

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 Risk Minimization

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

Block Features

$$\begin{aligned} \Sigma_x &= \text{blockdiag}(\psi_1 \mathbb{1}_{d_1}, \dots, \psi_k \mathbb{1}_{d_k}), \\ \Sigma_\delta &= \text{blockdiag}(\Delta_1 \mathbb{1}_{d_1}, \dots, \Delta_k \mathbb{1}_{d_k}), \\ \Sigma_v &= \text{blockdiag}(\Upsilon_1 \mathbb{1}_{d_1}, \dots, \Upsilon_k \mathbb{1}_{d_k}), \\ \Sigma_\theta &= \text{blockdiag}(t_1 \mathbb{1}_{d_1}, \dots, t_k \mathbb{1}_{d_k}), \end{aligned} \quad (2)$$

Usefulness and robustness

$$\mathcal{U}_{\theta_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{x,y} [y \theta_0^\top x], \quad (3)$$

$$\mathcal{R}_{\theta_0} = \frac{1}{\sqrt{d}} \mathbb{E}_{x,y} \left[\inf_{\|\delta\|_{\Sigma_v^{-1}} \leq \varepsilon_g} y \theta_0^\top (x + \delta) \right]. \quad (4)$$

Main Theorem For the ERM estimator of the risk function with ℓ_2 regularisation $r(\theta) = \frac{\lambda}{2} \|\theta\|_2^2$ and $\lambda \geq 0$, under the data model defined in ?? and in the high dimensional proportional limit, the generalisation error E_{gen} and the boundary error E_{bnd} concentrate to

$$E_{\text{gen}} = \frac{1}{\pi} \arccos\left(m / \sqrt{(\rho + \tau^2)q}\right), \quad (5)$$

$$E_{\text{bnd}} = \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \text{erfc}\left(\frac{-\frac{m}{\sqrt{q}} \nu}{\sqrt{2(\rho + \tau^2 - m^2/q)}}\right) \frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}} d\nu, \quad (6)$$

and the adversarial generalisation error concentrates to $E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}}$.

The values of m and q are the solutions of a system of eight self-consistent equations for the unknowns $(m, q, V, P, \hat{m}, \hat{q}, \hat{V}, \hat{P})$. The first four equations are dependant on the loss function g

and the adversarial training strength ε_t and read

$$\begin{cases} \hat{m} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \partial_\omega \mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P) \right] \\ \hat{q} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \mathcal{Z}_0 f_g^2(y, \sqrt{q}\xi, P) \right] \\ \hat{V} = -\alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \mathcal{Z}_0 \partial_\omega f_g(y, \sqrt{q}\xi, P) \right] \\ \hat{P} = -\frac{\varepsilon_t}{2\sqrt{P}} \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy y \mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P) \right] \end{cases}, \quad (7)$$

where $\xi \sim \mathcal{N}(0, 1)$ and $\mathcal{Z}_0 = 1/2 \text{erfc}(-y\omega/\sqrt{2(V+\tau^2)})$ and $f_g(y, \omega, V, P) = (\mathcal{P}(\omega) - \omega)/V$, where \mathcal{P} is the following proximal operator

$$\mathcal{P}(\omega) = \min_x \left[\frac{(x - \omega)^2}{2V} + g(yx - \varepsilon_t \sqrt{P}) \right]. \quad (8)$$

The second set of equation depend on the spectral distribution of the matrices Σ_x, Σ_δ and on the limiting distribution of the elements of $\bar{\theta}$. The equations read

$$\begin{cases} m = \mathbb{E}_\mu \left[\frac{\hat{m} \bar{\theta}^2}{\lambda + \hat{V} \omega + \hat{P} \delta} \right] \\ q = \mathbb{E}_\mu \left[\frac{\hat{m}^2 \bar{\theta}^2 \omega + \hat{q} \omega^2}{(\lambda + \hat{V} \omega + \hat{P} \delta)^2} \right] \\ V = \mathbb{E}_\mu \left[\frac{\omega}{\lambda + \hat{V} \omega + \hat{P} \delta} \right] \\ P = \mathbb{E}_\mu \left[\zeta \frac{\hat{m}^2 \bar{\theta}^2 + \hat{q} \omega^2}{(\lambda + \hat{V} \omega + \hat{P} \delta)^2} \right] \end{cases}. \quad (9)$$

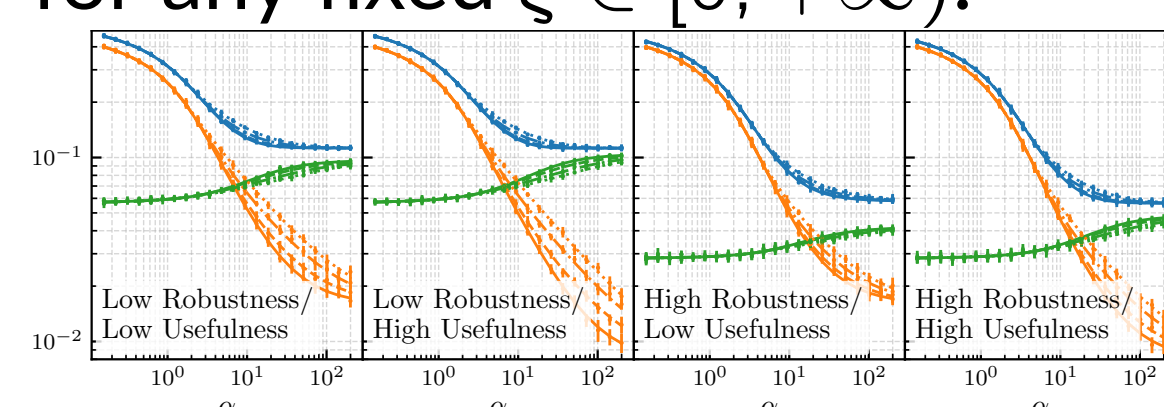
The value of A can be obtained from the solution of the same system of self consistent equations as

$$A = \mathbb{E}_\mu \left[v \frac{\hat{m}^2 \bar{\theta}^2 \omega + \hat{q} \omega^2}{(\lambda + \hat{V} \omega + \hat{P} \delta)^2} \right]. \quad (10)$$

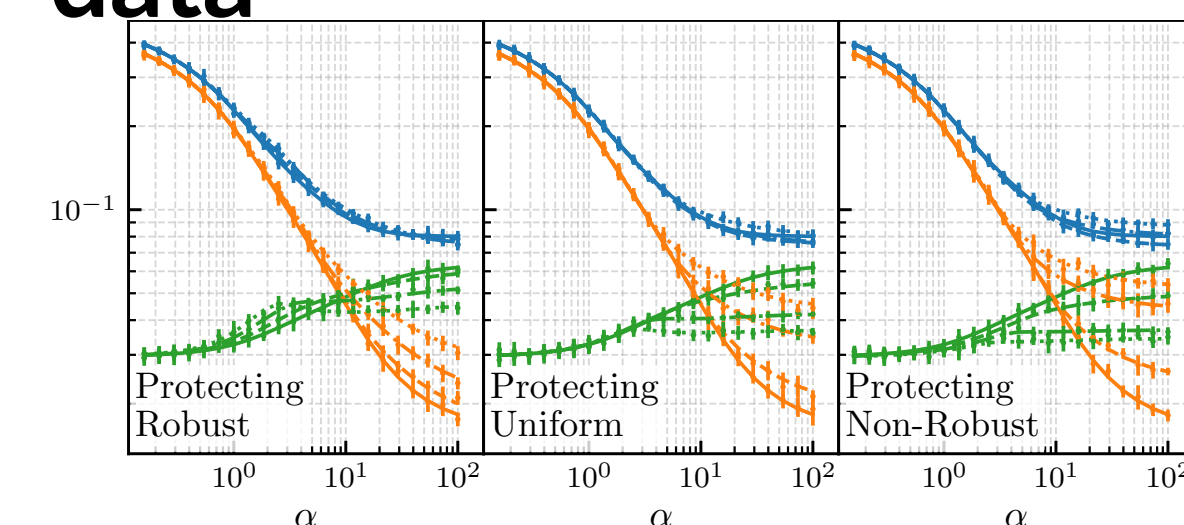
Trade-Offs

$$E_{\text{adv}} = E_{\text{gen}}(\vartheta, \mathcal{U}_{\theta_0}) + \int_0^{\varepsilon_g \kappa} f(\xi; \vartheta, \mathcal{U}_{\theta_0}) d\xi, \quad (11)$$

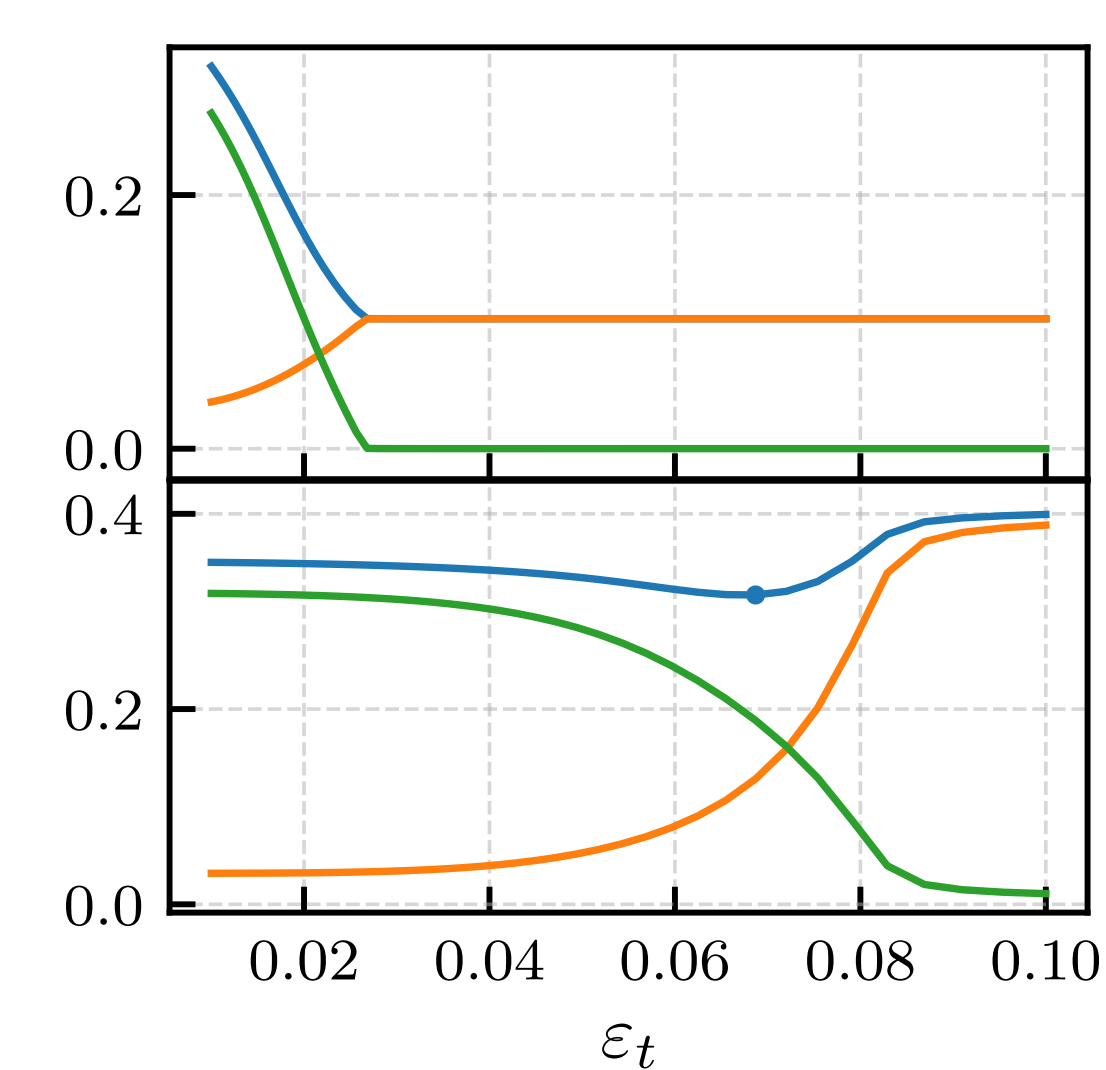
where we introduce the variable $\vartheta = m/\sqrt{\rho q}$ and $\kappa = \sqrt{A}/\sqrt{q}$. ϑ is the cosine of the angle between the teacher weights θ_0 and the student estimate $\hat{\theta}$ in the geometry of Σ_x and κ is the norm of $\hat{\theta}$ under the attack matrix. The function $f(\xi; \vartheta)$ is positive $\forall \vartheta, \forall \xi \in [0, +\infty)$ and it is strictly increasing in ϑ for any fixed $\xi \in [0, +\infty)$.



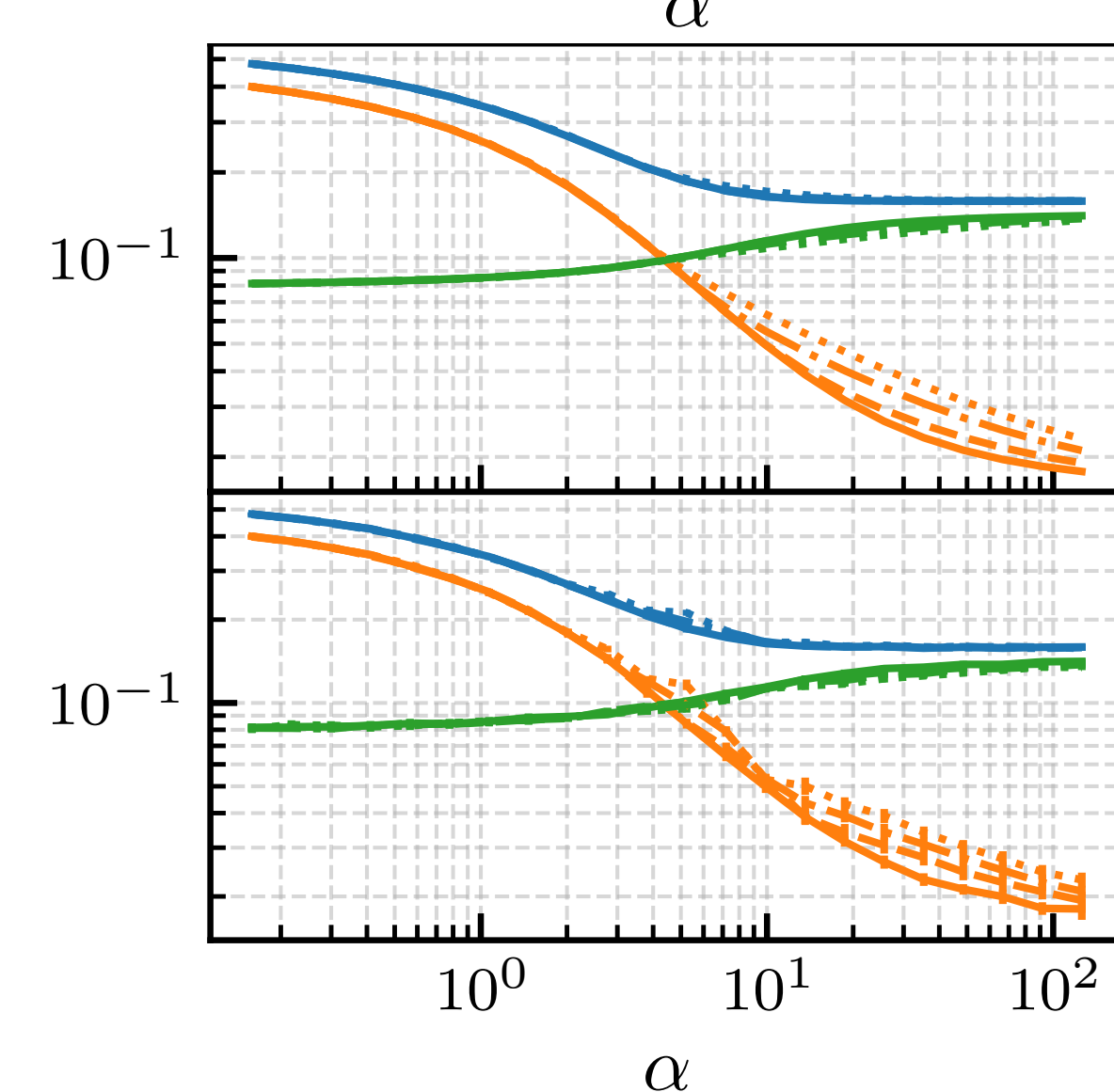
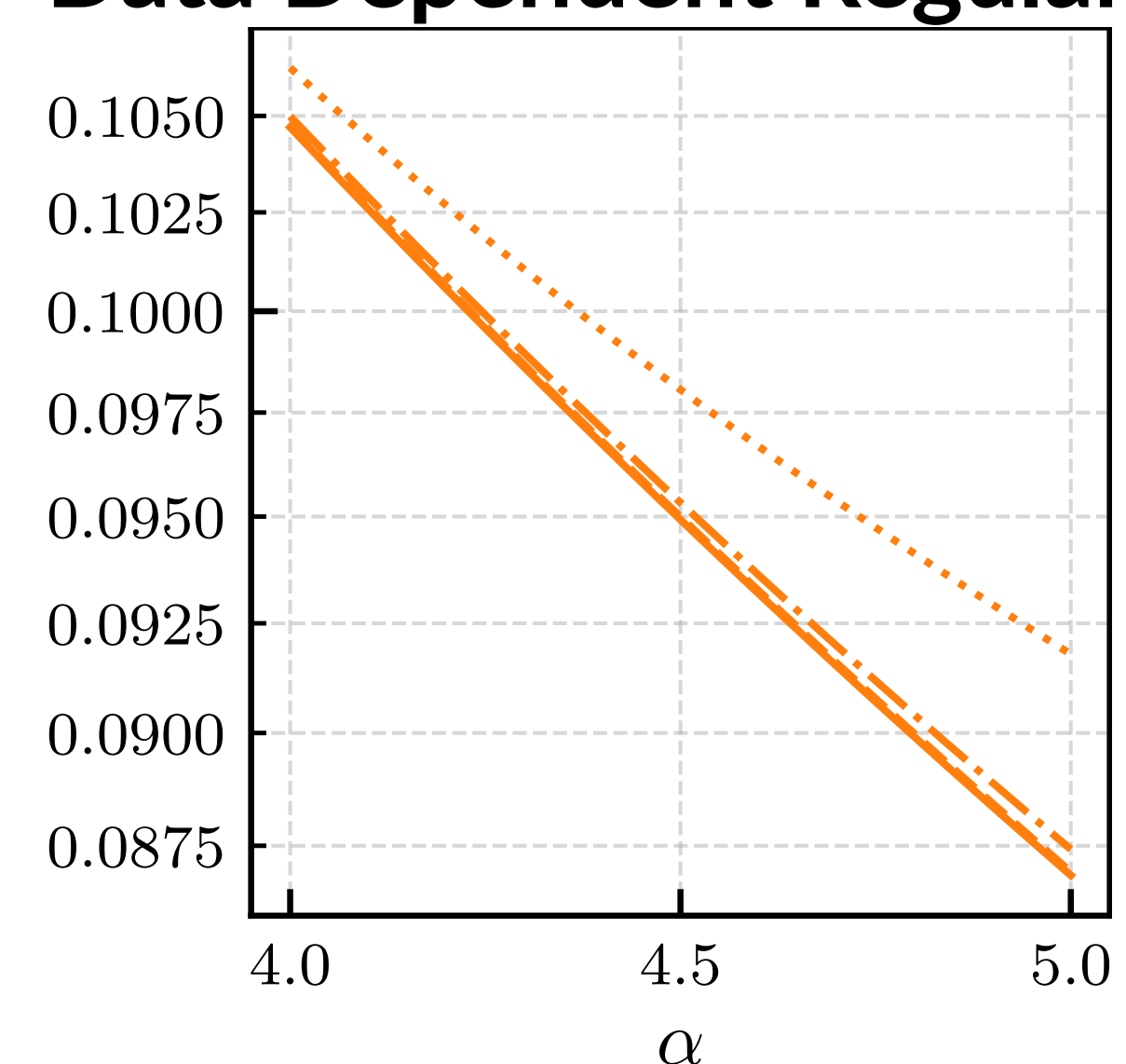
Directional Defences and structured data



Tradeoff directions and innocuous directions



Data Dependent Regularisation



Acknowledgements



Kasimir Tanner
Matteo Vilucchio
Bruno Loureiro
Florent Krzakala

EPFL