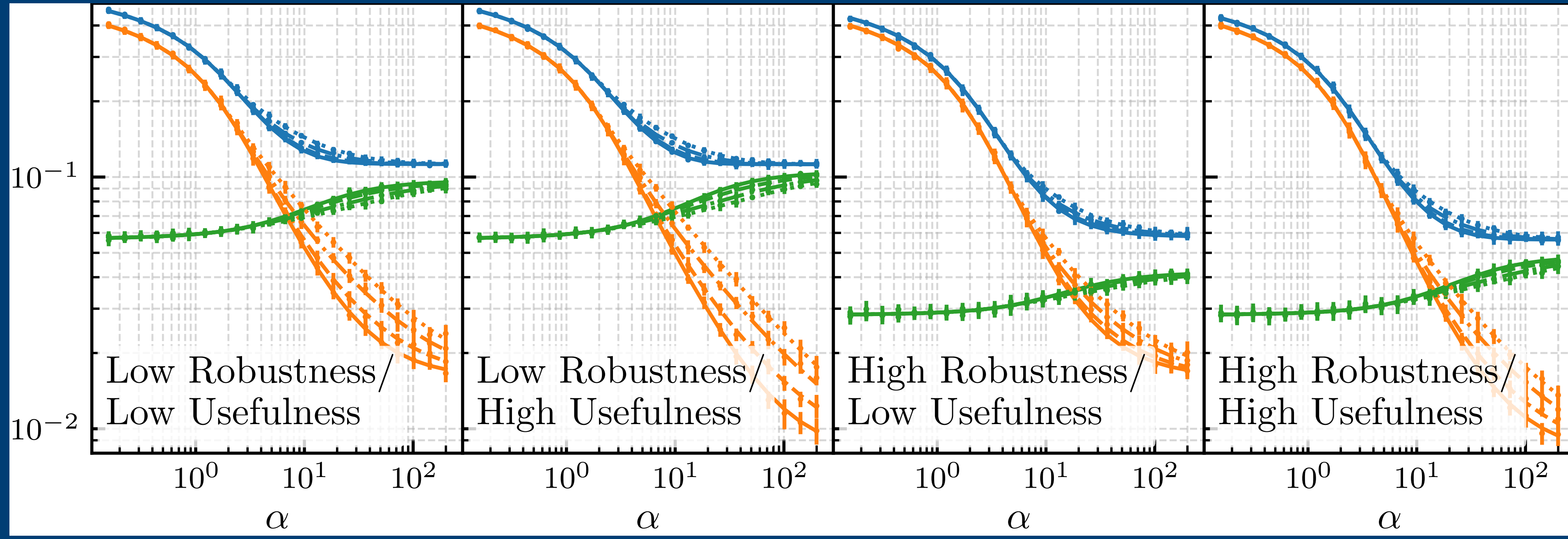


# A rigorous, closed-form characterisation of adversarial generalisation errors.



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

### Problem Setup

#### Binary Classification Setting:

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

#### Metrics of Interest:

- Generalisation Error:

$$E_{\text{gen}} = \mathbb{E}_{y,x} [\mathbb{1}(y \neq \hat{y}(\theta, x))] \quad (1)$$

- Adversarial Generalisation Error:

$$E_{\text{adv}} = \mathbb{E}_{y,x} \left[ \max_{\|\delta\|_{\Sigma_\delta^{-1}} \leq \epsilon_g} \mathbb{1}(y \neq \hat{y}(\theta, x + \delta)) \right] \quad (2)$$

- Boundary Error:

$$E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}} \quad (3)$$

where  $E_{\text{bnd}}$  are the attackable samples.

- Usefulness and Robustness:

$$\mathcal{U}_{\theta_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{x,y} [y \theta_0^\top x] \quad (4)$$

$$\mathcal{R}_{\theta_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{x,y} \left[ \inf_{\|\delta\|_{\Sigma_\delta^{-1}} \leq \epsilon_g} y \theta_0^\top (x + \delta) \right] \quad (5)$$

#### Adversarial ERM:

$$\sum_{i=1}^n g \left( y_i \frac{\theta^\top x_i}{\sqrt{d}} - \epsilon_t \frac{\sqrt{\theta^\top \Sigma_\delta \theta}}{\sqrt{d}} \right) + r(\theta) \quad (6)$$

### 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:

$$E_{\text{gen}} = \frac{1}{\pi} \arccos \left( m / \sqrt{(\rho + \tau^2)q} \right) \quad (7)$$

$$E_{\text{bnd}} = \int_0^{\epsilon_g \sqrt{A}} \text{erfc} \left( \frac{-\frac{m}{\sqrt{q}} \nu}{\sqrt{2(\rho + \tau^2 - m^2/q)}} \right) \frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}} d\nu \quad (8)$$

### Implications

#### Trade-off between Usefulness and Robustness :

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

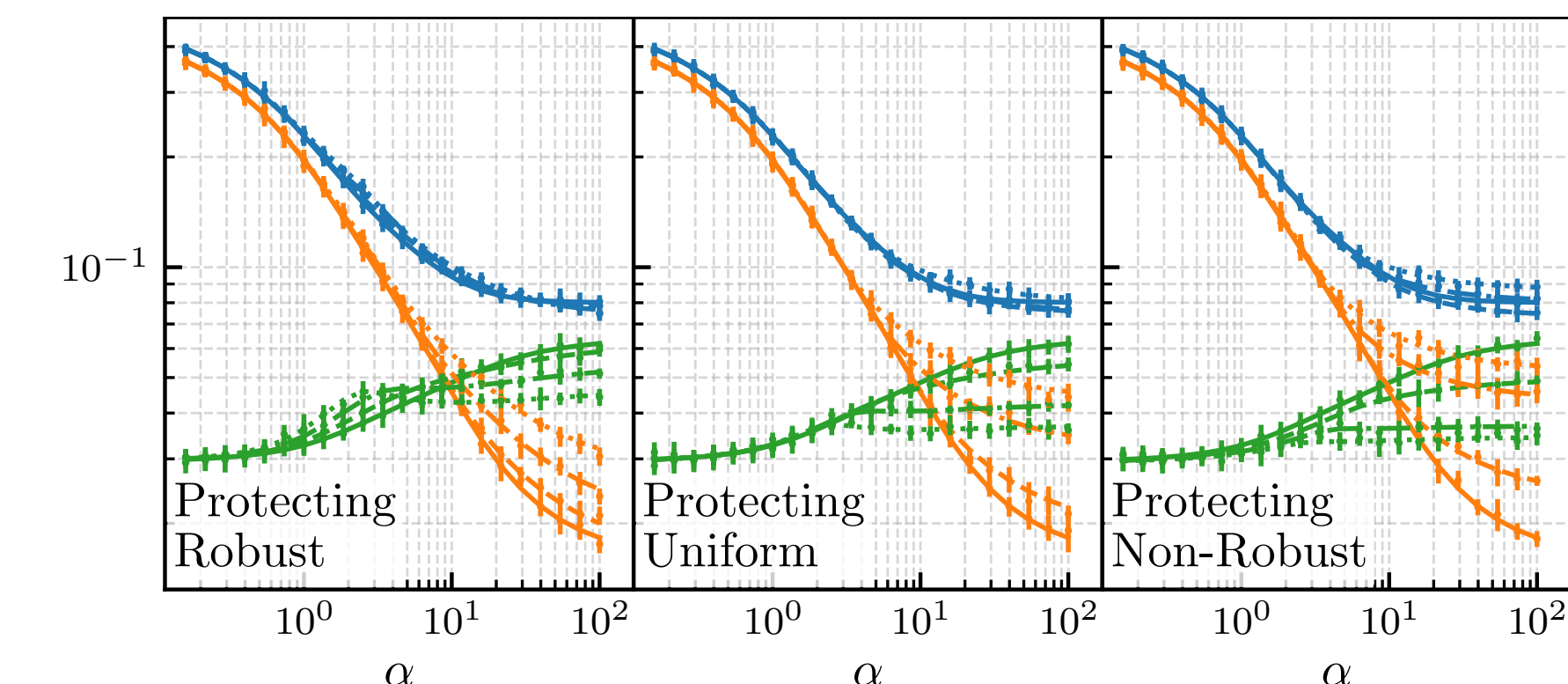
#### Key Bounds:

$$E_{\text{gen}} \geq \frac{1}{\pi} \arccos \left( \sqrt{\frac{\pi}{2\rho}} \mathcal{U}_{\theta_0} \right) \quad (9)$$

$$E_{\text{bnd}} \leq 2T(\epsilon_g A B, A^{-1}) - \frac{1}{\pi} \arctan(A^{-1}) - \frac{1}{\pi} \text{erf} \left( \frac{\epsilon_g B}{\sqrt{2}} \right) \text{erfc} \left( \frac{\epsilon_g A B}{\sqrt{2}} \right) \quad (10)$$

### Directional Defences and structured data

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



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

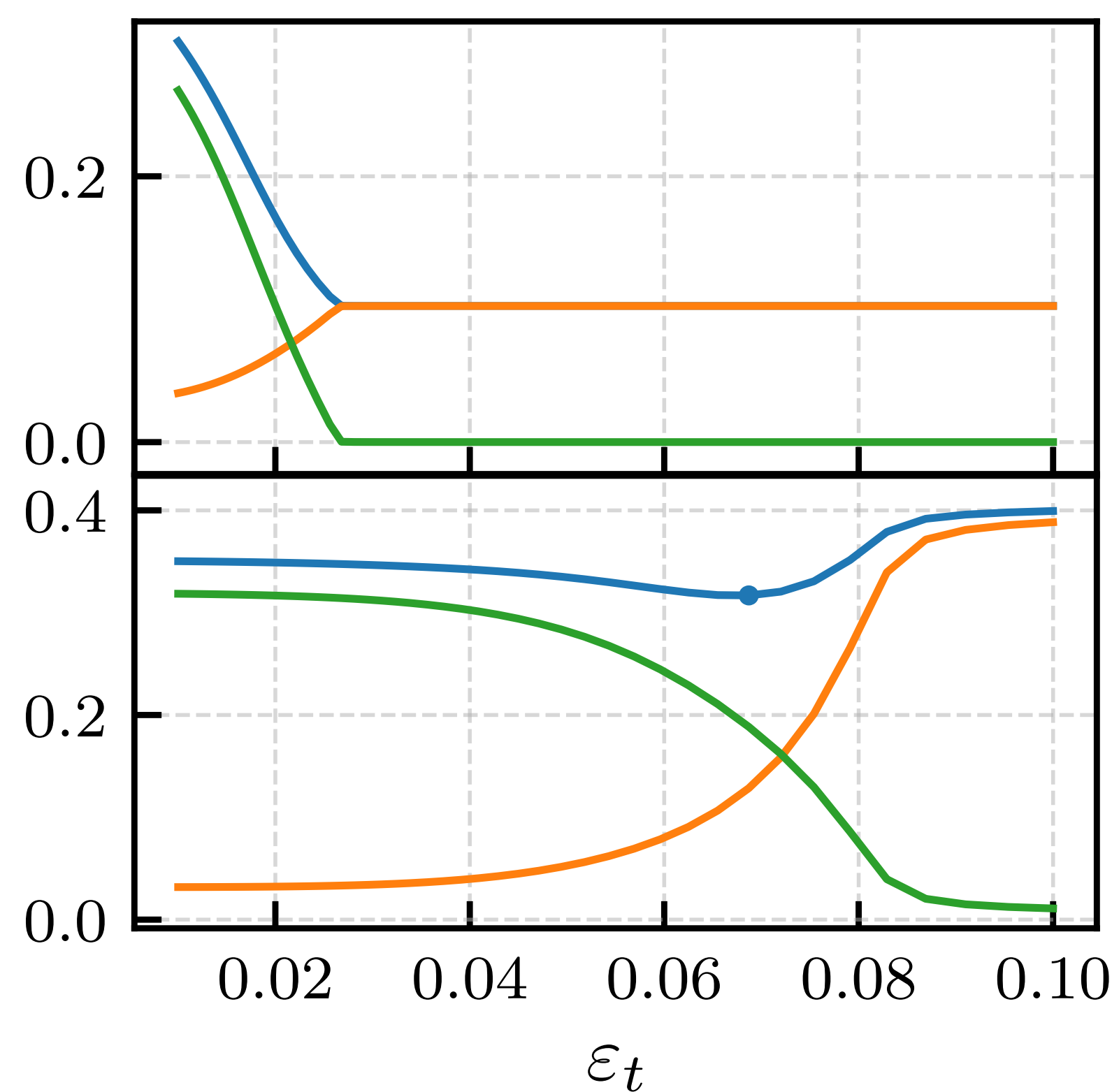
- **Defending robust features:** Low  $E_{\text{gen}}$  but high  $E_{\text{bnd}}$
- **Uniform defense:** Better balance, improves overall  $E_{\text{adv}}$
- **Defending non-robust features:** Increases  $E_{\text{gen}}$  while decreasing  $E_{\text{bnd}}$

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

- Always increases  $E_{\text{gen}}$  and decreases  $E_{\text{bnd}}$
- Can improve  $E_{\text{adv}}$  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:



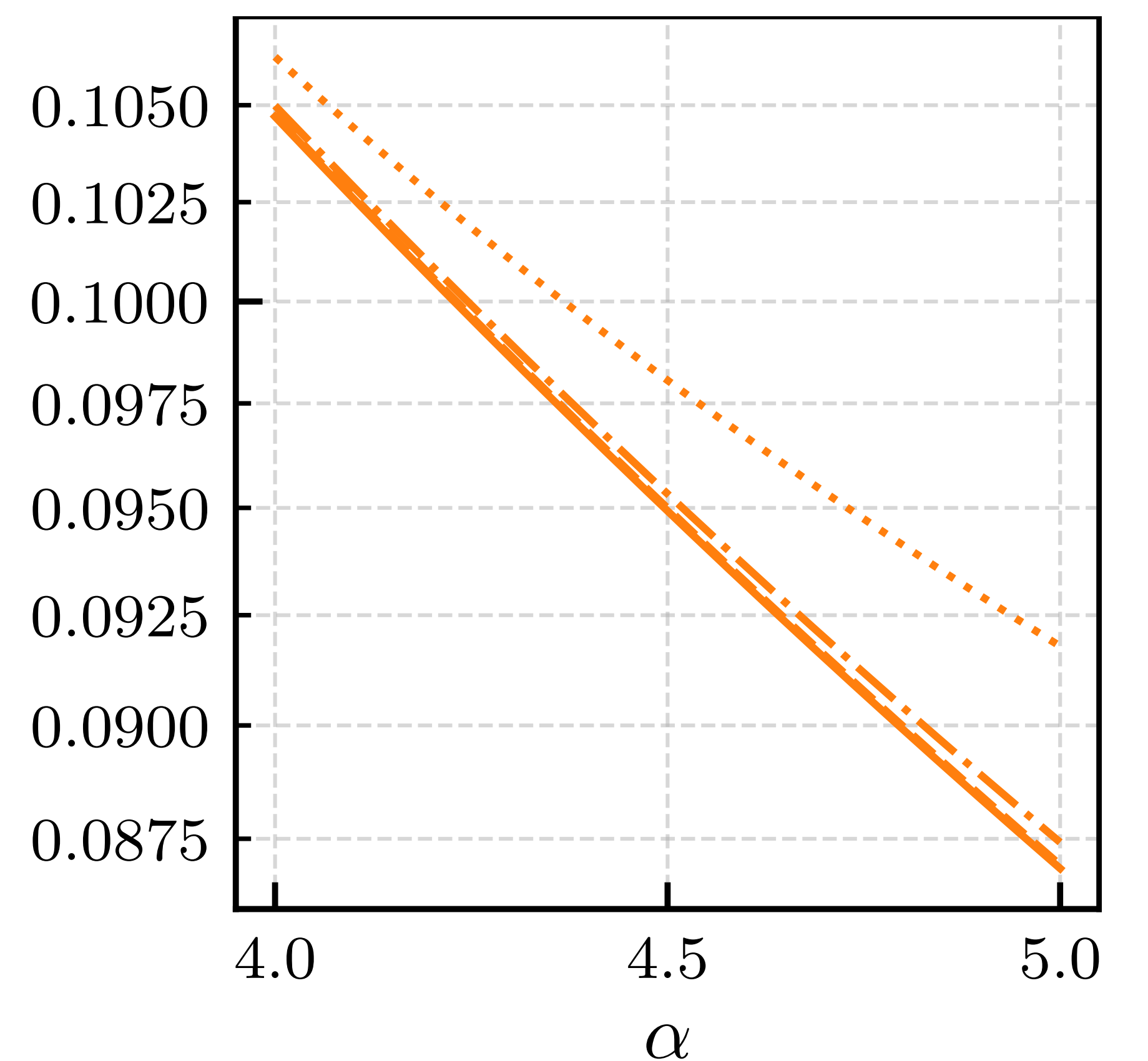
Impact of adversarial training on features with different geometries

#### Two Distinct Cases:

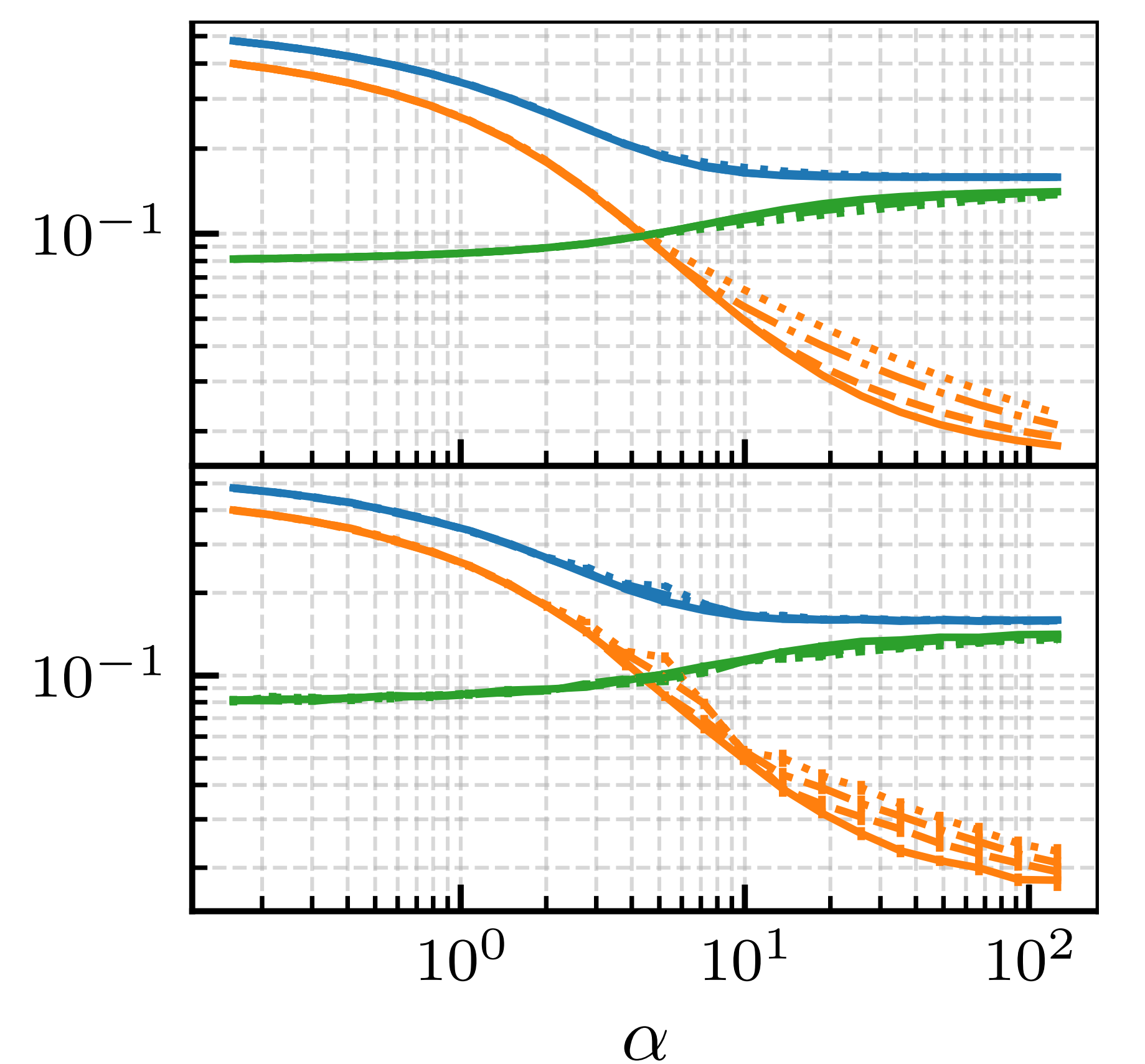
- **Innocuous Features** (orthogonal to teacher):
  - Attack can be completely neutralized
  - $E_{\text{adv}} \rightarrow E_{\text{gen}}$  as  $\epsilon_t$  increases
  - $E_{\text{bnd}} \rightarrow 0$  with sufficient training
- **Trade-off Features** (aligned with teacher):
  - Fundamental trade-off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

### Data Dependent Regularisation

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



Adversarial training is not just an  $\ell_2$  regularisation



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

#### Approximate Loss:

$$\sum_{i=1}^n g \left( y_i \frac{\theta^\top x_i}{\sqrt{d}} \right) + \tilde{\lambda}_1 \sqrt{\theta^\top \Sigma_\delta \theta} + \tilde{\lambda}_2 \theta^\top \Sigma_\delta \theta \quad (11)$$

#### Key Properties:

- **Not just  $\ell_2$ :** Performance depends on  $\epsilon_t$  even with optimal  $\lambda$
- **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