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

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

Binary Classification

Teacher-Student Setting: Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{-1, +1\}$ where $\boldsymbol{x}_i \sim \mathcal{N}(0, \Sigma_{\boldsymbol{x}})$ and $y_i \sim \mathbb{P}(y|\boldsymbol{\theta}_0^\top \boldsymbol{x}_i)$ for fixed $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.
Focus on probit model $\mathbb{P}(y|z) = {}^1\!/_2 \operatorname{erfc}(-{}^z\!/_{\sqrt{2}\tau})$, with noise parameter $\tau > 0$.
Problem Statement: Learn $\hat{y}(\hat{\boldsymbol{\theta}}(\mathcal{D}), \boldsymbol{x}) = \operatorname{sign}(\hat{\boldsymbol{\theta}}(\mathcal{D}) \cdot \boldsymbol{x}/\sqrt{d})$ using adversarial training.
Generalisation Error:

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

Adversarial Attack: Chose $\|\boldsymbol{v}\|_{\Sigma_{\boldsymbol{v}}^{-1}} \leq \varepsilon_g$ to misclassify the sample $\boldsymbol{x}_i$: $y(\boldsymbol{x}_i) \neq \hat{y}(\hat{\boldsymbol{\theta}}, \boldsymbol{x}_i + \boldsymbol{v}_i)$.
Adversarial Generalisation Error:

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

Note, $E_{\text{gen}} = E_{\text{adv}}(\varepsilon_g = 0)$.
Boundary Error: $E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}}$ where E_{bnd} are the attackable samples.

Adversarial ERM

$$\sum_{i=1}^n \max_{\|\boldsymbol{\delta}_i\|_{\Sigma_{\boldsymbol{\delta}}^{-1}} \leq \varepsilon_t} g\left(y_i \frac{\boldsymbol{\theta}^\top(\boldsymbol{x}_i + \boldsymbol{\delta}_i)}{\sqrt{d}}\right) + r(\boldsymbol{\theta}), \tag{3}$$

where g is a convex loss, $r(\boldsymbol{\theta})$ is a convex regularisation and $\Sigma_{\boldsymbol{\delta}}$ is positive definite. We set $r(\boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$.
The inner maximisation has a closed form solution:

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}} - \varepsilon_t \frac{\sqrt{\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}} \boldsymbol{\theta}}}{\sqrt{d}}\right) + r(\boldsymbol{\theta}). \tag{4}$$

The High-Dimensional Proportional Limit (hdl). The dimension d and the number of training samples n are large $d, n \to \infty$, but the sample complexity is fixed $\alpha := n/d$.

Usefulness and robustness Usefulness relates to generalisation error and robustness relates to boundary error.

$$\mathcal{U}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{x},y}[y\boldsymbol{\theta}_0^\top \boldsymbol{x}], \tag{5}$$

$$\mathcal{R}_{\boldsymbol{\theta}_0} = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{x},y}\left[\inf_{\|\boldsymbol{\delta}\|_{\Sigma_{\boldsymbol{v}}^{-1}} \leq \varepsilon_g} y\boldsymbol{\theta}_0^\top(\boldsymbol{x} + \boldsymbol{\delta})\right]. \tag{6}$$

Block Features Craft artificial usefulness and robustness:

$$\begin{aligned}
\Sigma_{\boldsymbol{x}} &= \operatorname{blockdiag}(\psi_1 \mathbb{1}_{d_1}, \ldots, \psi_k \mathbb{1}_{d_k}), \\
\Sigma_{\boldsymbol{\delta}} &= \operatorname{blockdiag}(\Delta_1 \mathbb{1}_{d_1}, \ldots, \Delta_k \mathbb{1}_{d_k}), \\
\Sigma_{\boldsymbol{v}} &= \operatorname{blockdiag}(\Upsilon_1 \mathbb{1}_{d_1}, \ldots, \Upsilon_k \mathbb{1}_{d_k}), \\
\Sigma_{\boldsymbol{\theta}} &= \operatorname{blockdiag}(t_1 \mathbb{1}_{d_1}, \ldots, t_k \mathbb{1}_{d_k}),
\end{aligned} \tag{7}$$

Main Theorem **Assumptions:**

- Well defined spectral distribution in the hdl $\Sigma_{\boldsymbol{x}} = \mathsf{S}^\top \operatorname{diag}(\omega_i)\mathsf{S}$, $\zeta_i = \operatorname{diag}(\mathsf{S}\Sigma_{\boldsymbol{\delta}}\mathsf{S}^\top)_i$ and $\upsilon_i = \operatorname{diag}(\mathsf{S}\Sigma_{\boldsymbol{v}}\mathsf{S}^\top)_i$.
- Assume that $\boldsymbol{\theta}_0^\top \Sigma_{\boldsymbol{x}} \boldsymbol{\theta}_0/d$ converges to ρ in the limit and that the entries of $\bar{\boldsymbol{\theta}} = \mathsf{S}\Sigma_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0/\sqrt{\rho}$ converge as well to a limiting distribution.
- Assume that in the hdl the spectral distributions for the matrices and the distributions of the elements of the vectors just defined converge jointly to a p.d.f., *i.e.* ${}^1\!/_d \sum_{i=1}^d \delta(\omega - \omega_i)\delta(\bar{\theta} - \bar{\theta}_i)\delta(\zeta - \zeta_i)\delta(\upsilon - \upsilon_i) \to \mu(\omega, \bar{\theta}, \zeta, \upsilon)$.

For $\lambda \geq 0$

$$E_{\text{gen}} = \frac{1}{\pi}\arccos\left(m/\sqrt{(\rho + \tau^2)q}\right), \tag{8}$$

$$E_{\text{bnd}} = \int_0^{\varepsilon_g \frac{\sqrt{A}}{\sqrt{q}}} \operatorname{erfc}\left(\frac{-\frac{m}{\sqrt{q}}\nu}{\sqrt{2(\rho + \tau^2 - m^2/q)}}\right)\frac{e^{-\frac{\nu^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}\nu, \tag{9}$$

and $E_{\text{adv}} = E_{\text{gen}} + E_{\text{bnd}}$.
$(m, q, V, P, \hat{m}, \hat{q}, \hat{V}, \hat{P})$ are the solutions of the following system of equations. Four are dependent on the loss function g and the adversarial training strength ε_t:

$$\begin{cases}
\hat{m} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P)\right] \\
\hat{q} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0 f_g^2(y, \sqrt{q}\xi, P)\right] \\
\hat{V} = -\alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0 \partial_\omega f_g(y, \sqrt{q}\xi, P)\right] \\
\hat{P} = -\frac{\varepsilon_t}{2\sqrt{P}}\alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, y\mathcal{Z}_0 f_g(y, \sqrt{q}\xi, P)\right]
\end{cases}, \tag{10}$$

where $\xi \sim \mathcal{N}(0, 1)$ and $\mathcal{Z}_0 = {}^1\!/_2 \operatorname{erfc}({}^{-y\omega}\!/_{\sqrt{2(V+\tau^2)}})$ and $f_g(y, \omega, V, P) = (\mathcal{P}(\omega) - \omega)/V$, where $\mathcal{P}$ is the following proximal operator

$$\mathcal{P}(\omega) = \min_x\left[\frac{(x - \omega)^2}{2V} + g(yx - \varepsilon_t\sqrt{P})\right]. \tag{11}$$

The other four are dependent on the spectral distribution of the matrices $\Sigma_{\boldsymbol{x}}, \Sigma_{\boldsymbol{\delta}}$ and on the limiting distribution of the elements of $\bar{\boldsymbol{\theta}}$:

$$\begin{cases}
m = \mathbb{E}_\mu\left[\frac{\hat{m}\bar{\theta}^2}{\lambda + \hat{V}\omega + \hat{P}\delta}\right] \\
q = \mathbb{E}_\mu\left[\frac{\hat{m}^2\bar{\theta}^2\omega + \hat{q}\omega^2}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right] \\
V = \mathbb{E}_\mu\left[\frac{\omega}{\lambda + \hat{V}\omega + \hat{P}\delta}\right] \\
P = \mathbb{E}_\mu\left[\zeta\frac{\hat{m}^2\bar{\theta}^2 + \hat{q}\omega^2}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right]
\end{cases}. \tag{12}$$

A is given by

$$A = \mathbb{E}_\mu\left[\upsilon\frac{\hat{m}^2\bar{\theta}^2\omega + \hat{q}\omega^2}{(\lambda + \hat{V}\omega + \hat{P}\delta)^2}\right]. \tag{13}$$

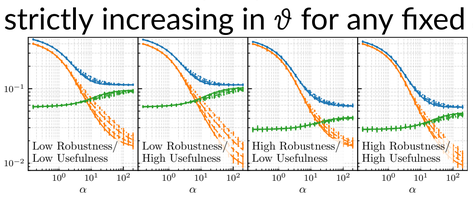
m, q, P and A are interpretable:

$$\begin{aligned}
m &= \mathbb{E}_{\mathcal{D}}\left[\tfrac{1}{d}\boldsymbol{\theta}_0^\top \Sigma_{\boldsymbol{x}}\hat{\boldsymbol{\theta}}\right], \quad q = \mathbb{E}_{\mathcal{D}}\left[\tfrac{1}{d}\hat{\boldsymbol{\theta}}^\top \Sigma_{\boldsymbol{x}}\hat{\boldsymbol{\theta}}\right], \\
P &= \mathbb{E}_{\mathcal{D}}\left[\tfrac{1}{d}\hat{\boldsymbol{\theta}}^\top \Sigma_{\boldsymbol{\delta}}\hat{\boldsymbol{\theta}}\right], \quad A = \mathbb{E}_{\mathcal{D}}\left[\tfrac{1}{d}\hat{\boldsymbol{\theta}}^\top \Sigma_{\boldsymbol{v}}\hat{\boldsymbol{\theta}}\right].
\end{aligned} \tag{14}$$

Trade-Offs

$$E_{\text{adv}} = E_{\text{gen}}(\vartheta, \mathcal{U}_{\boldsymbol{\theta}_0}) + \int_0^{\varepsilon_g \varkappa} f(\xi; \vartheta, \mathcal{U}_{\boldsymbol{\theta}_0})\,\mathrm{d}\xi, \tag{15}$$

where we introduce the variable $\vartheta = m/\sqrt{\rho q}$ and $\varkappa = \sqrt{A}/\sqrt{q}$. ϑ 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 ϑ for any fixed $\xi \in [0, +\infty)$.



We notice that the values for E_{gen} and E_{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

$$E_{\text{gen}} \geq \frac{1}{\pi}\arccos\left(\sqrt{\frac{\pi}{2\rho}}\mathcal{U}_{\boldsymbol{\theta}_0}\right). \tag{16}$$

We note that ρ 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

$$\begin{aligned}
E_{\text{bnd}} \leq\ & 2\mathsf{T}\left(\varepsilon_g \mathcal{A}\mathcal{B}, \mathcal{A}^{-1}\right) - \frac{1}{\pi}\arctan\left(\mathcal{A}^{-1}\right) \\
& - \frac{1}{\pi}\operatorname{erf}\left(\frac{\varepsilon_g \mathcal{B}}{\sqrt{2}}\right)\operatorname{erfc}\left(\frac{\varepsilon_g \mathcal{A}\mathcal{B}}{\sqrt{2}}\right),
\end{aligned} \tag{17}$$

where $\mathcal{B} = \max_i \sqrt{(\Sigma_{\boldsymbol{v}})_{ii}/(\Sigma_{\boldsymbol{x}})_{ii}}$, $\mathcal{A} = \sqrt{\pi}\mathcal{U}_{\boldsymbol{\theta}_0}/\sqrt{2\rho}$ and 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 α big enough exist two positive numbers M_1, M_2 such that

$$\begin{aligned}
|E_{\text{adv}}(\varepsilon_g, \varepsilon_t) - E_{\text{adv}}(\varepsilon_g, \varepsilon_t = 0)| &< M_1/\alpha, \\
|E_{\text{gen}}(\varepsilon_t) - E_{\text{gen}}(\varepsilon_t = 0)| &< M_2/\alpha,
\end{aligned} \tag{18}$$

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 ε_t and evaluated for ε_g.

Directional Defences and structured data

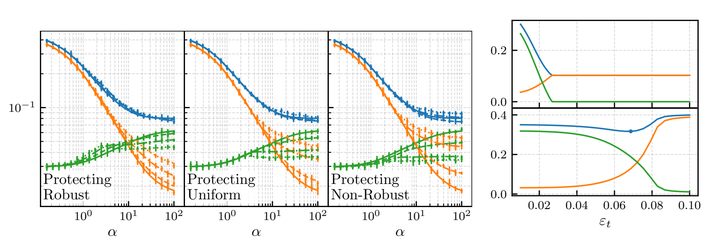
Consider the SWFM defined in **??** where the defence matrix is $\Sigma_{\boldsymbol{\delta}} = \operatorname{blockdiag}((\Delta_1 + \delta_1\varrho)\mathbb{1}_{d_1}, (\Delta_2 + \delta_2\varrho)\mathbb{1}_{d_2})$, with ϱ the parameter that makes the defence matrix change. Assume also that $\psi_1 > \psi_2$, $\Delta_2\psi_1 \geq \Delta_1\psi_2$ 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

$$\begin{aligned}
E_{\text{bnd}}(\varrho) &= E_{\text{bnd}}^0 + E_{\text{bnd}}^1\varrho + \mathcal{O}(\varrho^2), \\
E_{\text{gen}}(\varrho) &= E_{\text{gen}}^0 + E_{\text{gen}}^1\varrho + \mathcal{O}(\varrho^2),
\end{aligned} \tag{19}$$

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_{adv} at order ϱ iff the following condition is satisfied

$$\frac{\varepsilon_g}{\sqrt{2}}\operatorname{erfc}\left(-\frac{\vartheta_0 u_0 \varepsilon_g}{\sqrt{2 - 2\vartheta_0^2}}\right) < \frac{e^{-\frac{\vartheta_0^2 u_0^2 \varepsilon_g^2}{2(1-\vartheta_0^2)}}}{\sqrt{\pi}\sqrt{1 - \vartheta_0^2}}, \tag{20}$$

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 ε_g small enough this condition is always verified.



TODO add caption

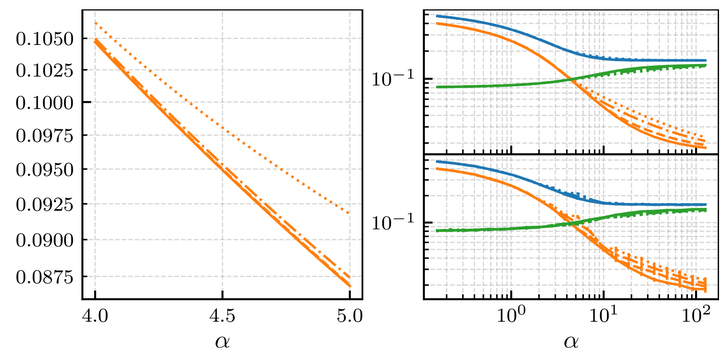
Tradeoff directions and innocuous directions We now investigate the effect that different types of geometries have on the trade-off between E_{gen} and E_{bnd}. Depending on the attack geometry $\Sigma_{\boldsymbol{v}}$ one can choose different defence geometries $\Sigma_{\boldsymbol{\delta}}$ and ask if and for which, protection without trade-off is possible. Any attack matrix $\Sigma_{\boldsymbol{v}}$ 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 ε_t on the errors for different choices of matrices $\Sigma_{\boldsymbol{\delta}} = \Sigma_{\boldsymbol{v}}$. 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 ε_t is

$$\sum_{i=1}^n g\left(y_i \frac{\boldsymbol{\theta}^\top \boldsymbol{x}_i}{\sqrt{d}}\right) + \tilde{\lambda}_1\sqrt{\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}}\boldsymbol{\theta}} + \tilde{\lambda}_2\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\delta}}\boldsymbol{\theta} \tag{21}$$

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.



TODO better caption

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