**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

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

## Block Features

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

## Usefulness and robustness

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

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

**Main Theorem** For the ERM estimator of the risk function with $\ell_2$ regularisation $r(\boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ and $\lambda \geq 0$, under the data model defined in **??** and in the high dimensional proportional limit, the generalisation error $E_{\text{gen}}$ and the boundary error $E_{\text{bnd}}$ concentrate to

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

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

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 $\varepsilon_t$ and read

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

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

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

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

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

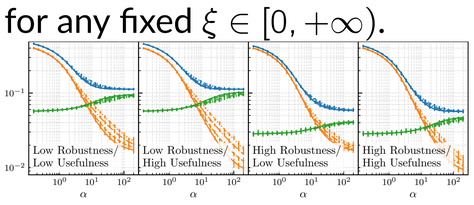
The value of $A$ can be obtained from the solution of the same system of self consistent equations as

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

## Trade-Offs

$$E_{\text{adv}} = E_{\text{gen}}(\vartheta, \mathcal{U}_{\boldsymbol{\theta}_0}) + \int_0^{\varepsilon_g \varkappa} f(\xi; \vartheta, \mathcal{U}_{\boldsymbol{\theta}_0})\, \mathrm{d}\xi\,, \tag{11}$$

where we introduce the variable $\vartheta = m/\sqrt{\rho q}$ and $\varkappa = \sqrt{A}/\sqrt{q}$. $\vartheta$ is the cosine of the angle between the teacher weights $\boldsymbol{\theta}_0$ and the student estimate $\hat{\boldsymbol{\theta}}$ in the geometry of $\Sigma_{\boldsymbol{x}}$ and $\varkappa$ is the norm of $\hat{\boldsymbol{\theta}}$ under the attack matrix. The function $f(\xi; \vartheta)$ is positive $\forall \vartheta, \forall \xi \in [0, +\infty)$ and it is strictly increasing in $\vartheta$ for any fixed $\xi \in [0, +\infty)$.




We notice that the values for $E_{\text{gen}}$ and $E_{\text{bnd}}$ change by varying the usefulness and robustness of the features for fixed types of attacks. Intuitively, we have that the more usefulness one has the less generalisation error one makes, indeed we can write a lower bound for the generalisation error

$$E_{\text{gen}} \geq \frac{1}{\pi} \arccos\left(\sqrt{\frac{\pi}{2\rho}} \mathcal{U}_{\boldsymbol{\theta}_0}\right). \tag{12}$$

We note that $\rho$ and $\mathcal{U}_{\boldsymbol{\theta}_0}$ only depend on $\Sigma_{\boldsymbol{x}}$ and $\boldsymbol{\theta}_0$. Robustness only affects the boundary error. High robustness implies less sensibility to adversarial attacks: robust features have less samples within an attack range of the student decision boundary. The highest value that the boundary error can achieve is limited by both the robustness and the usefulness as

$$\begin{aligned}
E_{\text{bnd}} \leq\ & 2\mathsf{T}\left(\varepsilon_g \mathcal{A}\mathcal{B}, \mathcal{A}^{-1}\right) - \frac{1}{\pi} \arctan\left(\mathcal{A}^{-1}\right) \\
& - \frac{1}{\pi} \text{erf}\left(\frac{\varepsilon_g \mathcal{B}}{\sqrt{2}}\right) \text{erfc}\left(\frac{\varepsilon_g \mathcal{A}\mathcal{B}}{\sqrt{2}}\right),
\end{aligned} \tag{13}$$

where $\mathcal{B} = \max_i \sqrt{(\Sigma_{\boldsymbol{\upsilon}})_{ii}/(\Sigma_{\boldsymbol{x}})_{ii}}$, $\mathcal{A} = \sqrt{\pi}\mathcal{U}_{\boldsymbol{\theta}_0}/\sqrt{2\rho}$ and $\mathsf{T}$ is the Owen function. This previous bound is a decreasing function of the robustness.

Under the same setting as **??** and considering a BFM with a single type of feature, i.e. $k = 1$ one has that $\forall \varepsilon_g, \varepsilon_t \geq 0$ for $\alpha$ big enough exist two positive numbers $M_1, M_2$ such that

$$\begin{aligned}
|E_{\text{adv}}(\varepsilon_g, \varepsilon_t) - E_{\text{adv}}(\varepsilon_g, \varepsilon_t = 0)| < M_1/\alpha\,, \\
|E_{\text{gen}}(\varepsilon_t) - E_{\text{gen}}(\varepsilon_t = 0)| < M_2/\alpha\,,
\end{aligned} \tag{14}$$

where $E_{\text{adv}}(\varepsilon_g, \varepsilon_t)$ and $E_{\text{gen}}(\varepsilon_t)$ define the adversarial and generalisation error of $\hat{\boldsymbol{\theta}}$ trained with $\varepsilon_t$ and evaluated for $\varepsilon_g$.

## Directional Defences and structured data

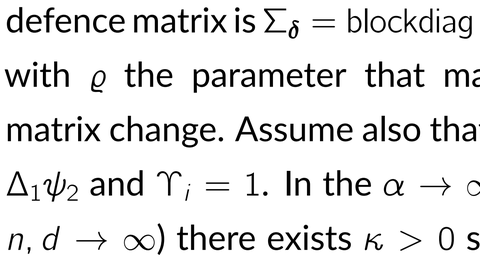
Consider the SWFM defined in **??** where the defence matrix is $\Sigma_{\boldsymbol{\delta}} = \text{blockdiag}\left((\Delta_1 + \delta_1 \varrho)\mathbb{1}_{d_1}, (\Delta_2 + \delta\right.$ with $\varrho$ the parameter that makes the defence matrix change. Assume also that $\psi_1 > \psi_2$, $\Delta_2\psi_1 \geq \Delta_1\psi_2$ and $\Upsilon_i = 1$. In the $\alpha \to \infty$ (taken after the $n, d \to \infty$) there exists $\kappa > 0$ such that $\forall \delta_1 > \kappa$, $\delta_2 = -\delta_1$ one has that

$$\begin{aligned}
E_{\text{bnd}}(\varrho) &= E_{\text{bnd}}^0 + E_{\text{bnd}}^1 \varrho + \mathcal{O}(\varrho^2)\,, \\
E_{\text{gen}}(\varrho) &= E_{\text{gen}}^0 + E_{\text{gen}}^1 \varrho + \mathcal{O}(\varrho^2)\,,
\end{aligned} \tag{15}$$

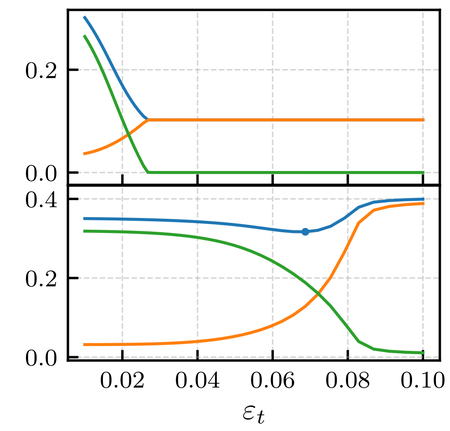
where $E_{\text{gen}}^1 > 0$, $E_{\text{bnd}}^1 < 0$ and $E_{\text{bnd}}^0$, $E_{\text{gen}}^0$ are the errors when $\varrho = 0$. Additionally, this leads to an improved value of $E_{\text{adv}}$ at order $\varrho$ iff the following condition is satisfied

$$\frac{\varepsilon_g}{\sqrt{2}} \text{erfc}\left(-\frac{\vartheta_0 u_0 \varepsilon_g}{\sqrt{2 - 2\vartheta_0^2}}\right) < \frac{e^{-\frac{\vartheta_0^2 u_0^2 \varepsilon_g^2}{2(1-\vartheta_0^2)}}}{\sqrt{\pi}\sqrt{1-\vartheta_0^2}}\,, \tag{16}$$

where $\vartheta_0 = m_0/\sqrt{\rho q_0}$ and $u_0 = \sqrt{A_0}/\sqrt{q_0}$ the solution of the problem with $\varrho = 0$. Notice that for $\varepsilon_g$ small enough this condition is always verified.




**Tradeoff directions and innocuous directions** We now investigate the effect that different types of geometries have on the trade-off between $E_{\text{gen}}$ and $E_{\text{bnd}}$. Depending on the attack geometry $\Sigma_{\boldsymbol{\upsilon}}$ one can choose different defence geometries $\Sigma_{\boldsymbol{\delta}}$ and ask if and for which, protection without trade-off is possible. Any attack matrix $\Sigma_{\boldsymbol{\upsilon}}$ eigenvalues can be split into directions orthogonal to the teacher and directions aligned with the teacher. **??** (Left) considers the effect of the adversarial training strength $\varepsilon_t$ on the errors for different choices of matrices $\Sigma_{\boldsymbol{\delta}} = \Sigma_{\boldsymbol{\upsilon}}$. In the the top we consider matrices whose biggest eigenvalues are orthogonal to the teacher vector and in the bottom one matrices where there is a leading eigenvector in the direction of the teacher.

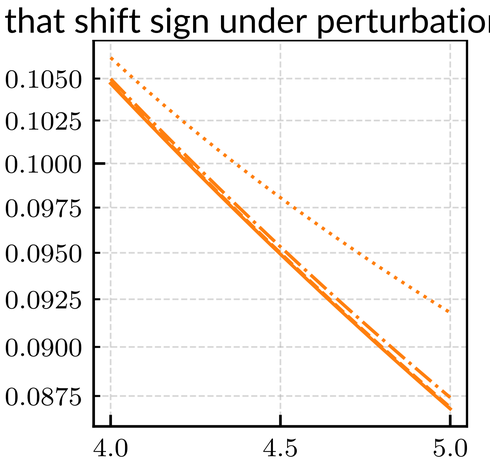


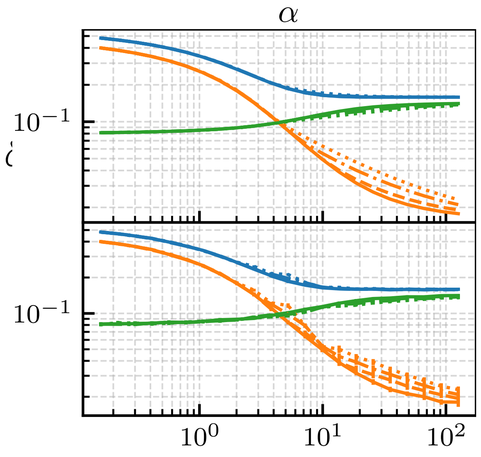

## Data Dependent Regularisation

The form for the approximate loss in the case of small $\varepsilon_t$ is

$$\sum_{i=1}^{n} g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}}\right) + \tilde{\lambda}_1 \sqrt{\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta}} + \tilde{\lambda}_2 \boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta} \tag{17}$$

where $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ depend on the model's parameters and perturbed margins of the points that shift sign under perturbation.





## Acknowledgements

We thank Lenka Zdeborová for fruitful discussions and insightful ideas regarding a class-preserving error, Lucas Clarte for useful discussions about the relevant literature, Guillaume Dalle for help in with the numerical implementation, Nikolaos Tsilivis for the discussion during the Cargese 2023 Workshop *Statistical Physics and Machine Learning back together again*, Pierre Mergny for the always helpful clarifications about Random Matrix Theory, Paul Krzakala for pointing relevant literature on adversarial training, Julia Kempe for the fruitful discussions and Vittorio Erba for rereading of the manuscript. BL acknowledges support from the *Choose France - CNRS AI Rising Talents* program, and FK from the Swiss National Science Foundation grant SNFS OperaGOST (grant number 200390).

Kasimir Tanner
Matteo Vilucchio
Bruno Loureiro
Florent Krzakala