Adversarial training protects the **non-robust** features. A **trade-off** emerges if those features are **useful**.

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

Empirical Misk Minimization

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

0.1 Block Features

$$\begin{aligned}
\Sigma_{\boldsymbol{x}} &= \mathrm{blockdiag}\left(\psi_1 \mathbb{1}_{d_1}, \ldots, \psi_k \mathbb{1}_{d_k}\right), \\
\Sigma_{\boldsymbol{\delta}} &= \mathrm{blockdiag}\left(\Delta_1 \mathbb{1}_{d_1}, \ldots, \Delta_k \mathbb{1}_{d_k}\right), \\
\Sigma_{\boldsymbol{\upsilon}} &= \mathrm{blockdiag}\left(\Upsilon_1 \mathbb{1}_{d_1}, \ldots, \Upsilon_k \mathbb{1}_{d_k}\right), \\
\Sigma_{\boldsymbol{\theta}} &= \mathrm{blockdiag}\left(t_1 \mathbb{1}_{d_1}, \ldots, t_k \mathbb{1}_{d_k}\right),
\end{aligned} \tag{2}$$

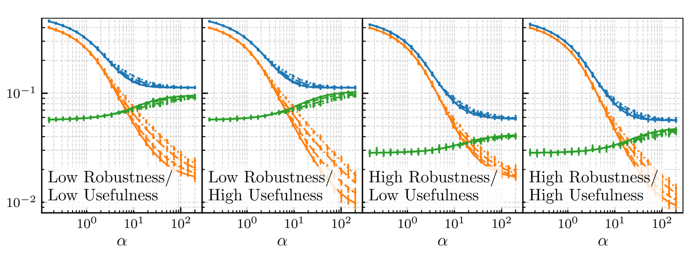
0.2 Usefulness and robustness

$$\mathcal{U}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}[y \boldsymbol{\theta}_0^\top \boldsymbol{x}] \,, \tag{3}$$

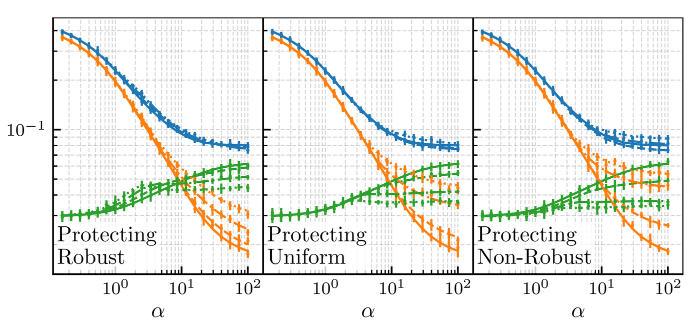
$$\mathcal{R}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{x},y}\left[\inf_{\|\boldsymbol{\delta}\|_{\Sigma_{\boldsymbol{\upsilon}}^{-1}} \le \varepsilon_g} y \boldsymbol{\theta}_0^\top (\boldsymbol{x} + \boldsymbol{\delta}) \right] . \tag{4}$$

Main Theorem

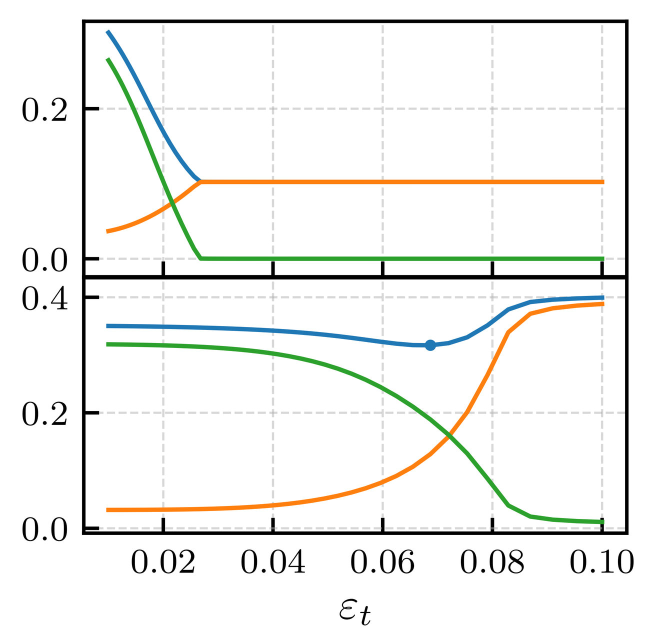
Trade-Offs



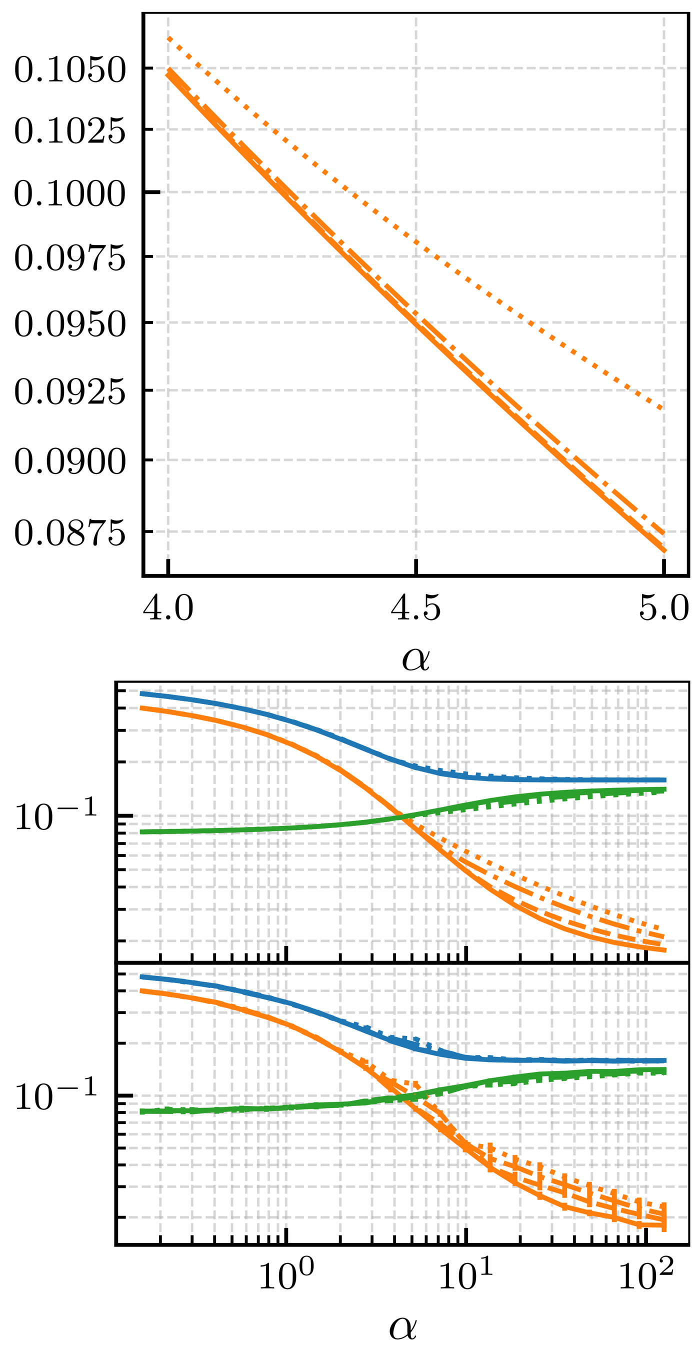
0.3 Directional Defences and structured data



0.4 Tradeoff directions and innocuous directions



0.5 Data Dependent Regularisation



Acknowledgements

Kasimir Tanner
Matteo Vilucchio
Bruno Loureiro
Florent Krzakala