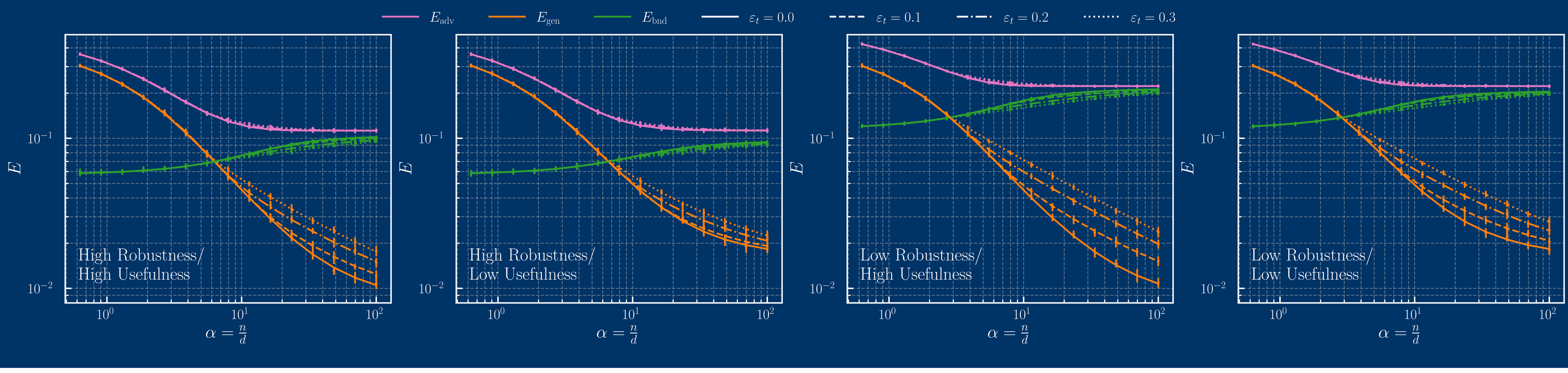
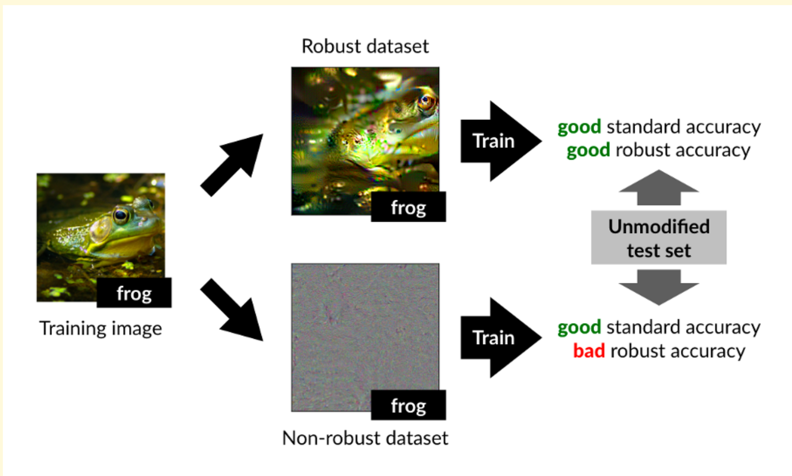
"Exact Asymptotics of **Adversarial Training**: Characterizing the Geometry of **Robustness-Accuracy Trade-Offs**"



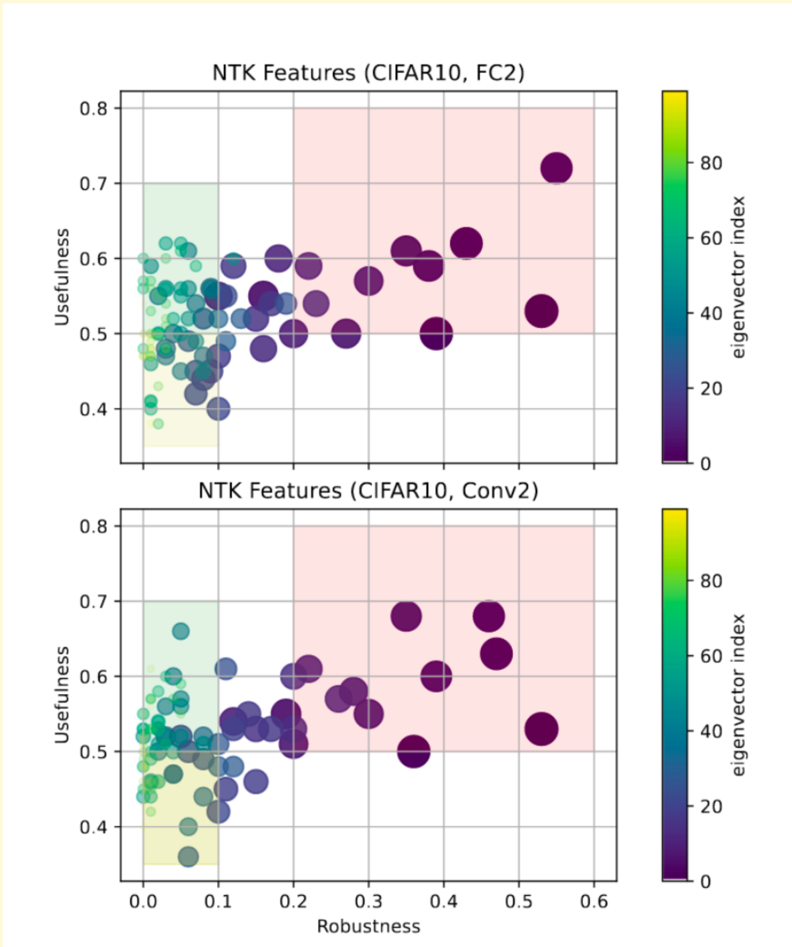
A High-dimensional Statistical Model for Adversarial Training: Geometry and Trade-Offs

Key Previous Results:

- Empirical findings: useful and robust features (Ilyas et al., 2019)
- Theory for the NTK regime (Tsilivis & Kempe, 2022)



Ilyas et al. (2019) - Disentangle features to robust and non-robust



Tsilivis & Kempe (2022) - Distinction of useful and robust features

Problem Setup (Binary Classification)

- Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{-1, +1\}$
- Student $\hat\omega$ and teacher ω^* vectors in $\mathbb{R}^d$
- High-dimensional limit: $d, n \to \infty$ with fixed $\alpha = n/d$
- Structured data with block features: covariance matrices $\Sigma_x, \Sigma_\delta, \Sigma_\upsilon, \Sigma_\theta$ are block diagonal with k blocks of sizes $d_1, \dots, d_k$

$$\sum_{i=1}^{n} \max_{\|\delta_i\|_{\Sigma_\delta^{-1}} \le \varepsilon_t} g\left(y_i \frac{\omega^\top (x_i + \delta_i)}{\sqrt{d}}\right) + r(\omega)$$

Metrics of Interest:

Generalisation Error:

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

Adversarial Generalisation Error:

$$E_{\mathrm{adv}} = \mathbb{E}_{y,x}[\max_{\|\delta\|_{\Sigma_\upsilon^{-1}} \le \varepsilon_g} \mathbb{1}(y \neq \hat y(\hat w, x + \delta))]$$

Boundary Error (attackable samples): $E_{\mathrm{adv}} = E_{\mathrm{gen}} + E_{\mathrm{bnd}}$

Here, usefulness and robustness

$$\mathcal{U}_{\omega^*} = \frac{1}{\sqrt d}\mathbb{E}_{x,y}[y\omega^{*\top} x] = \sqrt{\frac{2}{\pi}} \frac{\rho}{\sqrt{\rho + \tau^2}},$$
$$\mathcal{R}_{\omega^*} = \frac{1}{\sqrt d}\mathbb{E}_{x,y}[\inf_{\|\delta\|_{\Sigma_\upsilon^{-1}} \le \varepsilon_g} y\omega^{*\top}(x + \delta)] = \mathcal{U}_{\omega^*} - \frac{\varepsilon_g}{\sqrt d}\mathbb{E}[\sqrt{\omega^\top \Sigma_\upsilon \omega}].$$

Theoretical Results

Theorem: Adversarial generalization errors are 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: (see main figure)

$$E_{\mathrm{gen}} = \frac{1}{\pi}\arccos\left(\frac{m}{\sqrt{(\rho+\tau^2)q}}\right)$$
$$E_{\mathrm{bnd}} = \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \mathrm{erfc}\left(\frac{-\frac{m}{\sqrt q}\nu}{\sqrt{2(\rho + \tau^2 - m^2/q)}}\right) \frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}} d\nu$$

These quantities are interpretable as:

$$m = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{d} w^{*\top}\Sigma_x \hat w\right], q = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{d}\hat w^\top \Sigma_x \hat w\right], P = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{d}\hat w^\top \Sigma_\delta \hat w\right]$$

Trade-off between Usefulness and Robustness

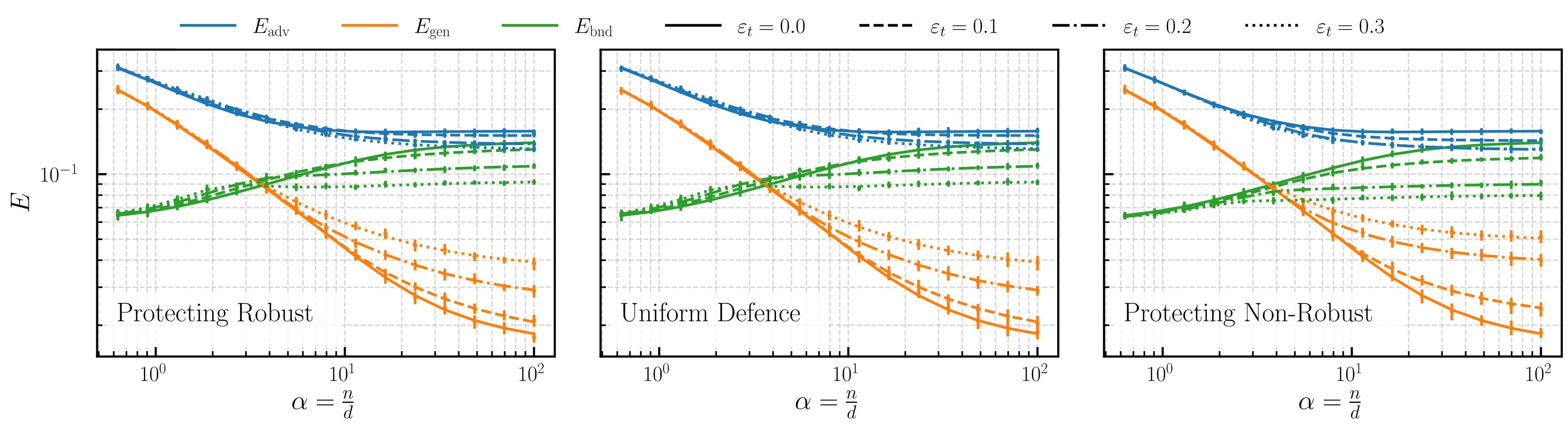
- Usefulness relates to generalisation error and robustness relates to boundary error
- Trade-off emerges when protecting useful but non-robust features

The errors scale inversly with $\alpha = \frac{n}{d}$

Proposition: Consider a single feature $k = 1$, $\forall \varepsilon_g, \varepsilon_t \ge 0$ for α big enough $\exists M_1, M_2$ such that:

$$|E_{\mathrm{adv}}(\varepsilon_g, \varepsilon_t) - E_{\mathrm{adv}}(\varepsilon_g, \varepsilon_t = 0)| < M_1/\alpha$$
$$|E_{\mathrm{gen}}(\varepsilon_t) - E_{\mathrm{gen}}(\varepsilon_t = 0)| < M_2/\alpha$$

Non-robust features can sometimes be protected without hurting accuracy



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

Proposition:

Assume two equally useful features, then defend the non-robust features more strongly.

Finding: Generalisation error increases linearly in defense strength and boundary error decreases linearly.

$$E_{\mathrm{bnd}}(\varrho) = E^0_{\mathrm{bnd}} + E^1_{\mathrm{bnd}}\varrho + \mathcal{O}(\varrho^2)$$
$$E_{\mathrm{gen}}(\varrho) = E^0_{\mathrm{gen}} + E^1_{\mathrm{gen}}\varrho + \mathcal{O}(\varrho^2)$$

where $E^1_{\mathrm{gen}} > 0$, $E^1_{\mathrm{bnd}} < 0$ and $E^0_{\mathrm{bnd}}, E^0_{\mathrm{gen}}$ errors at $\varrho = 0$

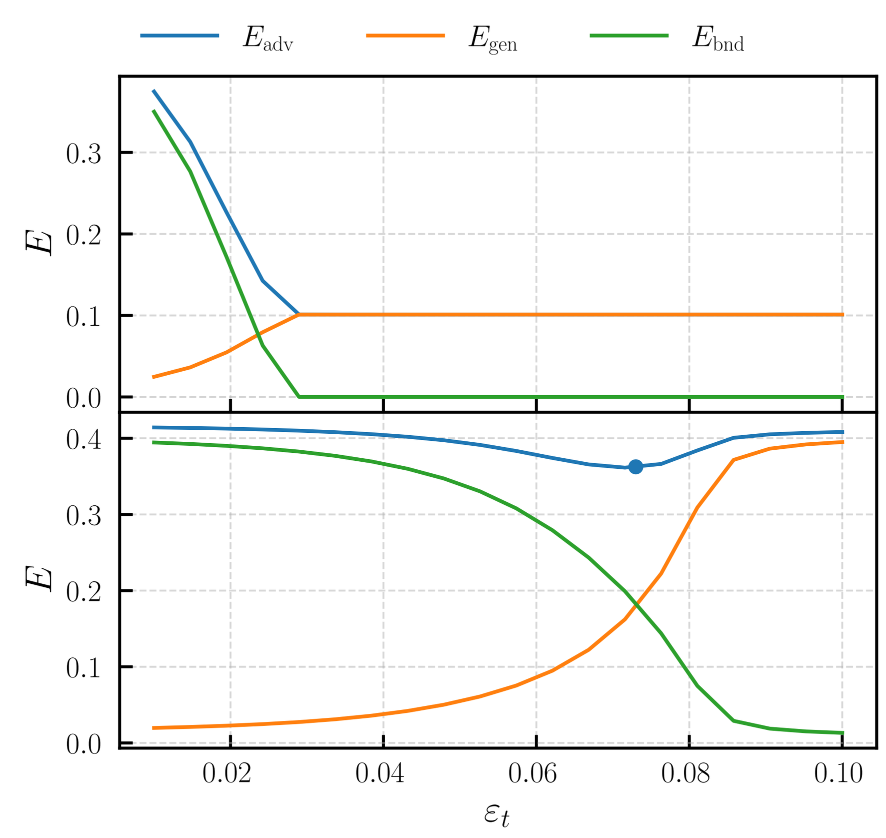
This improves E_{adv} at order ϱ iff

$$\frac{\varepsilon_g}{\sqrt 2}\mathrm{erfc}\left(-\frac{\vartheta_0 u_0 \varepsilon_g}{\sqrt{2 - 2\vartheta_0^2}}\right) < \frac{\exp\left(-\frac{\vartheta_0^2 u_0^2 \varepsilon_g^2}{2(1-\vartheta_0^2)}\right)}{\sqrt{\pi}\sqrt{1 - \vartheta_0^2}}$$

where $\vartheta_0 = m_0/\sqrt{\rho q_0}$ and $u_0 = \sqrt{A_0}/\sqrt{q_0}$ are the solution at $\varrho = 0$.

Defending the non-robust sub-space can improve robustness without hurting accuracy, especially under mild attacks

The attack geometry defines the adversarial trade-off



Impact of adversarial training on features with different geometries

Innocuous Features (orthogonal to teacher): Attack can be completely neutralized

Trade-off Features (aligned with teacher): Optimal performance at specific adversarial training cost

Kasimir Tanner
Matteo Vilucchio
Bruno Loureiro
Florent Krzakala

IdePHIX INFORMATION, LEARNING & PHYSICS LAB.
EPFL
ENS
PSL