Heavy-tailed covariates in high dimensions

Gabriele Sicuro

University of Bologna

1st of October 2024

Urte Adomaityte
KCL

Pierpaolo Vivo
KCL

Leonardo Defilippis
ENS

Bruno Loureiro
ENS

The problem in this talk: a supervised learning task

Given a dataset

$$\mathcal{D} = \{(y_\nu, \boldsymbol{x}_\nu)\}_{\nu=1}^{n}, \quad (y_\nu, \boldsymbol{x}_\nu) \sim \mathbb{P}(\mathcal{Y} \times \mathbb{R}^d) \quad \text{i.i.d.}$$

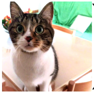
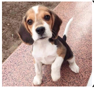
construct a predictor in the form

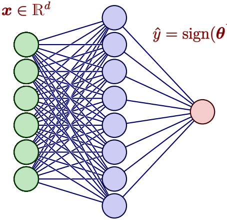
$$\hat{y} = \hat{y}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} \in \mathbb{R}^p$$

where $\hat{\boldsymbol{\theta}}$ has to be found by minimizing some *empirical risk function*,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\frac{1}{n} \sum_{\nu=1}^{n} \underbrace{\ell(y_\nu, \boldsymbol{x}_\nu; \boldsymbol{\theta})}_{\text{loss}} + \underbrace{\lambda \|\boldsymbol{\theta}\|^2}_{\text{reg.}} \right]$$

For example

$(y_\nu, \boldsymbol{x}_\nu) =$ e.g. $\Big(+1,$  $\Big)$ or $\Big(-1,$  $\Big)$ $\Longrightarrow$



The problem in this talk: a supervised learning task

Given a dataset

$$\mathcal{D} = \{(y_\nu, \boldsymbol{x}_\nu)\}_{\nu=1}^{n}, \quad (y_\nu, \boldsymbol{x}_\nu) \sim \mathbb{P}(\mathcal{Y} \times \mathbb{R}^d) \quad \text{i.i.d.}$$

construct a predictor in the form

$$\hat{y} = \hat{y}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} \in \mathbb{R}^p$$

where $\hat{\boldsymbol{\theta}}$ has to be found by minimizing some *empirical risk function*,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\frac{1}{n} \sum_{\nu=1}^{n} \underbrace{\ell(y_\nu, \boldsymbol{x}_\nu; \boldsymbol{\theta})}_{\text{loss}} + \underbrace{\lambda \|\boldsymbol{\theta}\|^2}_{\text{reg.}} \right]$$

We are interested in the statistics of $\hat{\boldsymbol{\theta}}$ over the ensemble induced by $\mathbb{P}$ as

$$n, d, p \to +\infty \quad \text{with} \quad {}^{n}\!/\!{}_{d} = \Theta(1) \quad {}^{p}\!/\!{}_{d} = \Theta(1)$$

and in particular in its *asymptotic performance*

$$\epsilon_\ell := \lim_n \frac{1}{n} \sum_{\nu=1}^{n} \ell(y_\nu, \boldsymbol{x}_\nu; \hat{\boldsymbol{\theta}}), \quad \epsilon_t := \lim_n \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{I}\left(y_\nu \neq \hat{y}(\boldsymbol{x}_\nu; \hat{\boldsymbol{\theta}})\right), \quad \epsilon_g := \mathbb{E}\left[\mathbb{I}\left(y \neq \hat{y}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})\right)\right].$$

The problem in this talk: a supervised learning task

A very large number of works focused on this setting. The *theoretical analysis* goes through modeling choices of different ingredients.

- The "**architecture**" through the design of risk/predictor.
- The **optimization algorithm** adopted to find $\hat{\boldsymbol{\theta}}$.
- The **dataset structure**.

The problem in this talk: a supervised learning task

A very large number of works focused on this setting. The *theoretical analysis* goes through modeling choices of different ingredients.

- The **"architecture"** through the design of risk/predictor.
- The **optimization algorithm** adopted to find $\hat{\boldsymbol{\theta}}$.
- The **dataset structure**.

A Leitmotif of the theoretical investigations in Statistics and Statistical Physics since pioneering works, has been a *Gaussian design* for the covariates $\{\boldsymbol{x}_\nu\}_\nu$:

- $\boldsymbol{x}_\nu \sim P$ where P is a *Gaussian distribution* or a *Gaussian mixture*, possibly in presence of *correlation*;

  Mei and Montanari (2022); Gerace et al. (2020); Mignacco et al. (2020); Baldassi et al. (2020); Loureiro et al. (2021)…
- a *Gaussian Equivalence Principle* allows an effectively Gaussian description.

  Montanari et al. (2019); Mei and Montanari (2022); Goldt et al. (2020,2022)…

The problem in this talk: a supervised learning task

A very large number of works focused on this setting. The *theoretical analysis* goes through modeling choices of different ingredients.

- The **"architecture"** through the design of risk/predictor.
- The **optimization algorithm** adopted to find $\hat{\boldsymbol{\theta}}$.
- The **dataset structure**.

A Leitmotif of the theoretical investigations in Statistics and Statistical Physics since pioneering works, has been a *Gaussian design* for the covariates $\{\boldsymbol{x}_\nu\}_\nu$:

- $\boldsymbol{x}_\nu \sim P$ where P is a *Gaussian distribution* or a *Gaussian mixture*, possibly in presence of *correlation*;

  Mei and Montanari (2022); Gerace et al. (2020); Mignacco et al. (2020); Baldassi et al. (2020); Loureiro et al. (2021)...
- a *Gaussian Equivalence Principle* allows an effectively Gaussian description.

  Montanari et al. (2019); Mei and Montanari (2022); Goldt et al. (2020,2022)...

Are these Gaussian assumptions "good enough"?

Hints of Gaussian equivalence

Gaussian equivalence hypothesis:

$$P(\boldsymbol{x}) \Longrightarrow \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}], \quad \boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal] \quad ?$$

Theorem (*informal*)
Hu and Lu (2022); Montanari and Saeed (2022); Dandi, Stephan, Krzakala, Loureiro, Zdeborová (2023)

In an ERM task, training loss and test error are universal and **corresponding to a Gaussian equivalent setting** as long as the features are such that

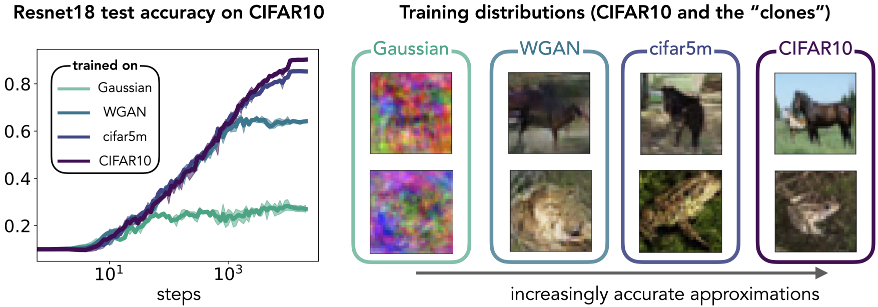
$$\sup_{\|\boldsymbol{v}\|\leq 1} \|\boldsymbol{v}^\intercal \boldsymbol{x}\|_{\psi_2} < +\infty, \qquad \lim_d \sup_{\boldsymbol{v}} |\mathbb{E}[f(\boldsymbol{v}^\intercal \boldsymbol{x})] - \mathbb{E}[f(\boldsymbol{v}^\intercal \boldsymbol{z})]| = 0$$

for f bounded Lipschitz and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

☞ The subgaussianity condition is not an artifact of the proof: it is a *necessary* condition!

See also: Donoho and Tanner (2009); Bordelon, Canatar, Pehlevan (2020); Spigler, Geiger, Wyart (2020); Jacot, Şimşek, Spadaro, Hongler, Gabriel (2020); Seddik, Louart, Couillet, and Tamaazousti (2020); Loureiro, Gerbelot, Cui, Goldt, Krzakala, Mézard, Zdeborová (2021); Loureiro, Sicuro, Gerbelot, Pacco, Krzakala, Zdeborová (2021)

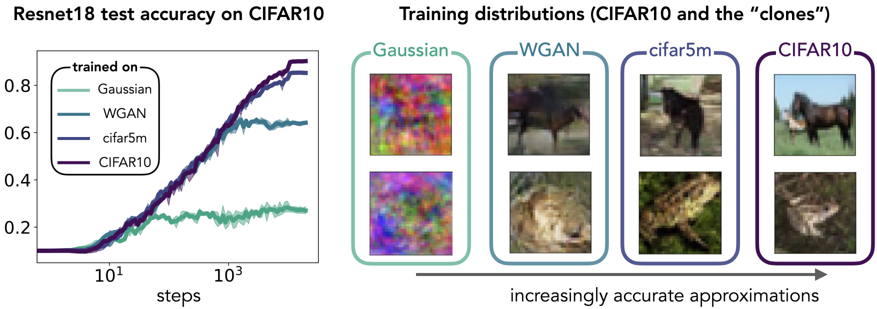
Relevance of HOCs



Refinetti, Ingrosso, Goldt (2022)

Real data are not Gaussian and neural networks can "see" (and exploit) that.

Relevance of HOCs



Refinetti, Ingrosso, Goldt (2022)

Real data are not Gaussian and neural networks can "see" (and exploit) that.

What is the simplest model for non-Gaussian covariates
that "breaks" Gaussian universality?

Gaussian Scale Mixtures

Gaussian scale mixtures *aka* Elliptic distributions *aka* Superstatistics

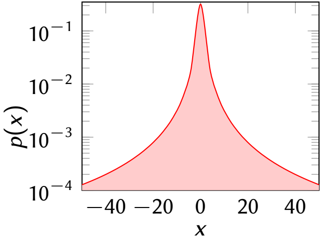
Gaussian scale mixtures (GSMs) [Andrews, Mallows (1974)] have the form

$$\boxed{\boldsymbol{x} \overset{\mathrm{d}}{=} \frac{1}{\sqrt{d}}\sigma \boldsymbol{z}}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \qquad \boxed{\sigma \sim \varrho \quad \text{positive}}$$

Theorem [Andrews, Mallows (1974)]

$x \sim \sigma z$ with $z \sim \mathcal{N}(0,1)$ for some ϱ iff $\mathbb{E}[\mathrm{x}] = 0$ and $(-\partial_y)^k p_{\mathrm{x}}(\sqrt{y}) \geq 0$ for all $k \in \mathbb{N}_0$.

- Covariance $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T}] = \frac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$: possibly infinite covariance when $\mathbb{E}[\sigma^2] = +\infty$.
- $\sup_{\|\boldsymbol{v}\| \leq 1} \|\boldsymbol{v}^\mathsf{T}\boldsymbol{x}\|_{\psi_2} = +\infty$ if σ has unbounded support.



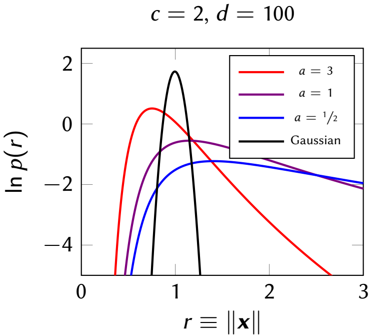
$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2} = \int_0^\infty \frac{\mathrm{e}^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \underbrace{\sqrt{\frac{2}{\pi}}\frac{\mathrm{e}^{-\frac{1}{2\sigma^2}}}{\sigma^2}}_{\varrho(\sigma)} \mathrm{d}\sigma.$$

The geometry of GSM

$$\boxed{\boldsymbol{x} \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{d}}\sigma \boldsymbol{z}}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \qquad \boxed{\sigma \sim \varrho \quad \text{positive}}$$

As $d \to +\infty$,

$$\Pr[\|\boldsymbol{x}\| > r] \to \Pr[\sigma > r].$$



Example

$$\varrho(\sigma) \propto \frac{\exp\left(-\frac{c}{\sigma^2}\right)}{\sigma^{2a+1}} \Rightarrow p(\boldsymbol{x}) \propto \left(1 + \frac{d\|\boldsymbol{x}\|^2}{2c}\right)^{-a-\frac{d}{2}}$$

$$\sigma_0^2 := \mathbb{E}[\|\boldsymbol{x}\|^2] = \begin{cases} \frac{c}{a-1} & \text{if } a > 1 \\ +\infty & \text{if } 0 < a \leq 1 \end{cases}$$

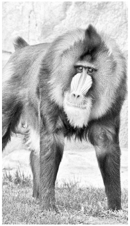
We can recover the Gaussian limit as

$$\varrho(\sigma) \xrightarrow[\sigma_0^2 = \frac{c}{a-1}]{a \to +\infty} \delta\left(\sigma^2 - \sigma_0^2\right).$$

A motivation

GSMs are not just a theoretical dataset model. In 1999, M.J. Wainwright and E.P. Simoncelli observed that GSMs appears in the statistics of natural images.

By analysing the distribution of a wavelet subband of natural images they observed a striking GSM structure.



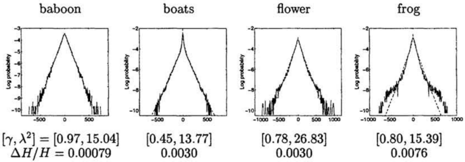


Figure 1. GSMs (dashed lines) fitted to empirical histograms (solid lines). Below each plot are the parameter values, and the relative entropy between the histogram (with 256 bins) and the model, as a fraction of the histogram entropy.

GSMs features are therefore not that artificial.

Classification of heavy-tailed clouds

with Urte Adomaityte and Pierpaolo Vivo

Set-up for classification

We consider a database of features $\mathcal{D} = \{(y_\nu, \boldsymbol{x}_\nu)\}_{\nu=1}^n$ generated via

$$\boldsymbol{x}_\nu = y_\nu \boldsymbol{\mu} + \frac{1}{\sqrt{d}}\sigma_\nu \boldsymbol{z}_\nu, \qquad y_\nu \sim \text{Rad}, \quad \boxed{\sigma_\nu \sim \varrho}, \quad \boldsymbol{z}_\nu \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad \|\boldsymbol{\mu}\|^2 = {}^1\!/_d.$$

The goal is to find $\hat{\boldsymbol{\theta}}$ such that

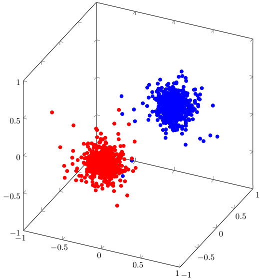
$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{\nu=1}^n \ell(y_\nu \boldsymbol{\theta}^\intercal \boldsymbol{x}_\nu) + \lambda \|\boldsymbol{\theta}\|^2\right]$$

We considered **square** and **logistic** loss:

$$\ell(y\boldsymbol{\theta}^\intercal \boldsymbol{x}) = (1 - y\boldsymbol{\theta}^\intercal \boldsymbol{x})^2, \qquad \ell(y\boldsymbol{\theta}^\intercal \boldsymbol{x}) = \ln(1 + \mathrm{e}^{-y\boldsymbol{\theta}^\intercal \boldsymbol{x}})$$

and given a new datapoint $\boldsymbol{x} \in \mathbb{R}^d$, the prediction will be given by

$$\boldsymbol{x} \mapsto \hat{y} = \operatorname{sign}\left(\hat{\boldsymbol{\theta}}^\intercal \boldsymbol{x}\right) \in \{-1, 1\}.$$



We work in the proportional regime, $n, d \to +\infty$, $\alpha = {}^n\!/_d = \Theta(1)$.

How to solve the problem: the replica method in a nutshell

$$\hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(y_\nu \boldsymbol{\theta}^\intercal \boldsymbol{x}_\nu\right) + \lambda\|\boldsymbol{\theta}\|^2 \Rightarrow \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta})$$

We aim at computing, in the proportional limit

$$\min_{\boldsymbol{\theta}} \hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta}) = -\lim_{\beta\to+\infty} \frac{1}{\beta} \ln \underbrace{\int \mathrm{e}^{-\beta\hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta})}\,\mathrm{d}\,\boldsymbol{\theta}}_{Z_{\mathcal{D}}(\beta)}.$$

To do so, we make, first of all, a *concentration assumption*

$$\lim_n \min_{\boldsymbol{\theta}} \hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta}) = -\lim_n \lim_{\beta\to+\infty} \frac{\mathbb{E}[\ln Z_{\mathcal{D}}(\beta)]}{\beta}.$$

Then we apply the so-called *replica trick*

$$-\lim_n \mathbb{E}[\ln Z_{\mathcal{D}}(\beta)] = \lim_{s\to 0} \frac{1-\lim_n \mathbb{E}[Z^s_{\mathcal{D}}(\beta)]}{s}.$$

What it is found in the end is that the problem can be rewritten in terms of a *low-dimensional expression* depending on a set of *order parameters*

$$\lim_n \min_{\boldsymbol{\theta}} \hat{\mathcal{R}}_{\mathcal{D}}(\boldsymbol{\theta}) = \min_{q,m,v} \Phi(q,m,v).$$

A set of equations for the order parameters

$$q = \lim_{d\to+\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}\|^2, \qquad m = \lim_{d\to+\infty} \boldsymbol{\mu}^\intercal \hat{\boldsymbol{\theta}}.$$

plus an auxiliary one, v, solving $[\zeta \sim \mathcal{N}(0,1)]$

$$m=\frac{\hat m}{\lambda+\hat v}, \qquad \hat q=\alpha\mathbb{E}[\sigma^2 f^2], \qquad \omega:=m+\sigma\sqrt{q}\zeta,$$
$$q=\frac{\hat m^2+\hat q}{(\lambda+\hat v)^2}, \qquad \hat v=-\alpha\frac{\mathbb{E}[\sigma f\zeta]}{\sqrt{q}}, \qquad h:=\operatorname*{argmin}_u\left[\frac{(u-\omega)^2}{2\sigma^2 v}+\ell(u)\right],$$
$$v=\frac{1}{\lambda+\hat v}, \qquad \hat m=\alpha\mathbb{E}[f], \qquad f:=\frac{h-\omega}{\sigma^2 v}.$$

As in the Gaussian case

$$\epsilon_\ell = \mathbb{E}\left[\ell(-h)\right]$$
$$\epsilon_t = \mathbb{E}[\theta(-h)]$$
$$\epsilon_g = \mathbb{E}[\theta(-\omega)].$$

- $\hat{\boldsymbol{\theta}}$ is Gaussian

$$\hat{\boldsymbol{\theta}} \stackrel{\mathrm{d}}{=} m\sqrt{d}\boldsymbol{\mu} + \sqrt{q^2-m^2}\boldsymbol{\xi}, \qquad \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d).$$

- If $(y,\mathbf{x}) \notin \mathcal{D}$, then

$$\hat{\boldsymbol{\theta}}^\intercal\mathbf{x} \stackrel{\mathrm{d}}{=} ym+\sigma\sqrt{q}\zeta, \qquad \zeta\sim\mathcal{N}(0,1).$$

- If $(y,\mathbf{x}) \in \mathcal{D}$, then

$$\hat{\boldsymbol{\theta}}^\intercal\mathbf{x} \stackrel{\mathrm{d}}{=} \arg\min_u\left[\frac{(ym+\sqrt{q}\sigma\zeta-u)^2}{2\sigma^2 v}+\ell(yu)\right], \qquad \zeta\sim\mathcal{N}(0,1).$$

A set of equations for the order parameters

$$q = \lim_{d\to+\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}\|^2, \qquad m = \lim_{d\to+\infty} \boldsymbol{\mu}^\intercal \hat{\boldsymbol{\theta}}.$$

plus an auxiliary one, v, solving $[\zeta \sim \mathcal{N}(0,1)]$

$$\begin{aligned}
&m=\frac{\hat m}{\lambda+\hat v}, && \hat q=\alpha\mathbb{E}[\sigma^2 f^2], && \omega:=m+\sigma\sqrt{q}\zeta,\\
&q=\frac{\hat m^2+\hat q}{(\lambda+\hat v)^2}, && \hat v=-\alpha\frac{\mathbb{E}[\sigma f\zeta]}{\sqrt{q}}, && h:=\arg\min_u\left[\frac{(u-\omega)^2}{2\sigma^2 v}+\ell(u)\right],\\
&v=\frac{1}{\lambda+\hat v}, && \hat m=\alpha\mathbb{E}[f], && f:=\frac{h-\omega}{\sigma^2 v}.
\end{aligned}$$

As in the Gaussian case

$$\begin{aligned}
\epsilon_\ell &= \mathbb{E}\left[\ell(-h)\right]\\
\epsilon_t &= \mathbb{E}[\theta(-h)]\\
\epsilon_g &= \mathbb{E}[\theta(-\omega)].
\end{aligned}$$

- $\hat{\boldsymbol{\theta}}$ is Gaussian

$$\hat{\boldsymbol{\theta}} \stackrel{\mathrm{d}}{=} m\sqrt{d}\boldsymbol{\mu} + \sqrt{q^2-m^2}\boldsymbol{\xi}, \qquad \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d).$$

- If $(y,\mathbf{x})\notin\mathcal{D}$, then

$$\hat{\boldsymbol{\theta}}^\intercal\mathbf{x} \stackrel{\mathrm{d}}{=} ym+\sigma\sqrt{q}\zeta, \qquad \zeta\sim\mathcal{N}(0,1).$$

- If $(y,\mathbf{x})\in\mathcal{D}$, then

e.g., square loss $$\hat{\boldsymbol{\theta}}^\intercal\mathbf{x} \stackrel{\mathrm{d}}{=} \frac{m+\sigma^2 v}{1+\sigma^2 v}y + \frac{\sqrt{q}\sigma}{1+\sigma^2 v}\zeta, \qquad \zeta\sim\mathcal{N}(0,1).$$

"Breaking" Gaussian universality

Let us consider a mixture of clouds having the **same finite covariance**

$$\boldsymbol{\Sigma}_\pm = \tfrac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$$

Using

$$\varrho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \exp\left(-\frac{c}{\sigma^2}\right) \quad \text{with} \quad \frac{c}{a-1} = \mathbb{E}[\sigma^2]$$

"Breaking" Gaussian universality

Let us consider a mixture of clouds having the **same finite covariance**

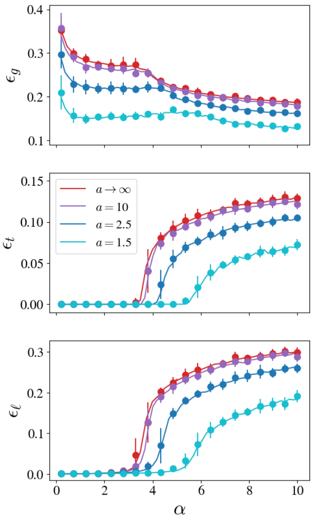
$$\boldsymbol{\Sigma}_{\pm} = \frac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$$

Using

$$\varrho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \exp\left(-\frac{c}{\sigma^2}\right) \quad \text{with} \quad \frac{c}{a-1} = \mathbb{E}[\sigma^2]$$

- **No "Gaussian universality"**: training error, training loss, test error all depend on the tail exponent a although matching first and second moments.

Logistic loss $\lambda = 10^{-4}$



“Breaking” Gaussian universality

Let us consider a mixture of clouds having the **same finite covariance**

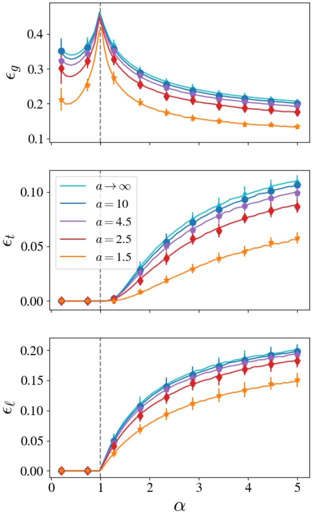
$$\boldsymbol{\Sigma}_\pm = \tfrac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$$

Using

$$\varrho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \exp\left(-\frac{c}{\sigma^2}\right) \quad \text{with} \quad \frac{c}{a-1} = \mathbb{E}[\sigma^2]$$

- **No “Gaussian universality”**: training error, training loss, test error all depend on the tail exponent a although matching first and second moments.

Square loss with $\lambda = 10^{-5}$.



“Breaking” Gaussian universality

Let us consider a mixture of clouds having the **same finite covariance**

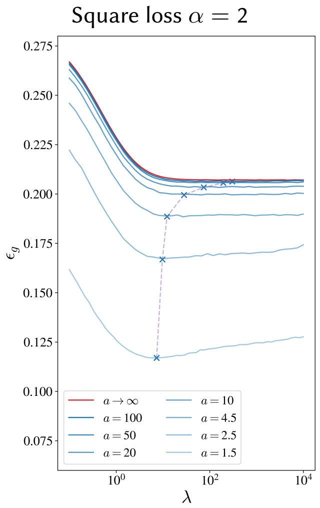
$$\boldsymbol{\Sigma}_{\pm} = \tfrac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$$

Using

$$\varrho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \exp\left(-\frac{c}{\sigma^2}\right) \quad \text{with} \quad \frac{c}{a-1} = \mathbb{E}[\sigma^2]$$

- **No “Gaussian universality”**: training error, training loss, test error all depend on the tail exponent a although matching first and second moments.
- **With square loss**, optimal performance for finite λ, with $\lambda \to +\infty$ in the limit of Gaussian clouds.

  Mignacco et al. (2020); Baldassi et al. (2020).



“Breaking” Gaussian universality

Let us consider a mixture of clouds having the **same finite covariance**

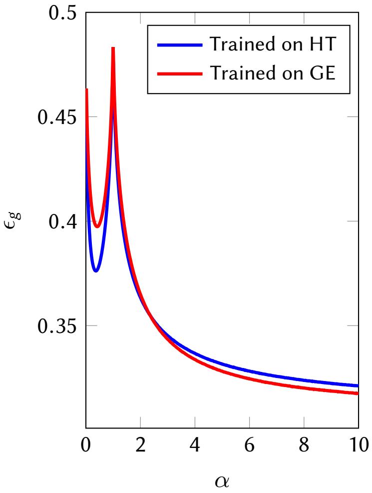
$$\mathbf{\Sigma}_\pm = \tfrac{1}{d}\mathbb{E}[\sigma^2]\boldsymbol{I}_d$$

Using

$$\varrho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \exp\left(-\frac{c}{\sigma^2}\right) \quad \text{with} \quad \frac{c}{a-1} = \mathbb{E}[\sigma^2]$$

- **No “Gaussian universality”**: training error, training loss, test error all depend on the tail exponent a although matching first and second moments.
- **With square loss**, optimal performance for finite λ, with $\lambda \to +\infty$ in the limit of Gaussian clouds.
  Mignacco et al. (2020); Baldassi et al. (2020).
- **No obvious performance ordering**: using a Gaussian equivalent setting is beneficial for large α.

Square loss with $\lambda = 10^{-5}$.
Test data with $a = c + 1 = 2$.



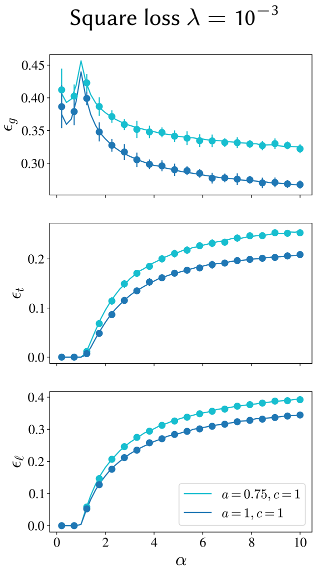
Very heavy-tailed classification

We can also consider two clouds having **no covariance**

$$\mathbb{E}[\sigma^2] = +\infty$$

Clouds have tail decay $\|\boldsymbol{x}\|^{-2a}$ in the radial direction with $a \in (0, 1]$.

In this case, heavier tail gives worse performances.

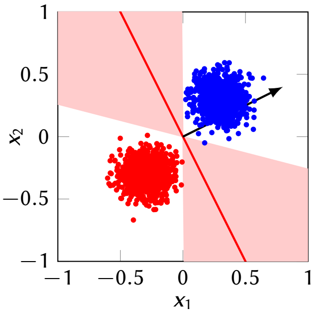


Tail effects on the separability transition

If we consider logistic loss at zero regularisation, this is equivalent to search for the *max-margin hyperplane* as

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{\nu=1}^{n}\ln\left(1+\mathrm{e}^{-y_\nu\boldsymbol{\theta}^\intercal\boldsymbol{x}_\nu}\right)+\lambda\|\boldsymbol{\theta}\|^2\right]$$

$$\xrightarrow[\lambda\to 0^+]{\boldsymbol{\theta}\to\lambda^{-1/2}\boldsymbol{\theta}} \arg\min_{\boldsymbol{\theta}}\sum_{\nu=1}^{n}\max\left\{0,-y_\nu\boldsymbol{\theta}^\intercal\boldsymbol{x}_\nu\right\}.$$



- In the *single* Gaussian case Sur and Candès (2019) has shown that there is a *phase transition* in $\alpha = {}^n/_d$: points separable for $\alpha < \alpha^\star$. In the limit of infinite covariance Cover (1965) had shown that $\alpha^\star = 2$.
- Explicit formula by Mignacco *et al.* (2020) for the separability of $K = 2$ clusters and implicit analytical criterion for the generic case of K Gaussian clusters by Loureiro *et al.* (2020).

Existence of the MLE

In our setting, separability is possible iff

$$\boxed{\alpha \le \max_{\theta\in(0,1]} \frac{1-\theta^2}{\mathcal{S}_\star(\theta)} =: \alpha^\star}$$

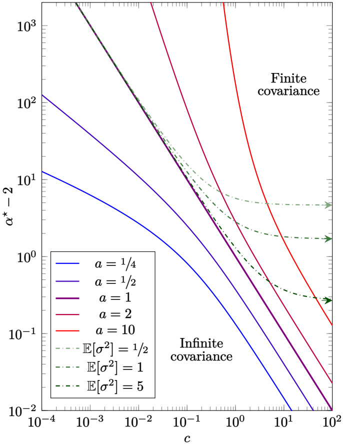
where

$$\mathcal{S}_\star(\theta) = \int_0^\infty z^2 \mathbb{E}\left[\mathcal{N}\left(z + \tfrac{\theta}{\sigma}; 0, 1\right)\right] \mathrm{d}\, z.$$

Using

$$\rho(\sigma) \propto \frac{1}{\sigma^{2a+1}}\, \mathrm{e}^{-c/\sigma^2}$$

- at given finite variance, the Gaussian threshold value is a *lower bound*.



For $a = 1$

$$\alpha^\star = 2 + \frac{1}{c}$$

Existence of the MLE

In our setting, separability is possible iff

$$\boxed{\alpha \leq \max_{\theta \in (0,1]} \frac{1-\theta^2}{\mathcal{S}_\star(\theta)} =: \alpha^\star}$$

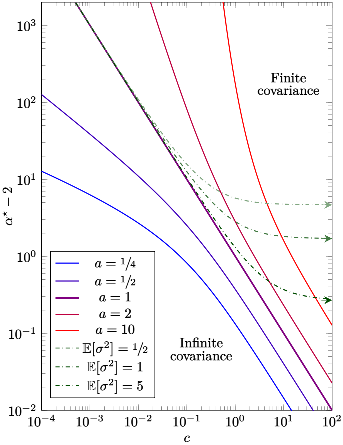
where

$$\mathcal{S}_\star(\theta) = \int_0^\infty z^2 \mathbb{E}\left[\mathcal{N}\left(z + \tfrac{\theta}{\sigma}; 0, 1\right)\right] \mathrm{d}\, z.$$

Using

$$\rho(\sigma) \propto \frac{1}{\sigma^{2a+1}} \,\mathrm{e}^{-c/\sigma^2}$$

- at given finite variance, the Gaussian threshold value is a *lower bound*.
- in the limit of infinite width $\alpha^\star \to 2$.



For $a = 1$

$$\alpha^\star = 2 + \frac{1}{c}$$

An extreme deconstruction: random labels

Suppose now that we assign the labels to our points **completely randomly**. Some **universality** emerges by effect of the lack of correlations label/structure.

With Gaussian clouds:

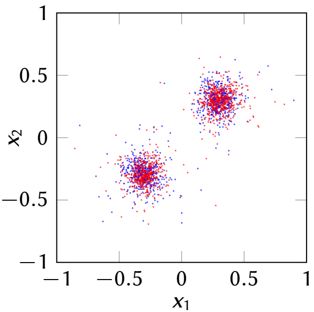
- Equivalent to single Gaussian cloud

$$P(\boldsymbol{x}) \approx \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{0}, \frac{\sigma^2}{d}\boldsymbol{I}_d\right).$$

- Using square loss and *random labels*, universal training loss for $\lambda \to 0$

$$\epsilon_\ell = \frac{1}{2}\left(1 - \frac{1}{\alpha}\right)_+.$$

Gerace *et al.* (2022); Pesce *et al.* (2023)



Why random labels?

- Adopted in capacity calculations by e.g. Gardner and Derrida (1989) and Vapnik (1989)
- Zhang *et al.* (2021) used the setting as a reference for worst-case analysis and in the study of training time vs training with informative labels.

An extreme deconstruction: random labels

Suppose now that we assign the labels to our points **completely randomly**. Some **universality** emerges by effect of the lack of correlations label/structure.

With heavy-tailed clouds:

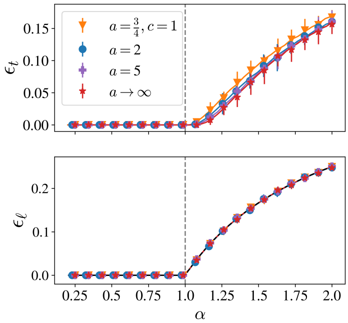
- ✗ Equivalent to single GSM cloud

$$P(\boldsymbol{x}) \approx \mathbb{E}\left[\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{0}, \frac{\sigma^2}{d}\boldsymbol{I}_d\right)\right].$$

- ✓ Using square loss and *random labels*, universal training loss for $\lambda \to 0$

$$\epsilon_\ell = \frac{1}{2}\left(1 - \frac{1}{\alpha}\right)_+ .$$

For $a > 1$, $a = c + 1$ so that $\boldsymbol{\Sigma} = {}^1\!/_d \boldsymbol{I}_d$. Square loss with $\lambda = 10^{-4}$ on random labels for $K = 2$ clouds vs prediction for $K = 1$ cloud.



Regression with heavy tails

with Urte Adomaityte, Bruno Loureiro, Leonardo Defilippis

Regression with heavy tails

Consider now a dataset $\mathcal{D} = \{(y_\nu, \boldsymbol{x}_\nu)\}_{\nu\in[n]}$ generated via a *linear model*

$$y = \boldsymbol{\theta}_0^\intercal \boldsymbol{x} + \sqrt{\Delta}\eta, \qquad \eta \sim \mathcal{N}(0,1), \quad \Delta > 0$$

where again

$$\boxed{\boldsymbol{x} \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{d}}\sigma\boldsymbol{z}}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \qquad \boxed{\sigma \sim \varrho(\sigma)}$$

Regression with heavy tails

Consider now a dataset $\mathcal{D} = \{(y_\nu, \boldsymbol{x}_\nu)\}_{\nu\in[n]}$ generated via a *linear model*

$$y = \boldsymbol{\theta}_0^\intercal \boldsymbol{x} + \sqrt{\Delta}\eta, \qquad \eta \sim \mathcal{N}(0,1), \quad \Delta > 0$$

where again

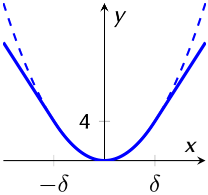
$$\boxed{\boldsymbol{x} \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{d}}\sigma \boldsymbol{z}}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \qquad \boxed{\sigma \sim \varrho(\sigma)}$$

Useful set-up to model a *variance contamination*,

$$\varrho(\sigma) = (1-\varepsilon)\underbrace{\delta(\sigma-\sigma_0)}_{\text{Gauss}} + \varepsilon\underbrace{\varrho_c(\sigma)}_{\text{contamination}}, \qquad \varepsilon \in [0,1].$$

To mitigate the presence of the contamination Huber (1965) proposed a differentiable, robust loss:

$$|y-\boldsymbol{\theta}^\intercal\boldsymbol{x}|_\delta := \begin{cases} (y-\boldsymbol{\theta}^\intercal\boldsymbol{x})^2 & \text{if } |y-\boldsymbol{\theta}^\intercal\boldsymbol{x}|<\delta, \\ 2\delta|y-\boldsymbol{\theta}^\intercal\boldsymbol{x}|-\delta^2 & \text{if } |y-\boldsymbol{\theta}^\intercal\boldsymbol{x}|\geq\delta. \end{cases}$$



Asymptotic properties for regression on GSMs studied by El Karoui et al. (2013, 2018) under the assumption

$$\mathbb{E}[\sigma^4] < +\infty.$$

Regression with heavy tails

Consider now a dataset $\mathcal{D}=\{(y_\nu,\boldsymbol{x}_\nu)\}_{\nu\in[n]}$ generated via a *linear model*

$$y=\boldsymbol{\theta}_0^\intercal\boldsymbol{x}+\sqrt{\Delta}\eta,\qquad \eta\sim\mathcal{N}(0,1),\quad \Delta>0$$

where again

$$\boxed{\boldsymbol{x}\stackrel{\mathrm{d}}{=}\frac{1}{\sqrt{d}}\sigma\boldsymbol{z}},\qquad \boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d),\qquad \boxed{\sigma\sim\varrho(\sigma)}$$

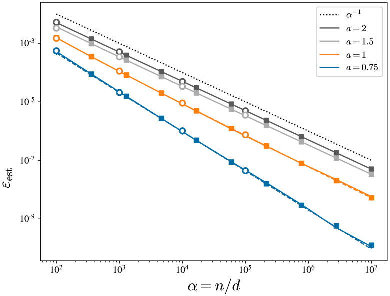
MSE rates for Huber loss

If $\varrho(\sigma)\sim\frac{1}{\sigma^{2a+1}}$ for $\sigma\gg1$, $\forall\delta$

$$\lim_d\frac{\|\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}\|^2}{\Delta}\underset{\alpha\gg1}{=}\begin{cases}\frac{1}{\sigma_0^2\alpha}+o\left(\frac{1}{\alpha}\right) & \text{if } a>1,\\ \frac{1}{\sigma_0^2\alpha\ln\alpha}+o\left(\frac{1}{\alpha\ln\alpha}\right) & \text{if } a=1,\\ \frac{1}{(\sigma_0^2\alpha)^{1/a}}+o\left(\frac{1}{\alpha^{1/a}}\right) & \text{if } a\in(0,1).\end{cases}$$

If $\delta\to+\infty$ (square loss)

$$\sigma_0^2=\lim_{x\to+\infty}\left(1-\mathbb{E}\left[\frac{x}{x+\sigma^2}\right]\right)\frac{x^{\min\{1,a\}}}{(\ln x)^{\delta_{a,1}}}.$$



Square loss (dashed) and optimal Huber (continuous) vs experiments (circles). Squares are the Bayes optimal bound.

Conclusions

Conclusions and perspectives

- "Doubly-random" models can be useful to study non-Gaussianity — see e.g., the recent work of Székely et al. (2024) on spiked models.
  GSMs are in particular a good theoretical setup for heavy tails, robustness and to go beyond the "Gaussian shell" geometry.
- Heavy tails **"break"** some recently found universality laws even in the simplest possible set-ups (convex ERM).
- Separability transitions/rates are affected by power-law tails.

Conclusions and perspectives

- "Doubly-random" models can be useful to study non-Gaussianity — see e.g., the recent work of Székely et al. (2024) on spiked models.
  GSMs are in particular a good theoretical setup for heavy tails, robustness and to go beyond the "Gaussian shell" geometry.
- Heavy tails **"break"** some recently found universality laws even in the simplest possible set-ups (convex ERM).
- Separability transitions/rates are affected by power-law tails.

Some open questions.

- What would happen using random features?
- What is the effect of fat tails (i.e., *outliers*) in fairness models?
- Rigorous proofs?
  Work in progress with Cédric Gerbelot.
- What about the dynamics?
  Building on results of Ben Arous, Bruna et al. (2023) have shown that the Gaussian picture is preserved in the GSM setting if $\mathbb{E}[\sigma^4] < \infty$: a proper *information exponent* determines the *out-of-mediocrity timescale* in online SGD.
  Also, work in progress with Urte Adomaityte, Bruno Loureiro and Pierfrancesco Urbani.

Thank you for your attention.

References

Adomaityte, Sicuro, Vivo — NeurIPS 2023 [arXiv:2304.02912]
Adomaityte, Defilippis, Loureiro, Sicuro — JSTAT 2024 [arXiv:2309.16476]