A rigorous, closed-form characterisation of adversarial generalisation errors.

A High Dimensional Statistical Model for Adversarial Training: Geometry and Trade-Offs

Problem Setup

Binary Classification Setting:

- Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{-1, +1\}$
- Probit model with noise parameter $\tau > 0$
- High-dimensional limit: $d, n \to \infty$ with fixed $\alpha = n/d$
- Structured data with block features: covariance matrices $\Sigma_{\boldsymbol{x}}, \Sigma_{\boldsymbol{\delta}}, \Sigma_{\boldsymbol{\upsilon}}, \Sigma_{\boldsymbol{\theta}}$ are block diagonal with k blocks of sizes $d_1, \ldots, d_k$

Metrics of Interest:

- Generalisation Error: $E_{\text{gen}} = \mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{1}(y \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}))\right]$
- Adversarial Generalisation Error: $E_{\text{adv}} = \mathbb{E}_{y,\boldsymbol{x}}\left[\max_{\|\boldsymbol{\delta}\|_{\Sigma_{\boldsymbol{\upsilon}}^{-1}} \leq \varepsilon_g} \mathbb{1}(y \neq \hat{y}(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{\delta}))\right]$
- Boundary Error: $E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}}$ where E_{bnd} are the attackable samples.
- Usefulness and Robustness:

$$\mathcal{U}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}[y \boldsymbol{\theta}_0^\top \boldsymbol{x}] \,, \tag{1}$$

$$\mathcal{R}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}\left[\inf_{\|\boldsymbol{\delta}\|_{\Sigma_{\boldsymbol{\upsilon}}^{-1}} \leq \varepsilon_g} y \boldsymbol{\theta}_0^\top (\boldsymbol{x} + \boldsymbol{\delta})\right] . \tag{2}$$

Adversarial ERM:

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - \varepsilon_t \frac{\sqrt{\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta}}}{\sqrt{d}}\right) + r(\boldsymbol{\theta}) \,. \tag{3}$$

Main Result

Theorem: Adversarial generalization errors are *provably* characterized by a system of 8 order parameters $(m, q, V, P, \hat{m}, \hat{q}, \hat{V}, \hat{P})$ and an additional parameter A through:

$$E_{\text{gen}} = \frac{1}{\pi} \arccos\left(m / \sqrt{(\rho + \tau^2) q}\right) , \tag{4}$$

$$E_{\text{bnd}} = \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \operatorname{erfc}\left(\frac{-\frac{m}{\sqrt{q}} \nu}{\sqrt{2(\rho + \tau^2 - m^2/q)}}\right) \frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\nu \,, \tag{5}$$

Implications

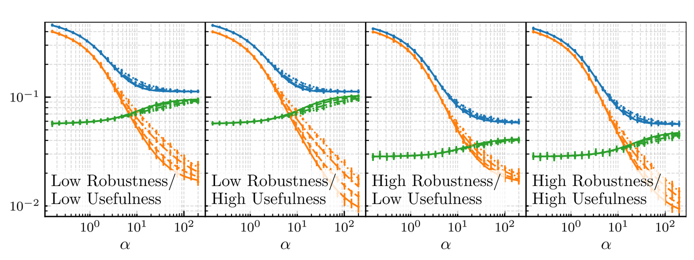
Trade-off between Usefulness and Robustness :

- Usefulness relates to generalisation error.
- Robustness relates to boundary error.
- Trade-off emerges when protecting useful but non-robust features.

Key Bounds:

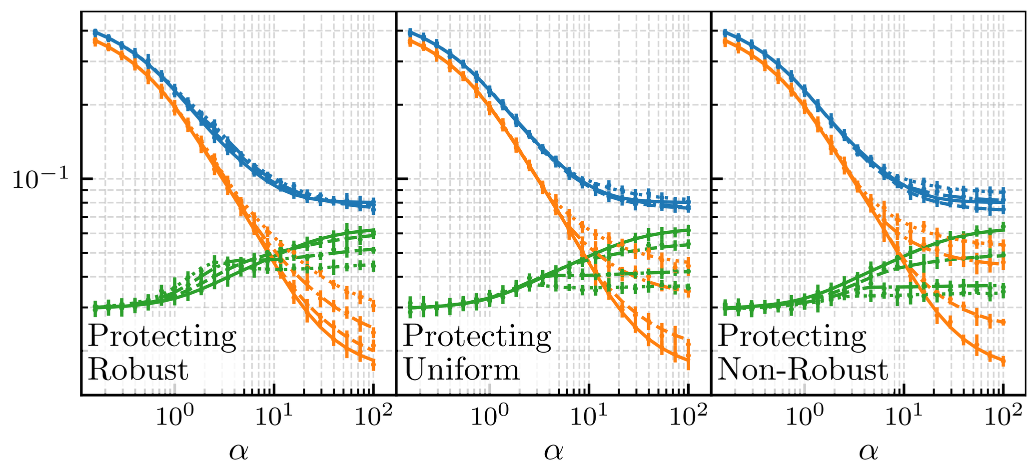
$$E_{\text{gen}} \geq \frac{1}{\pi} \arccos\left(\sqrt{\frac{\pi}{2\rho}} \mathcal{U}_{\boldsymbol{\theta}_0}\right) . \tag{6}$$

$$E_{\text{bnd}} \leq 2\mathrm{T}\left(\varepsilon_g \mathcal{A} \mathcal{B}, \mathcal{A}^{-1}\right) - \frac{1}{\pi} \arctan\left(\mathcal{A}^{-1}\right) - \frac{1}{\pi} \operatorname{erf}\left(\frac{\varepsilon_g \mathcal{B}}{\sqrt{2}}\right) \operatorname{erfc}\left(\frac{\varepsilon_g \mathcal{A} \mathcal{B}}{\sqrt{2}}\right) , \tag{7}$$



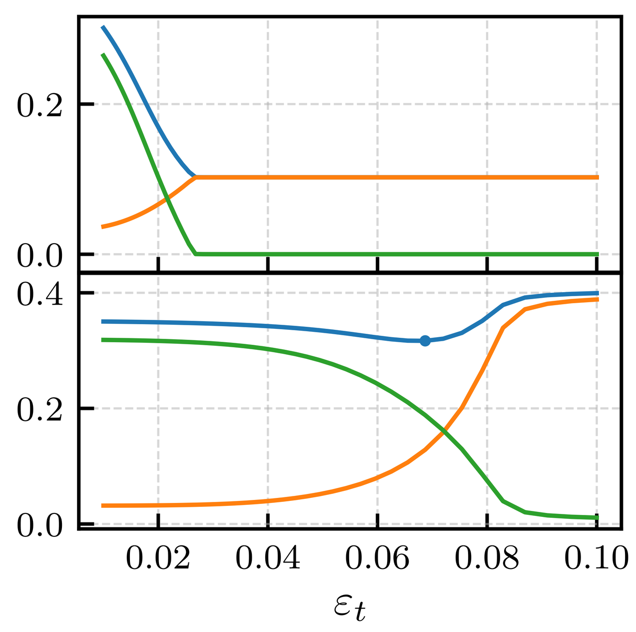
Directional Defences and structured data

Key Finding: The choice of defense strategy significantly impacts adversarial performance:



Impact of different defense strategies on generalization (E_{gen}) and boundary (E_{bnd}) errors

- **Defending robust features**: Low E_{gen} but high E_{bnd}
- **Uniform defense**: Better balance, improves overall E_{adv}
- **Defending non-robust features**: Increases E_{gen} while decreasing E_{bnd}



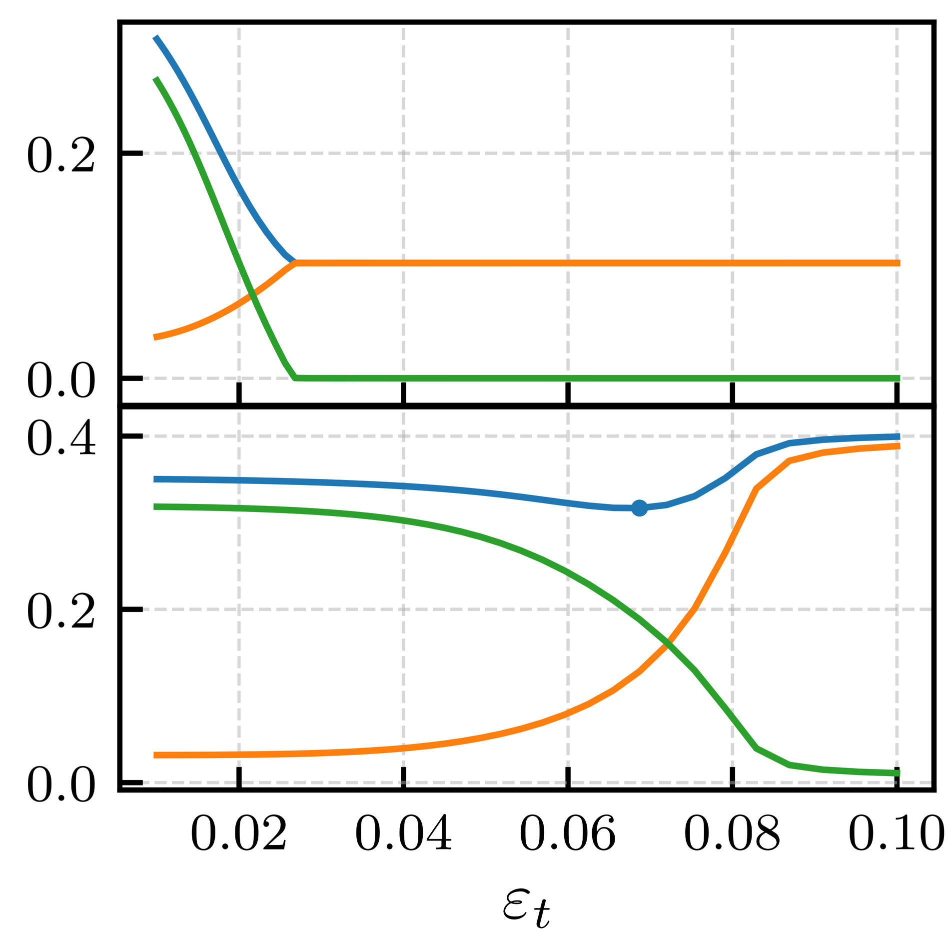
Optimal defense strategy depends on feature geometry

Analytical Result: For structured data with two feature blocks, we prove that protecting non-robust features:

- Always increases E_{gen} and decreases E_{bnd}
- Can improve E_{adv} when attack size is small enough

Tradeoff directions and innocuous directions

Key Insight: The geometry of features determines whether adversarial training leads to a trade-off:



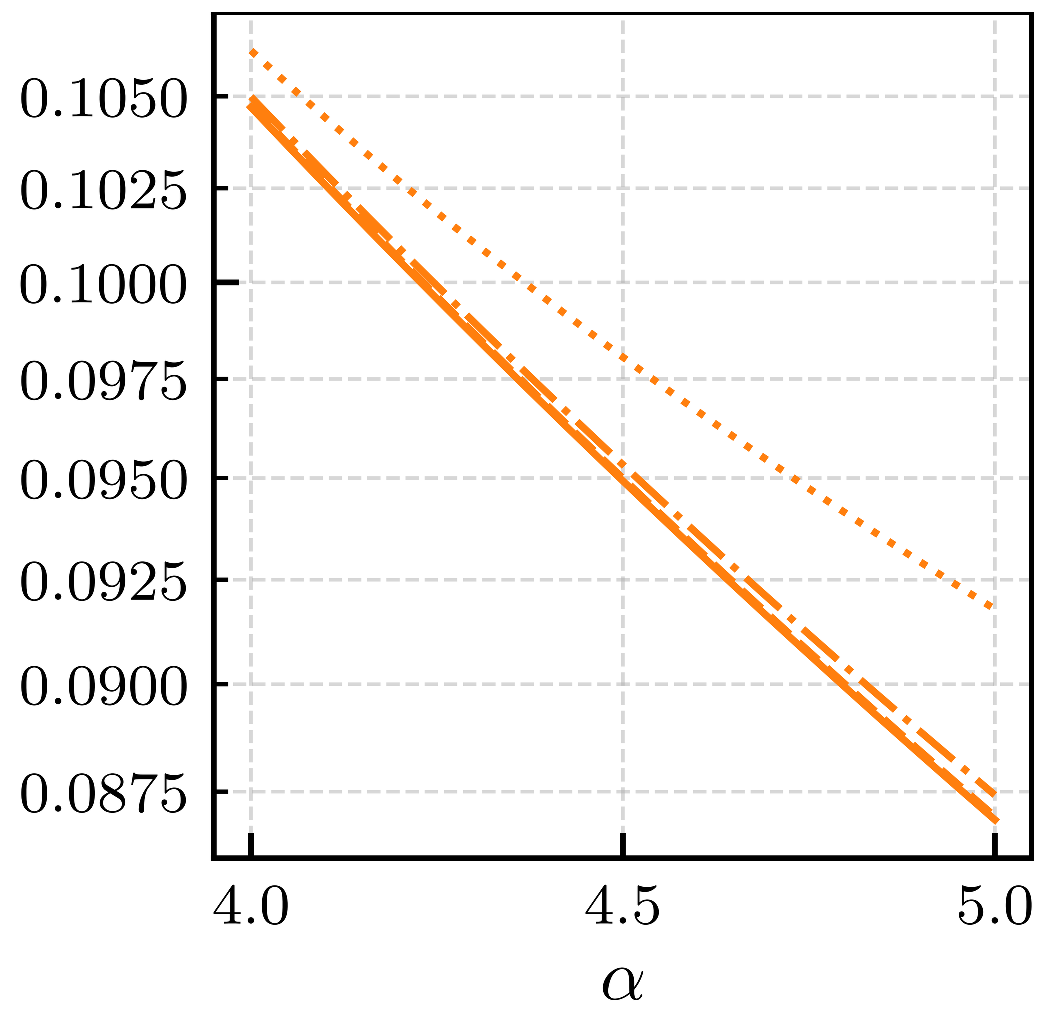
Impact of adversarial training on features with different geometries

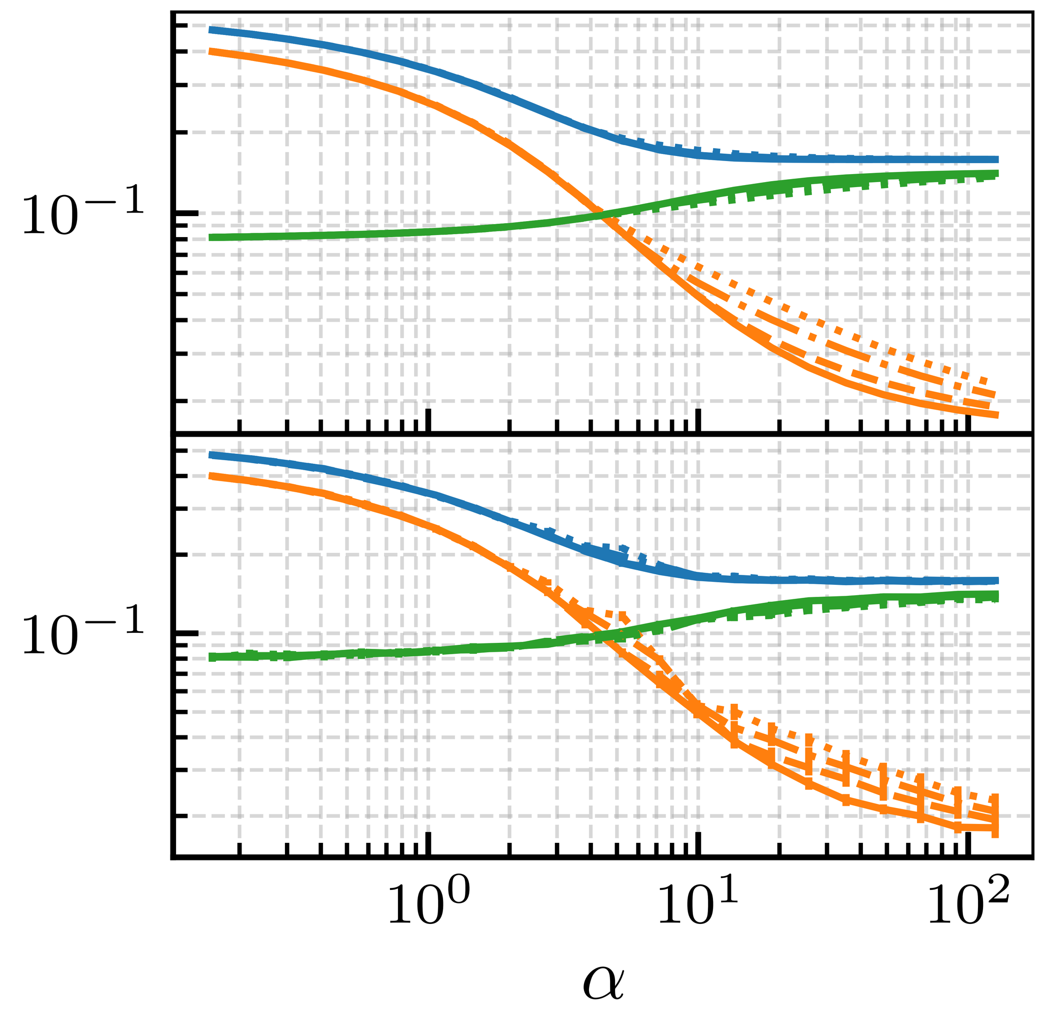
Two Distinct Cases:

- **Innocuous Features** (orthogonal to teacher):
  - Attack can be completely neutralized
  - $E_{\text{adv}} \to E_{\text{gen}}$ as ε_t increases
  - $E_{\text{bnd}} \to 0$ with sufficient training
- **Trade-off Features** (aligned with teacher):
  - Fundamental trade-off between E_{gen} and E_{bnd}
  - Optimal performance at specific ε_t
  - Requires careful hyperparameter tuning

Data Dependent Regularisation

Key Finding: Adversarial training can be approximated as a data-dependent regularisation:





Learning curves for adversarial training (top) and its regularisation approximation (bottom)

Approximate Loss:

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}}\right) + \tilde{\lambda}_1 \sqrt{\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta}} + \tilde{\lambda}_2 \boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta} \tag{8}$$

Key Properties:

- **Not just ℓ_2:** Performance depends on ε_t even with optimal λ
- **Effective Regularisation:** is a directional $\sqrt{\ell_2} + \ell_2$ regularisation.
- **Non-sparse:** $\sqrt{\ell_2}$ term provides linear scaling in the norm of the student vector without sparsity

Acknowledgements

Bruno Loureiro acknowledges support from the *Choose France - CNRS AI Rising Talents* program, and Florent Krzakala from the Swiss National Science Foundation grant SNFS OperaGOST (grant number 200390).

Kasimir Tanner
Matteo Vilucchio
Bruno Loureiro
Florent Krzakala

EPFL