10^3)
③ $\lvert f_\theta(x,t)\rvert^2$	7.28×10^{-4}	7.67×10^{-4}	5.55×10^{-3} (1.2×10^6)
(c) **Allen-Cahn** [Monaco and Apiletti, 2023]			
① $\lvert u_\theta(0,x) - u_0(x)\rvert^2$	1.60×10^{-3}	1.60×10^{-3}	1.61×10^{-3} (52.7)
② $\lvert u_\theta(t,-1) - u_\theta(t,1)\rvert^2$	5.66×10^{-6}	5.66×10^{-6}	5.66×10^{-6} (95.4)
③ $\lvert f_\theta(x,t)\rvert^2$	10.74	10.76	10.84 (6.7×10^5)
(d) **Diffusion-Sorption** [Takamoto et al., 2022]			
① $\lvert u_\theta(0,x)\rvert^2$	0.0	0.0	0.0 (0.2)
② $\lvert u_\theta(t,0) - 1\rvert^2$	4.22×10^{-4}	4.39×10^{-4}	1.09×10^{-3} (72.5)
② $\lvert u_\theta(t,1) - D\partial_x u_\theta(t,1)\rvert^2$	2.30×10^{-5}	2.34×10^{-5}	2.37×10^{-5} (226.4)
③ $\lvert f_\theta(x,t)\rvert^2$	1.10×10^{-3}	21.09	21.34 (2.4×10^6)

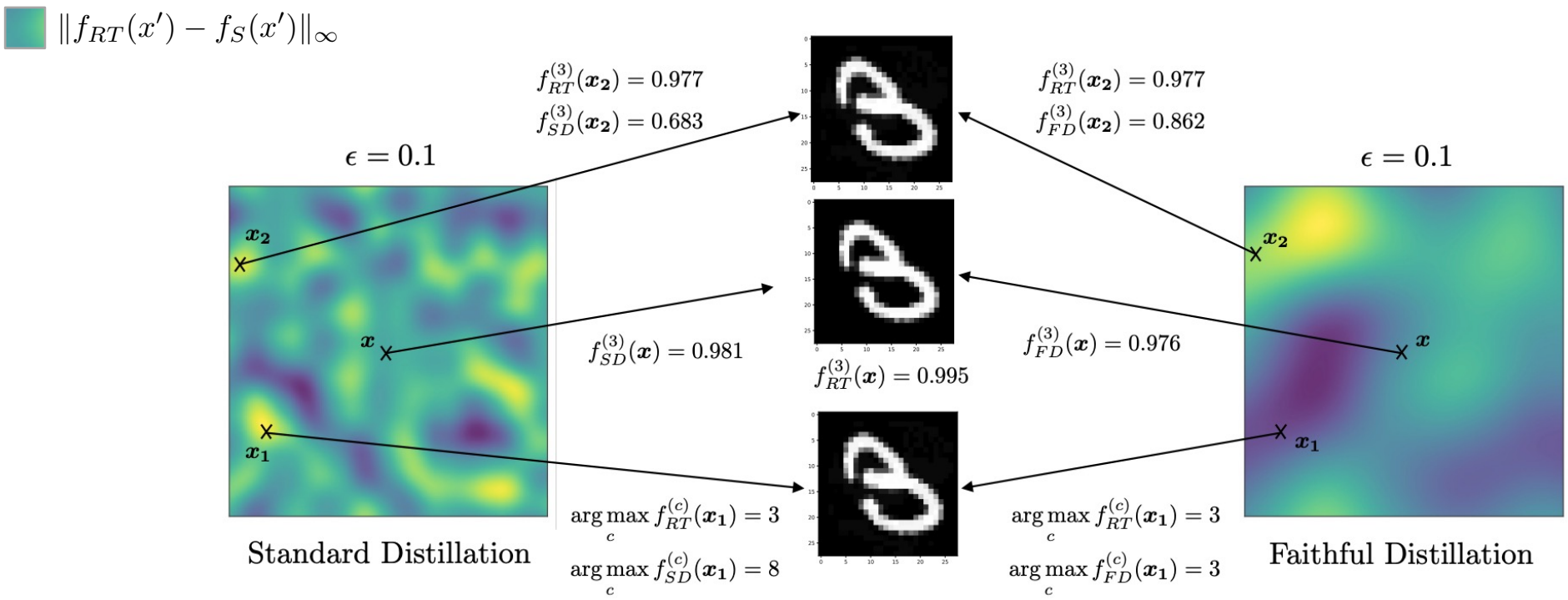
Raissi, Maziar, Paris Perdikaris, and George E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." Journal of Computational physics 378 (2019): 686-707.

Monaco, Simone, and Daniele Apiletti. "Training physics-informed neural networks: One learning to rule them all?." Results in Engineering 18 (2023): 101023.

Takamoto, Makoto, et al. "PDEBench: An extensive benchmark for scientific machine learning." Advances in Neural Information Processing Systems 35 (2022): 1596-1611.

Faithful Knowledge Distillation (FKD, 2023)

- *Can we be sure when we distil a teacher network that the student will match the **teacher's output and confidence**?*
- **Intuition**:



Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network." stat 1050 (2015): 9.

Faithful Knowledge Distillation (FKD, 2023)

UNIVERSITY OF OXFORD

Definition 2 (Faithful Imitator) *We say that* $\hat{f} : \mathbb{R}^n \to \mathbb{R}^d$ *is an* (ϵ, δ)*- faithful imitation of* $f : \mathbb{R}^n \to \mathbb{R}^d$ *around* $\boldsymbol{x}_0 \in \mathbb{R}^n$ *if:*

$$d_f(f(\boldsymbol{x}), \hat{f}(\boldsymbol{x})) \leq \delta, \quad \forall \boldsymbol{x} \in B_\epsilon(\boldsymbol{x}_0) \tag{7}$$

where $B_\epsilon(\boldsymbol{x}_0) = \{\boldsymbol{x}' \in \mathbb{R}^n \mid d_{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{x}') \leq \epsilon\}$, $d_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ *is a chosen metric function in the output space, and* $d_{\boldsymbol{x}} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ *is a metric function in the input space. We refer to any* δ *that bounds* $d_f(f(\boldsymbol{x}'), \hat{f}(\boldsymbol{x}'))$ *as a faithfulness bound for a given* ϵ*.*

- **Faithful Imitator for Knowledge Distillation**: *underestimate* δ empirically using a PGD-inspired method, *overestimate* it using an extension to CROWN.
- Compare standard distillation (SD), with other existing losses with similar goals (ARD, RSLAD) and our proposed faithful distillation loss (FD) which *empirically* minimizes KL divergence of teacher and student.

Table 2: *Faithfulness and relative calibration* - empirical lower bounds (EMPLB) and faithfulness upper bounds (FAITHUB) on MNIST, F-MNIST and CIFAR-10. *Lower is better.*

		EMPLB				FAITHUB			
	ϵ	f_{SD}	f_{ARD}	f_{RSLAD}	f_{FD}	f_{SD}	f_{ARD}	f_{RSLAD}	f_{FD}
MNIST	0.025	$0.042_{\pm 0.102}$	$0.045_{\pm 0.113}$	$0.039_{\pm 0.087}$	$\mathbf{0.033}_{\pm 0.079}$	$0.073_{\pm 0.150}$	$0.061_{\pm 0.143}$	$0.060_{\pm 0.118}$	$\mathbf{0.054}_{\pm 0.113}$
	0.05	$0.060_{\pm 0.129}$	$0.055_{\pm 0.130}$	$0.049_{\pm 0.101}$	$\mathbf{0.041}_{\pm 0.091}$	$0.179_{\pm 0.257}$	$0.097_{\pm 0.198}$	$0.101_{\pm 0.175}$	$\mathbf{0.094}_{\pm 0.171}$
	0.1	$0.106_{\pm 0.186}$	$0.078_{\pm 0.164}$	$0.072_{\pm 0.129}$	$\mathbf{0.061}_{\pm 0.117}$	$0.731_{\pm 0.343}$	$0.262_{\pm 0.337}$	$0.290_{\pm 0.319}$	$\mathbf{0.248}_{\pm 0.305}$
	0.15	$0.172_{\pm 0.246}$	$0.106_{\pm 0.197}$	$0.101_{\pm 0.160}$	$\mathbf{0.086}_{\pm 0.145}$	$0.974_{\pm 0.123}$	$0.636_{\pm 0.383}$	$0.668_{\pm 0.355}$	$\mathbf{0.588}_{\pm 0.367}$
	0.2	$0.258_{\pm 0.302}$	$0.140_{\pm 0.229}$	$0.138_{\pm 0.192}$	$\mathbf{0.118}_{\pm 0.174}$	$0.999_{\pm 0.017}$	$0.905_{\pm 0.230}$	$0.889_{\pm 0.229}$	$\mathbf{0.884}_{\pm 0.238}$
F-MNIST	4/255	$0.099_{\pm 0.159}$	$0.067_{\pm 0.112}$	$\mathbf{0.053}_{\pm 0.094}$	$0.060_{\pm 0.099}$	$0.215_{\pm 0.267}$	$0.128_{\pm 0.182}$	$\mathbf{0.118}_{\pm 0.170}$	$0.123_{\pm 0.171}$
	8/255	$0.155_{\pm 0.212}$	$0.096_{\pm 0.145}$	$\mathbf{0.081}_{\pm 0.130}$	$0.089_{\pm 0.131}$	$0.617_{\pm 0.393}$	$0.330_{\pm 0.342}$	$0.341_{\pm 0.341}$	$\mathbf{0.322}_{\pm 0.334}$
	12/255	$0.219_{\pm 0.265}$	$0.129_{\pm 0.179}$	$\mathbf{0.112}_{\pm 0.163}$	$0.122_{\pm 0.165}$	$0.892_{\pm 0.267}$	$0.678_{\pm 0.381}$	$0.690_{\pm 0.370}$	$\mathbf{0.668}_{\pm 0.380}$
	16/255	$0.297_{\pm 0.310}$	$0.165_{\pm 0.214}$	$\mathbf{0.148}_{\pm 0.197}$	$0.160_{\pm 0.199}$	$0.990_{\pm 0.071}$	$0.906_{\pm 0.235}$	$0.917_{\pm 0.214}$	$\mathbf{0.902}_{\pm 0.237}$
	20/255	$0.378_{\pm 0.343}$	$0.204_{\pm 0.243}$	$\mathbf{0.190}_{\pm 0.226}$	$0.202_{\pm 0.229}$	$0.999_{\pm 0.013}$	$0.986_{\pm 0.089}$	$0.988_{\pm 0.079}$	$\mathbf{0.985}_{\pm 0.090}$
CIFAR-10	4/255	$0.282_{\pm 0.145}$	$0.242_{\pm 0.139}$	$0.226_{\pm 0.126}$	$\mathbf{0.187}_{\pm 0.096}$	$0.671_{\pm 0.191}$	$0.546_{\pm 0.203}$	$0.497_{\pm 0.176}$	$\mathbf{0.474}_{\pm 0.182}$
	8/255	$0.388_{\pm 0.166}$	$0.317_{\pm 0.155}$	$0.291_{\pm 0.139}$	$\mathbf{0.245}_{\pm 0.114}$	$0.983_{\pm 0.033}$	$0.910_{\pm 0.127}$	$0.873_{\pm 0.133}$	$\mathbf{0.864}_{\pm 0.153}$
	12/255	$0.489_{\pm 0.175}$	$0.394_{\pm 0.169}$	$0.357_{\pm 0.148}$	$\mathbf{0.302}_{\pm 0.129}$	$1.000_{\pm 0.002}$	$1.000_{\pm 0.016}$	$0.990_{\pm 0.035}$	$\mathbf{0.989}_{\pm 0.038}$
	16/255	$0.583_{\pm 0.177}$	$0.462_{\pm 0.178}$	$0.415_{\pm 0.157}$	$\mathbf{0.359}_{\pm 0.142}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.002}$	$\mathbf{0.999}_{\pm 0.008}$	$\mathbf{0.999}_{\pm 0.009}$
	20/255	$0.666_{\pm 0.165}$	$0.530_{\pm 0.176}$	$0.474_{\pm 0.164}$	$\mathbf{0.416}_{\pm 0.151}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.001}$	$\mathbf{1.000}_{\pm 0.001}$

Goldblum, Micah, et al. "Adversarially robust distillation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.

Zi, Bojia, et al. "Revisiting adversarial robustness distillation: Robust soft labels make student better." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

UNIVERSITY OF OXFORD

01

Safety in Human Interaction/
Understanding Decision-Making Process

02

Correctness of ML-based Systems

UNIVERSITY OF OXFORD

01

Safety in Human Interaction/
Understanding Decision-Making Process

02

Correctness of ML-based Systems

UNIVERSITY OF OXFORD

01

Safety in Human Interaction/
Understanding Decision-Making Process

02

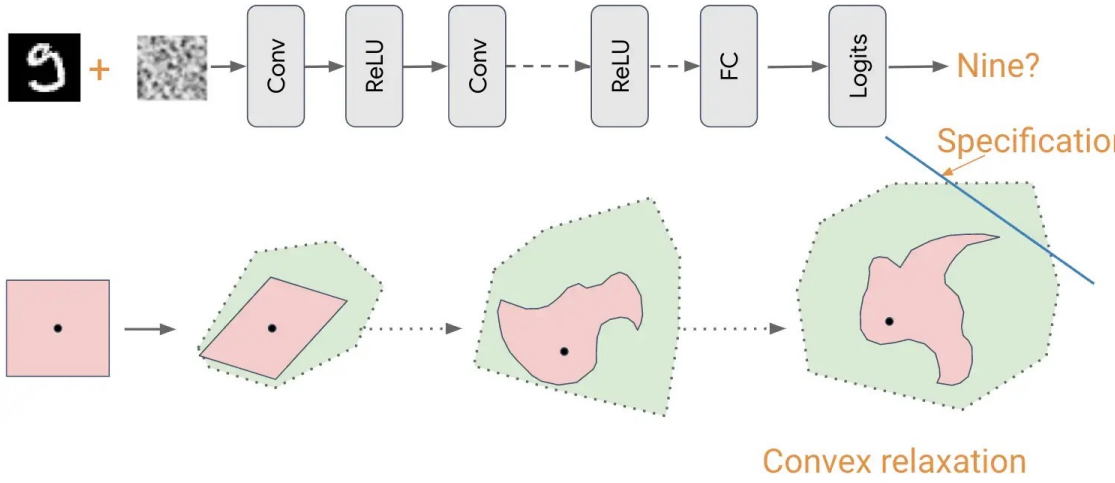
Correctness of ML-based Systems

Scalable Adversarial Robustness Verification/Certification

UNIVERSITY OF OXFORD

- Complete vs. incomplete methods:
  - *Can it certify a ResNet50? If YES → incomplete.*

- "Sound, but incomplete" methods:
  - Bound propagation (e.g., DiffAI, IBP)
  - Convex relaxation optimization-based methods (e.g. CROWN)
  - **Smoothing-based methods** (e.g. randomized smoothing)



Randomized Smoothing [Cohen et. al, 2019]

- Given a *base classifier* **f**, we can obtain a *smooth classifier*.

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\varepsilon) = c)$$
$$\text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

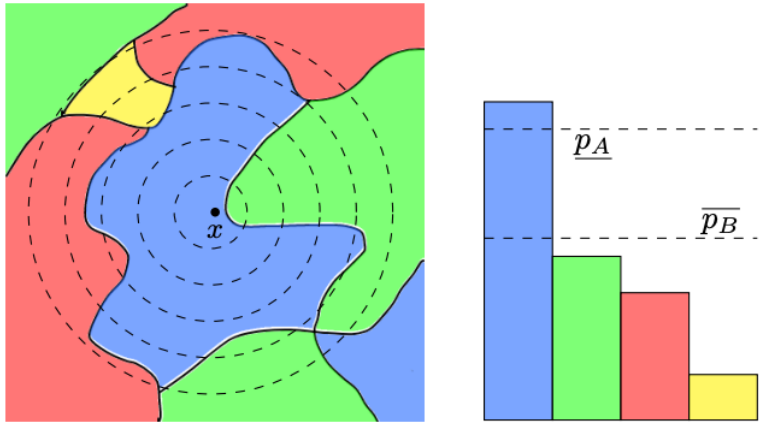
Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (2)$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (3)$$

Inverse Gaussian CDF



[Left] Decision boundaries of the base classifier **f** and level sets of the noise distribution [Right] output distribution given noise.

- Proof based on Neyman-Pearson lemma [Neyman & Pearson, 1933], as tight as possible for ℓ_2.
- Other works extend this to other ℓ_p norms [Yang et al., 2020].
- Can we **simplify** the analysis?

Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." international conference on machine learning. PMLR, 2019.
Neyman, Jerzy, and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231.694-706 (1933): 289-337.
Yang, Greg, et al. "Randomized smoothing of all shapes and sizes." International Conference on Machine Learning. PMLR, 2020.

A Lipschitz view of Randomized Smoothing (ANCER, 2022)

UNIVERSITY OF OXFORD

Proposition 1. *Consider a differentiable function* $g : \mathbb{R}^n \to \mathbb{R}$. *If* $sup_x \|\nabla g(x)\|_* \leq L$ *where* $\|\cdot\|_*$ *has a dual norm* $\|z\| = \max_x z^\top x \quad s.t. \ \|x\|_* \leq 1$, *then* g *is* L*-Lipschitz under norm* $\|\cdot\|_*$, *that is* $|g(x) - g(y)| \leq L\|x - y\|$.

Given the previous proposition, we formalize $\|\cdot\|$ certification as follows:

Theorem 1. *Let* $g : \mathbb{R}^n \to \mathbb{R}^K$, g^i *be* L*-Lipschitz continuous under norm* $\|\cdot\|_* \ \forall i \in \{1, \ldots, K\}$, *and* $c_A = arg\max_i g^i(x)$. *Then, we have* $arg\max_i g^i(x + \delta) = c_A$ *for all* δ *satisfying:*

$$\|\delta\| \leq \frac{1}{2L}\left(g^{c_A}(x) - \max_{c} g^{c \neq c_A}(x)\right).$$

- **Remark**: randomized smoothing is an instance of **Theorem 1** where the smooth classifier enjoys an analytical form for L by design.
  - Choose a smoothing distribution and compute an analytic Lipschitz constant under the dual norm for $g \rightarrow$ gives you a certificate under the norm by **Theorem 1**.
- Can recover the certificates from [Cohen et. al, 2019] for ℓ_2 and from [Yang et al., 2020] for several ℓ_p ones.

Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." international conference on machine learning. PMLR, 2019.
Yang, Greg, et al. "Randomized smoothing of all shapes and sizes." International Conference on Machine Learning. PMLR, 2020.

Anisotropic Certification (ANCER, 2022)

UNIVERSITY OF OXFORD

- Lipschitz analysis allows us to obtain **anisotropic** certificates
- **Ellipsoid certificates**: $g_\Sigma(x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\left[f(x+\epsilon)\right]$ $\quad\quad$ $\|\delta\|_{\Sigma,2} = \sqrt{\delta^\top\Sigma^{-1}\delta}$

Proposition 2. *$\Phi^{-1}(g_\Sigma(x))$ is 1-Lipschitz (i.e. $L=1$) under the $\|\cdot\|_{\Sigma^{-1},2}$ norm.*

Since Φ^{-1} is a strictly increasing function, by combining Proposition 2 with Theorem 1, we have:

Corollary 1. *Let $c_A = \arg\max_i g^i_\Sigma(x)$, then $\arg\max_i g^i_\Sigma(x+\delta) = c_A$ for all δ satisfying:*

$$\|\delta\|_{\Sigma,2} \leq \frac{1}{2}\left(\Phi^{-1}\left(g^{c_A}_\Sigma(x)\right) - \Phi^{-1}\left(\max_c g^{c\neq c_A}_\Sigma(x)\right)\right).$$

- **Generalized cross-polytope certificates**: $g_\Lambda(x) = \mathbb{E}_{\epsilon\sim\mathcal{U}[-1,1]^n}[f(x+\Lambda\epsilon)]$

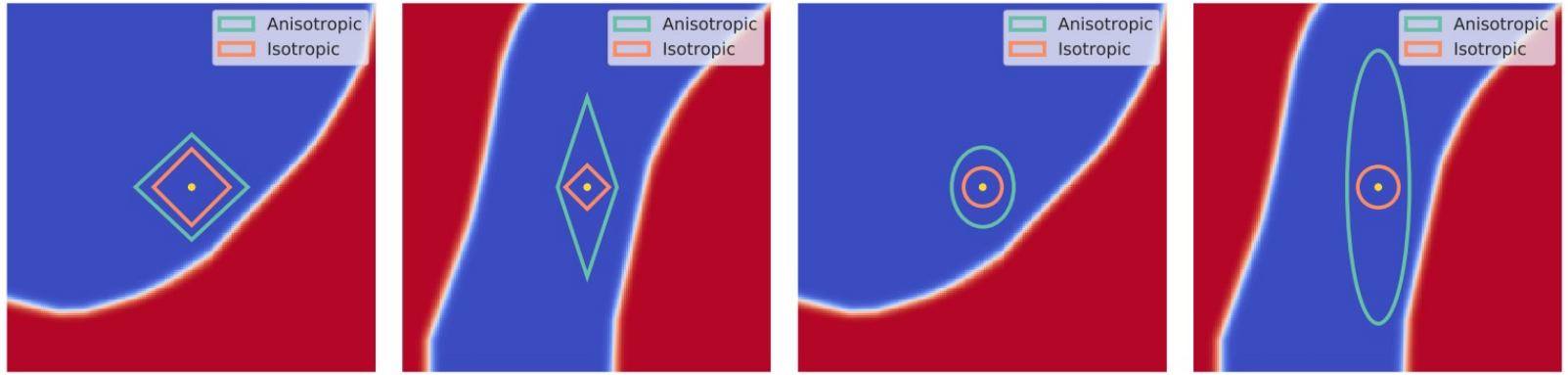
Proposition 3. *The classifier g_Λ is $1/2$-Lipschitz (i.e. $L=1/2$) under the $\|\Lambda x\|_\infty$ norm.*

Similar to Corollary 1, by combining Proposition 3 with Theorem 1, we have that:

Corollary 2. *Let $c_A = \arg\max_i g^i_\Lambda(x)$, then $\arg\max_i g^i_\Lambda(x+\delta) = c_A$ for all δ satisfying:*

$$\|\delta\|_{\Lambda,1} = \|\Lambda^{-1}\delta\|_1 \leq \left(g^{c_A}_\Lambda(x) - \max_c g^{c\neq c_A}_\Lambda(x)\right).$$

Anisotropic Certification (ANCER, 2022)



Examples of points in a toy 2D dataset, as well as potential isotropic and anisotropic certificates.

Visualization of natural CIFAR-10 images (top) and modified with an imperceptible change that is **not** inside the optimal isotropic certified region but **is** inside the optimal anisotropic one.

- Data-dependent certification requires memorization technique from [Alfarra et al., 2022], imposing a linear memory cost on the number of samples.

Alfarra, Motasem, et al. "Data dependent randomized smoothing." Uncertainty in Artificial Intelligence. PMLR, 2022.

Anisotropic Certification (ANCER, 2022)

UNIVERSITY OF OXFORD

Certificate volume

CIFAR-10	Certification	Accuracy @ ℓ_2 radius (%) 0.0	0.25	0.5	1.0	1.5	2.0	2.5	ℓ_2 ACR	ℓ_2^Σ $AC\tilde{R}$
COHEN Cohen et al. (2019)	Fixed σ	86	71	51	27	14	6	2	0.722	0.722
	Isotropic DD	82	76	62	39	24	14	8	1.117	1.117
	ANCER	86	85	77	53	31	17	10	**1.449**	**1.772**
SMOOTHADV Salman et al. (2019a)	Fixed σ	82	72	55	32	19	9	5	0.834	0.834
	Isotropic DD	82	75	63	40	25	15	7	1.011	1.011
	ANCER	83	81	73	48	30	17	8	**1.224**	**1.573**
MACER Zhai et al. (2019)	Fixed σ	87	76	59	37	24	14	9	0.970	0.970
	Isotropic DD	88	80	66	40	17	9	6	1.007	1.007
	ANCER	84	80	67	34	15	11	9	**1.136**	**1.481**

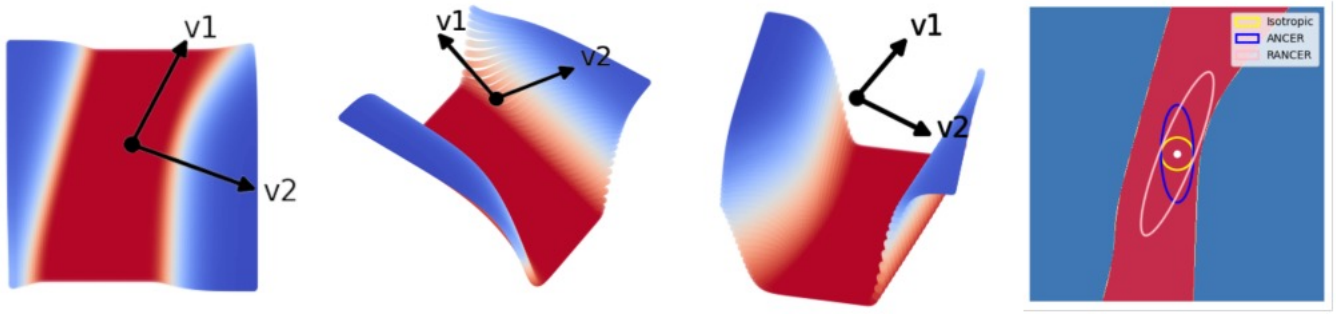
ImageNet	Certification	Accuracy @ ℓ_2 radius (%) 0.0	0.5	1.0	1.5	2.0	2.5	3.0	ℓ_2 ACR	ℓ_2^Σ $AC\tilde{R}$
COHEN Cohen et al. (2019)	Fixed σ	70	56	41	31	19	14	12	1.098	1.098
	Isotropic DD	71	59	46	36	24	19	15	1.234	1.234
	ANCER	70	70	62	61	42	36	29	**1.810**	**1.981**
SMOOTHADV Salman et al. (2019a)	Fixed σ	65	59	44	38	26	20	18	1.287	1.287
	Isotropic DD	66	62	53	41	32	24	20	1.428	1.428
	ANCER	66	66	62	58	44	37	32	**1.807**	**1.965**

CIFAR-10	Certification	Accuracy @ ℓ_1 radius (%) 0.0	0.25	0.5	0.75	1.0	1.5	2.0	ℓ_1 ACR	ℓ_1^Λ $AC\tilde{R}$
RS4A Yang et al. (2020)	Fixed σ	92	83	75	71	46	0	0	0.775	0.775
	Isotropic DD	92	89	82	76	58	6	2	0.946	0.946
	ANCER	92	90	84	80	63	6	2	**0.980**	**1.104**
ImageNet										
RS4A Yang et al. (2020)	Fixed σ	78	73	67	63	0	0	0	0.683	0.683
	Isotropic DD	79	76	70	65	46	0	0	0.729	0.729
	ANCER	78	76	70	66	48	0	0	**0.730**	**1.513**

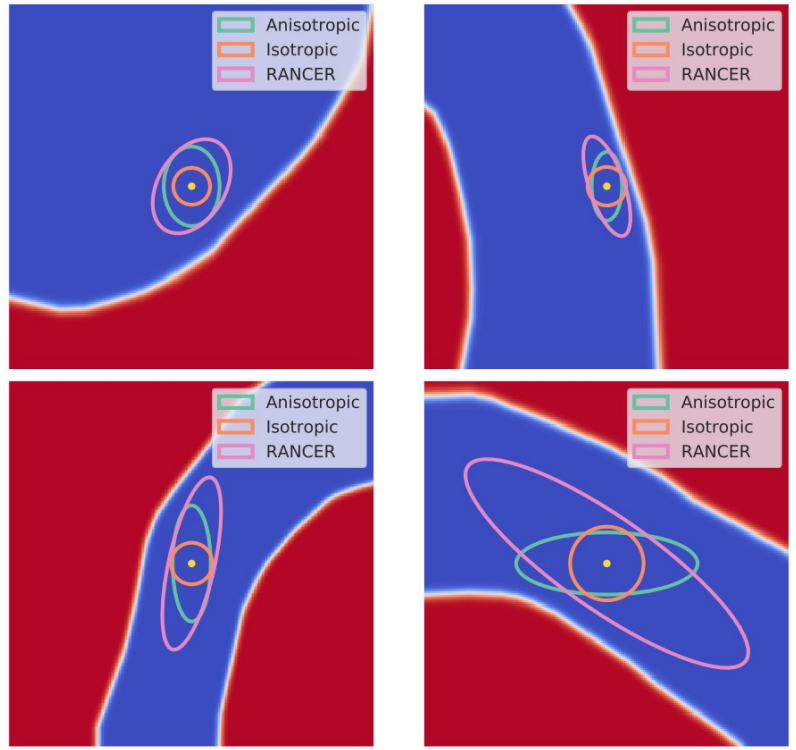
Comparison of top-1 certified accuracy, average certified radius and average certified proxy radius for different certification regimes.

RANCER: Non-Axis Aligned (RANCER, 2023)

- Anisotropic certification so far has been axis-aligned
- RANCER extends ANCER to optimize a full covariance matrix for Gaussian noise, resulting in a non-axis aligned certificate
- To achieve it, we use the eigen decomposition of the Hessian of the loss function at the certified point – corresponding to the "unsafe" directions of maximal change – which should be **smoothed more**
- Once rotated with that basis, optimized in the same way as (ANCER, 2022)



Safe and unsafe directions of the Hessian of the loss function at the point, eventually leading to the pink certified ellipse in the rightmost figure once region is optimized.



Examples of points in a toy 2D dataset.

Certifying Ensembles with $\mathcal{S}$-Lipschitzness (SEnsemb, 2023)

UNIVERSITY OF OXFORD

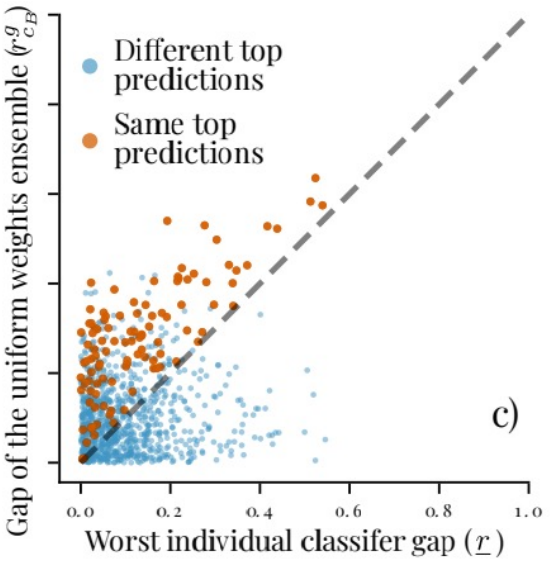
- A generalization of Lipschitzness to $\mathcal{S}$-Lipschitzness (avoiding symmetry requirements) enables tight analysis on the theoretical robustness of ensembles through $\mathcal{S}$-certificates.

Definition 2 ($\mathcal{S}$-Lipschitz function). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{S}$-Lipschitz for a bounded set $\mathcal{S} \subset \mathbb{R}^d$ if it holds that:*

$$-\rho_{\mathcal{S}}(x - y) \leq f(y) - f(x) \leq \rho_{\mathcal{S}}(y - x), \ \forall x, y \in \mathbb{R}^d,$$

with $\rho_{\mathcal{S}}(\delta) = \sup_{c \in \mathcal{S}} c^\top \delta$. If $\mathcal{S}$ is convex, then $\rho_{\mathcal{S}}$ corresponds to its support function.

- **Limitation of ensembling robust classifiers**:
  - One can gain robustness by ensembling... But the more robust the individual classifiers, the lower the possible improvement.
  - One can be worse off ensembling than using a single classifier... Even potentially losing all robustness (radius collapses to 0).
  - One will never be worse off if all classifiers predict the same class – but then why ensemble?
  - Usually end up losing more robustness than gaining → **ensembling robust classifiers reduces robustness**.



UNIVERSITY OF OXFORD

01

Safety in Human Interaction/
Understanding Decision-Making Process

02

Correctness of ML-based Systems

Certifiable ML – a network bounding problem

UNIVERSITY OF OXFORD

Safety/Robustness

(SEnsemb, 2023) Certifying Ensembles: A General Certification Theory with S-Lipschitzness
A Petrov, **F Eiras**, A Sanyal, PHS Torr, A Bibi (ICML)

(RANCER, 2023) RANCER: Non-Axis Aligned Anisotropic Certification With Randomized Smoothing
T Rumezhak, **F Eiras**, PHS Torr, A Bibi (WACV)

(ANCER, 2022) AnCer: Anisotropic certification via sample-wise volume maximization
F Eiras, M Alfarra, MP Kumar, PHS Torr, PK Dokania, B Ghanem, A Bibi (TMLR)

Reliability/Correctness

(PINN, 2023) Provably Correct Physics-Informed Neural Networks
F Eiras, A Bibi, R Bunel, KD Dvijotham, PHS Torr, MP Kumar (Pre-print)

(FKD, 2023) Faithful Knowledge Distillation
T Lamb, R Bunel, KD Dvijotham, PHS Torr, MP Kumar, **F Eiras** (Pre-print)

UNIVERSITY OF OXFORD

Thank you

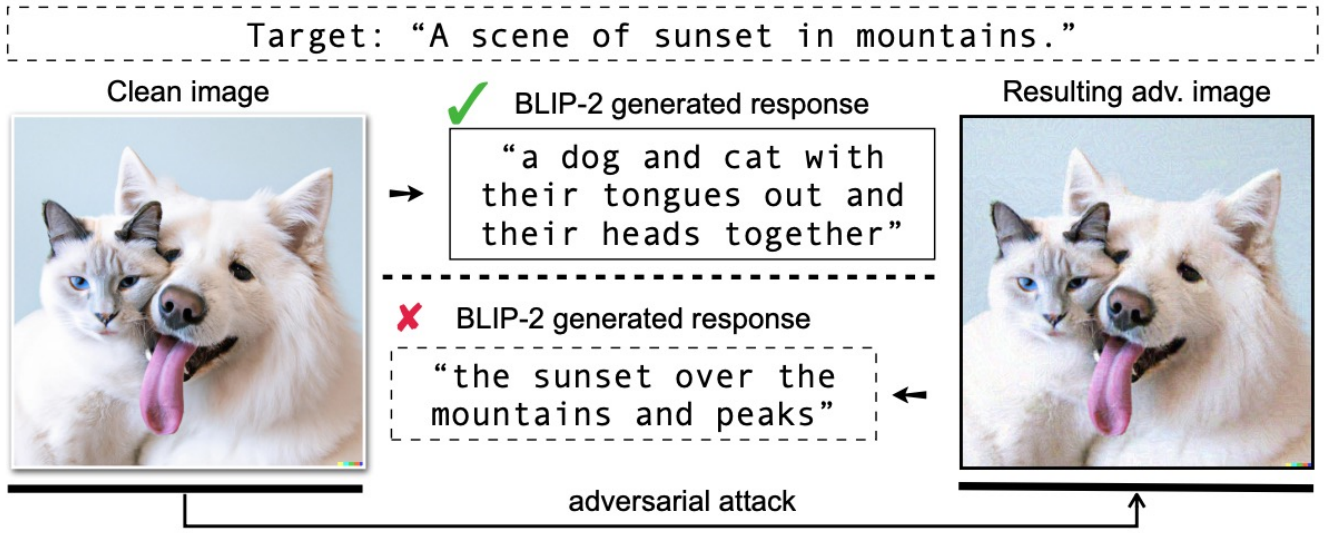
eiras@robots.ox.ac.uk

Twitter: @fgirbal

LinkedIn: franciscogirbaleiras

Adversarial Robustness of Vision-Language Models

UNIVERSITY OF OXFORD



A safety problem 😢

On Evaluating Adversarial Robustness of Large Vision-Language Models

Yunqing Zhao[*1], **Tianyu Pang**[*†2], **Chao Du**[†2], **Xiao Yang**[3], **Chongxuan Li**[4], **Ngai-Man Cheung**[†1], **Min Lin**[2]
[1]Singapore University of Technology and Design
[2]Sea AI Lab, Singapore
[3]Tsinghua University [4]Renmin University of China
yunqing_zhao@mymail.sutd.edu.sg, ngaiman_cheung@sutd.edu.sg
{tianyupang,duchao,linmin}@sea.com
yangxiao19@mails.tsinghua.edu.cn, chongxuanli@ruc.edu.cn

Abstract

Large vision-language models (VLMs) such as GPT-4 have achieved unprecedented performance in response generation, especially with visual inputs, enabling more creative and adaptable interaction than large language models such as ChatGPT. Nonetheless, multimodal generation exacerbates safety concerns, since adversaries may successfully evade the entire system by subtly manipulating the most vulnerable modality (e.g., vision). To this end, we propose evaluating the robustness of open-source large VLMs in the most realistic and high-risk setting, where adversaries have only *black-box* system access and seek to deceive the model into returning the *targeted* responses. In particular, we first craft targeted adversarial examples against pretrained models such as CLIP and BLIP, and then transfer these adversarial examples to other VLMs such as MiniGPT-4, LLaVA, UniDiffuser, BLIP-2, and Img2Prompt. In addition, we observe that black-box queries on these VLMs can further improve the effectiveness of targeted evasion, resulting in a surprisingly high success rate for generating targeted responses. Our findings provide a quantitative understanding regarding the adversarial vulnerability of large VLMs and call for a more thorough examination of their potential security flaws before deployment in practice. Code is at https://github.com/yunqing-me/AttackVLM.

1 Introduction

Large vision-language models (VLMs) have enjoyed tremendous success and demonstrated promising capabilities in text-to-image generation [52, 65, 69], image-grounded text generation (e.g., image captioning or visual question-answering) [2, 14, 40, 83], and joint generation [5, 30, 95] due to an increase in the amount of data, computational resources, and number of model parameters. Notably, after being finetuned with instructions and aligned with human feedback, GPT-4 [55] is capable of conversing with human users and, in particular, supports visual inputs.

From Zhao, Yunqing, et al. "On Evaluating Adversarial Robustness of Large Vision-Language Models." arXiv preprint arX