Concentration of the Measure in Machine Learning

Cosme Louart

Assistant Professor at CUHK Shenzhen

SCHOOL OF DATA SCIENCE

Concentration of the Measure in Machine Learning

Cosme Louart

Assistant Professor at CUHK Shenzhen

SCHOOL OF DATA SCIENCE

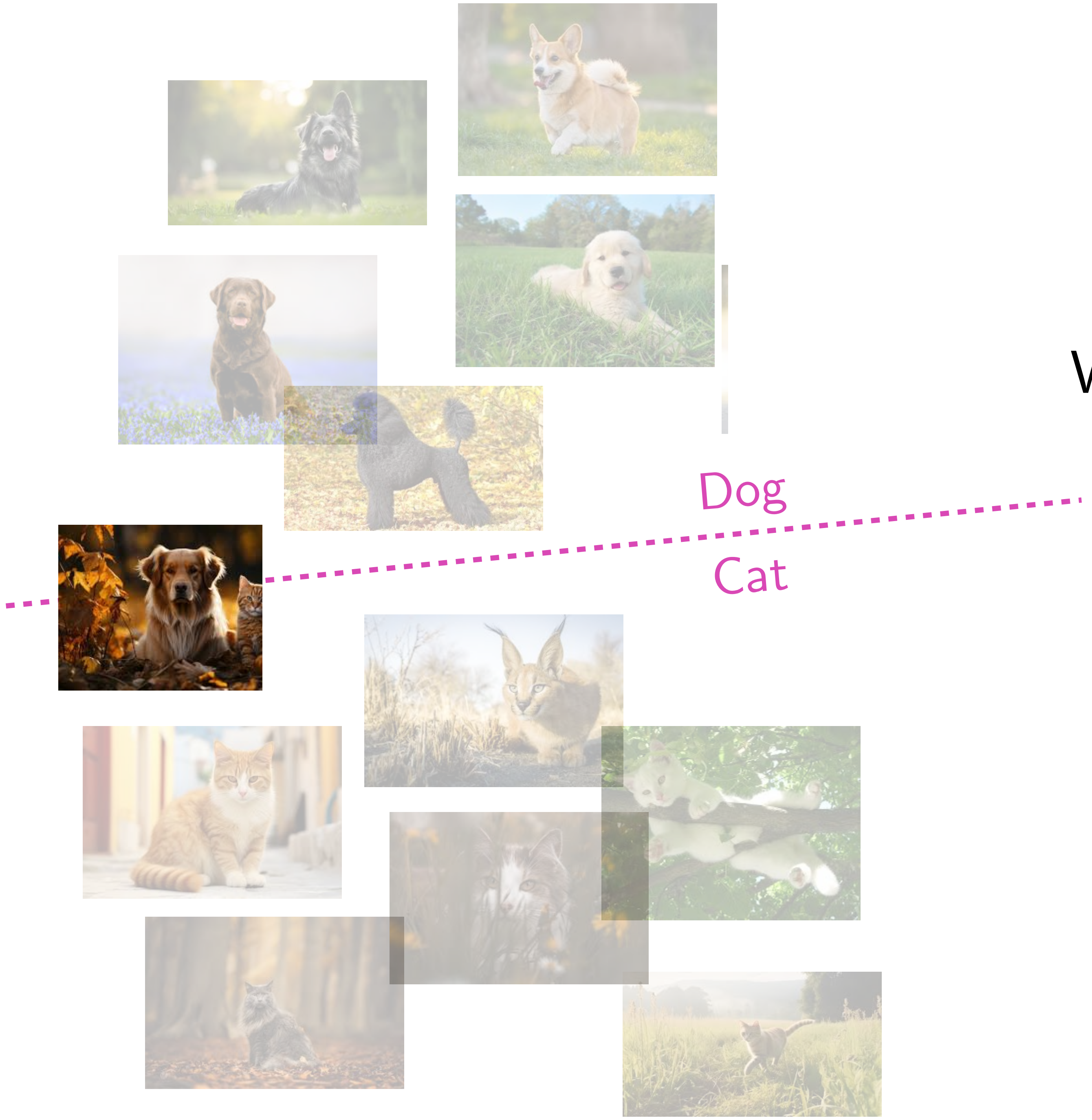
Content

I - **Motivation:** Probability in Machine Learning

"Almost all of machine learning can be viewed in probabilistic terms, making probabilistic thinking fundamental. It is, of course, not the only view. But it is through this view that we can connect what we do in machine learning to every other computational science, whether that be in stochastic optimisation, control theory, operations research, econometrics, information theory, statistical physics or bio-statistics. For this reason alone, mastery of probabilistic thinking is essential."

Shakir Mohamed, DeepMind

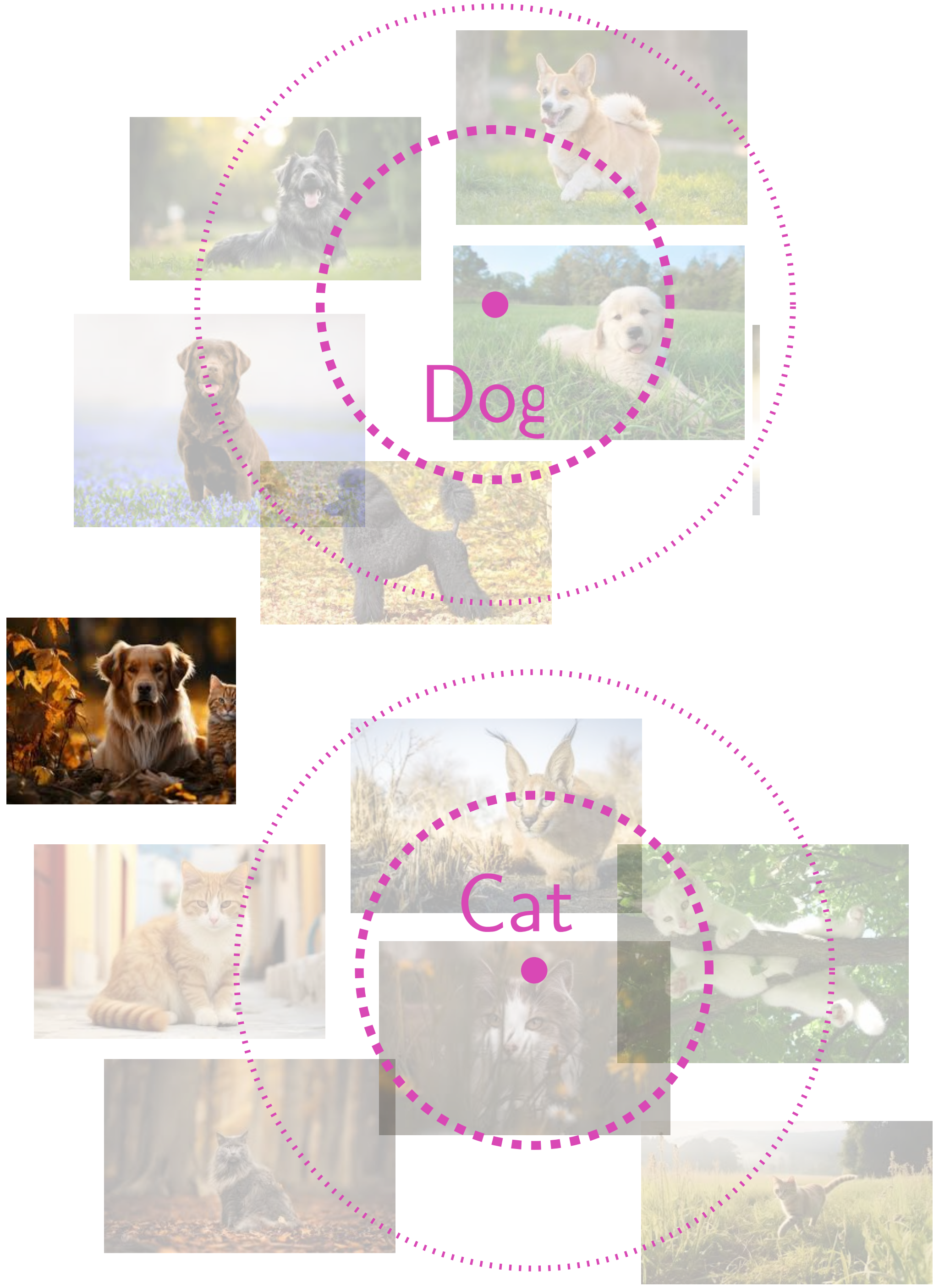
Geometric vs. Probabilistic view on data



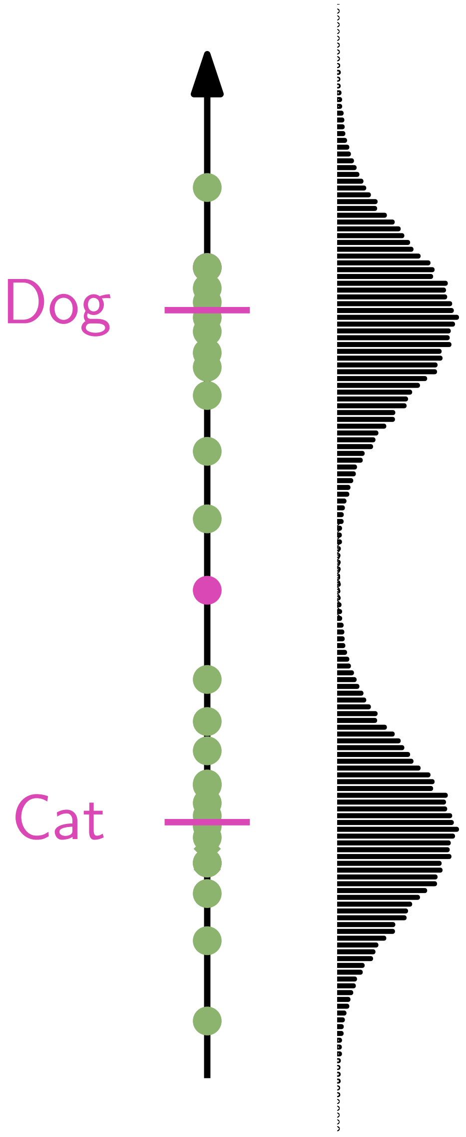
Who is dog/cat ?

→ ??

Geometric vs. Probabilistic view on data



Who is dog/cat ?

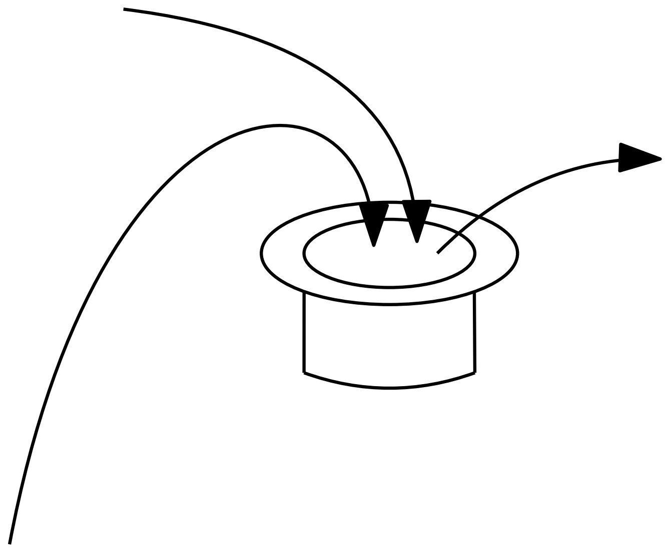


B - Dimension of input / Dimension of output

Requirements

Ex 1: Unsupervised learning

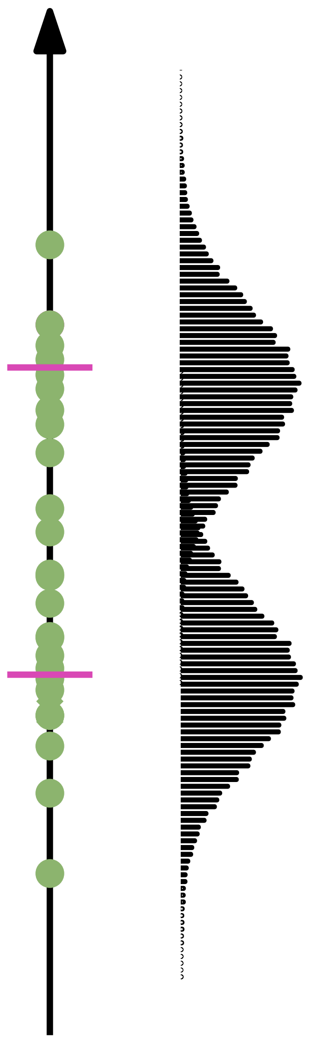
Class 1



Class 2

$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$:
Data random matrix

AI(X): random variable



$Y = (y_1, \ldots, y_n) \in \{-1, 1\}^n$

Estimate:

$\mathbb{E}[\|\text{AI}(X) - Y\|^2]$

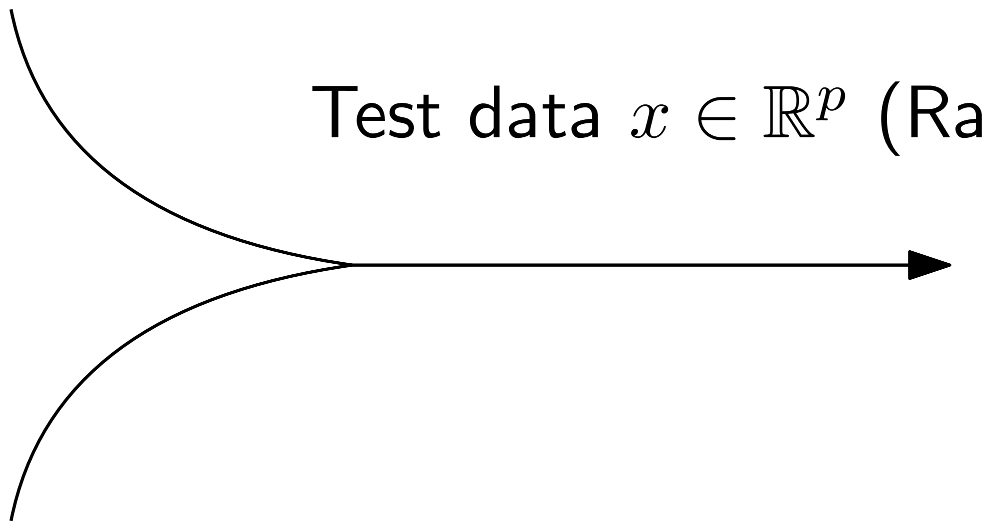
$\mathbb{P}(|\text{AI}(x_i) - y_i| \geq 1)$

Requirements

Ex 2: Supervised learning

Class 1

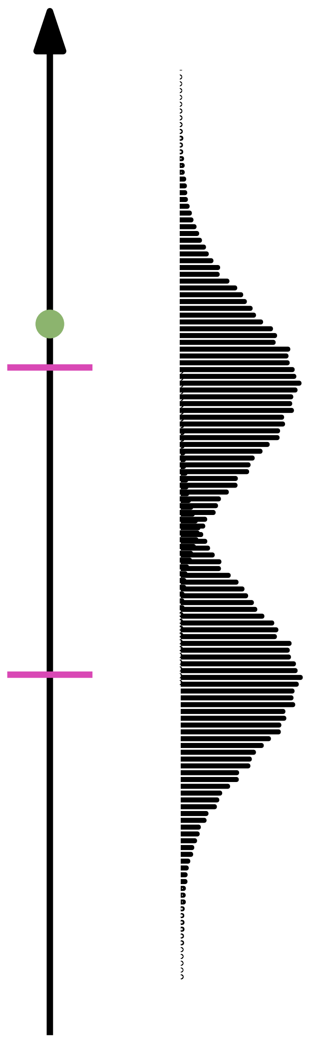
Test data $x \in \mathbb{R}^p$ (Random)



Class 2

$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$:
Training data (**random**)

AI$_X(x)$ (random)



$y \in \{-1, 1\}$: Label of x

Estimate:
$\mathbb{E}[\|\text{AI}_X(x) - y\|^2]$
$\mathbb{P}(|\text{AI}_X(x) - y| \geq 1)$

Conclusion

$X \mapsto \mathrm{AI}_X$ and $x \mapsto \mathrm{AI}_X(x)$

non linear

Test data $x \in \mathbb{R}^p$ (Random)

"Concentrated vectors"

Advantages:

- Larger hypothesis
- Flexible with non-linearities

AI_X

Ran

$\mathrm{AI}_X(x)$ (random)

$\mathrm{AI}_X(X)$ Concentrated as X, x!

$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$:

Training data (**random**)

II - Theory: Concentration of the measure

Theorem: Given $Z \sim \mathcal{N}(\mu, I_n)$, $\forall f : \mathbb{R}^n \to \mathbb{R}$, 1-Lipschitz:

$$\mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq t\right) \leq 2e^{-\frac{t^2}{2}}$$

Given $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ λ-Lipschitz and $f : \mathbb{R}^n \to \mathbb{R}$ 1-Lipschitz:

Application: $Z = (X, x)$ and $\Phi(Z) = \text{AI}_X(x)$

If $(X, x) \mapsto \text{AI}_X(x)$ C-Lipschitz:

$$\mathbb{P}\left(|\text{AI}_X(x) - \mathbb{E}[\text{AI}_X(x)]| \geq t\right) \leq 2e^{-\frac{t^2}{2}}$$

Theorem: **(Talagrand)**
Given $Z = (Z_1, \dots Z_n) \in [0,1]^n$ s.t. $Z_1, \dots, Z_n$ independent
$\forall f : \mathbb{R}^p \to \mathbb{R}$, 1-Lipschitz and convex:

$$\mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq t\right) \leq 2e^{-\frac{t^2}{4}}.$$

Michel Ledoux (2005) *The concentration of measure phenomenon.* vol. 89. Mathematical Surveys and Monographs. Providence, Rhode Island: American Math- ematical Society, page 181.

From Gaussian to realistic Hypothesis

Theorem: Given $Z^{(n)} \sim \mathcal{N}(\mu, I_n)$, , $\exists C, c > 0$, $\forall n \in \mathbb{N}$, $\forall f : \mathbb{R}^n \to \mathbb{R}$, 1-Lipschitz:

$$\mathbb{P}\left(\left|f(Z^{(n)}) - \mathbb{E}[f(Z^{(n)})]\right| \geq t\right) \leq Ce^{-ct^2}$$

$\longleftrightarrow$ $Z \propto \mathcal{E}_2$

Our "Gaussian like" setting

$\forall n, p \in \mathbb{N}$ $X^{(n,p)} \in \mathbb{R}^{n \times p}$

General hypothesis:

$\exists C, c > 0$ s.t. $\forall n, p \in \mathbb{N}$, $\forall f : \mathbb{R}^{n \times p} \to \mathbb{R}$ 1-Lipschitz:

$$\mathbb{P}\left(|f(X^{(n,p)}) - \mathbb{E}[f(X^{(n,p)})]| \geq t\right) \leq Ce^{-ct^2}$$

$\longleftrightarrow$ **Notation:** $X \propto \mathcal{E}_2$

In practice:

$$X, x \propto \mathcal{E}_2(\sigma) \implies \mathrm{AI}_X(x) \propto \mathcal{E}_2(\sigma)$$

From Gaussian to Realistic Hypothesis

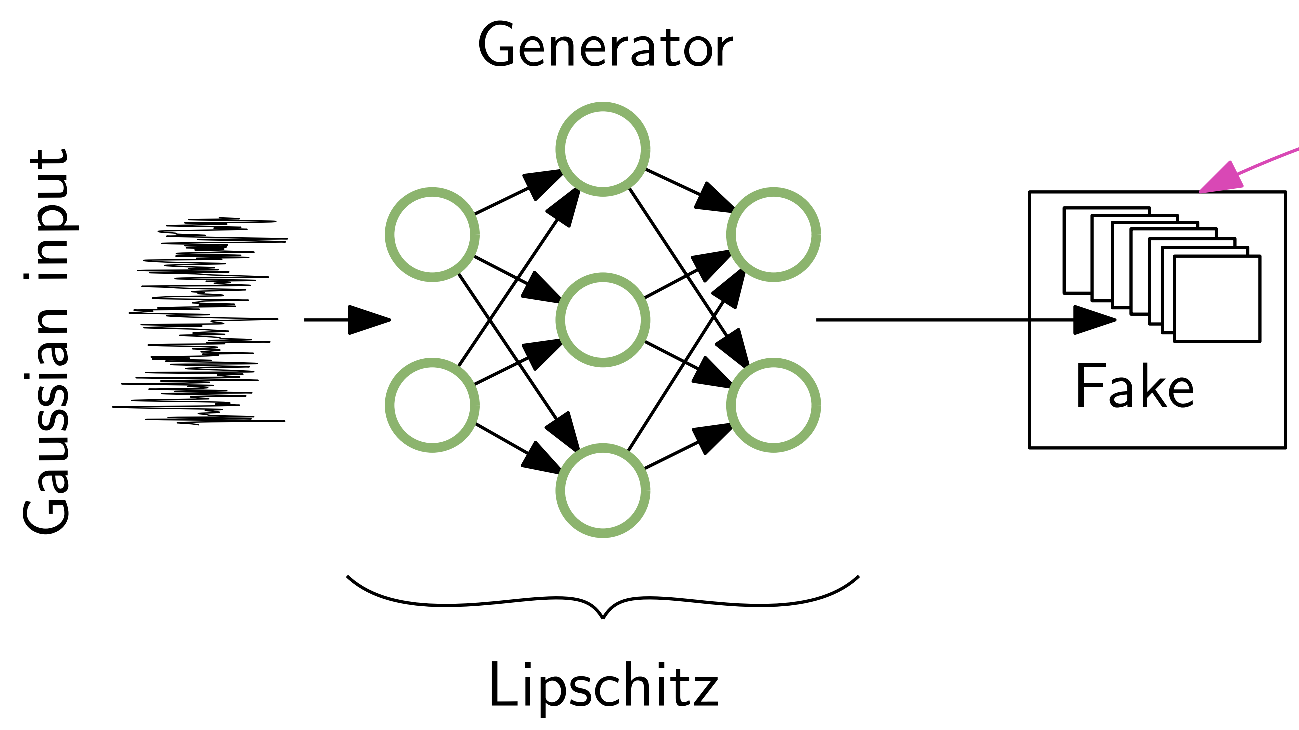
Theorem: Given $Z^{(n)} \sim \mathcal{N}(\mu, I_n)$,

$$Z \propto \mathcal{E}_2$$

Recall: $\forall \Phi : \mathbb{R}^n \to \mathbb{R}^q$ C-Lipschitz:

$$\Phi(Z) \propto \mathcal{E}_2$$

GAN generated images are concentrated vectors



Concentration of measure tools

Lemma:

$$X \propto \mathcal{E}_2(\sigma) \iff \begin{array}{c}\exists C > 0 \text{ s.t. } \forall n \in \mathbb{N}, \forall f : \mathbb{R}^n \to \mathbb{R} \text{ 1-Lipschitz:} \\ \mathbb{E}[|f(X^{(n)}) - \mathbb{E}[f(X^{(n)})]|^r] \leq C(\frac{r}{2})^{\frac{r}{2}} {\sigma_n}^r\end{array}$$

Consequence: σ measures the moments

For random variables:

$Z^{(n)}$: random variable, $\bar{Z}^{(n)}$: scalar.

$$\forall n \in \mathbb{N}, \forall t \geq 0 : \mathbb{P}(|Z^{(n)} - \bar{Z}^{(n)}| \geq t) \leq Ce^{-c(t/\sigma_n)^2} \qquad \longleftrightarrow \qquad Z \in \bar{Z} \pm \mathcal{E}_2(\sigma)$$

Lemma:

$$\begin{cases} Z_1 \in \bar{Z}_1 \pm \mathcal{E}_2(\sigma_1) \\ Z_2 \in \bar{Z}_2 \pm \mathcal{E}_2(\sigma_2) \end{cases} \iff \begin{cases} Z_1 + Z_2 \in \bar{Z}_1 + \bar{Z}_2 \pm \mathcal{E}_2(\sigma_1 + \sigma_2) \\ Z_1 \cdot Z_2 \in \bar{Z}_1 \cdot \bar{Z}_2 \pm \mathcal{E}_2(|\bar{Z}_1| \cdot \sigma_2 + \sigma_1 \cdot |\bar{Z}_2|) + \mathcal{E}_1(\sigma_1 \cdot \sigma_2) \end{cases}$$

Control of the norm

Infinite norm (Given $z \in \mathbb{R}^n$: $\|z\|_\infty = \max_{1\leq i\leq n} |z_i|$)

Given $Z^{(n)} \in \mathbb{R}^n$, $Z \propto \mathcal{E}_2(\sigma)$:

$$\begin{aligned}\mathbb{P}\left(\|Z - \tilde{Z}\|_\infty \geq t\right) &= \mathbb{P}\left(\sup_{1\leq i\leq p} e_i^T(Z - \tilde{Z}) \geq t\right) \\ &\leq p \sup_{1\leq i\leq p} \mathbb{P}\left(e_i^T(Z - \tilde{Z}) \geq \right. \\ &\leq Ce^{\log p - (t/c\sigma)^2} \leq C'e^{-(t/}\end{aligned}$$

III - **Application:** Regression

Classification problem with two classes.

2 laws in $\mathbb{R}^p$: $\mathcal{C}_+$; $\mathcal{C}_-$

$X = (x_1, \ldots, x_n) \in \mathcal{M}_{p,n}$: data matrix, $x_i \sim \mathcal{C}_+$ or $x_i \sim \mathcal{C}_-$

notation: $\mu_\pm = \mathbb{E}[x_i]$, $\Sigma_\pm = \mathbb{E}[x_i x_i^T]$, for $x_i \sim \mathcal{C}_\pm$

$Y \in \{-1, 1\}^n$: label vector $x_i \sim \mathcal{C}_\pm \Rightarrow y_i = \pm 1$

Regression problem

Ridge Regression:

$$\text{Minimise } \frac{1}{n} \left\| \beta^T X - Y \right\|^2 + \gamma \left\| \beta \right\|^2$$

γ : regularising parameter

Robust Regression:

$$\text{Minimise } \frac{1}{n} \sum_{i=1}^{n} f(y_i \beta^T x_i) + \gamma \|\beta\|^2$$

$f : \mathbb{R} \to \mathbb{R}$: loss function

Ridge Regression

Minimise $\frac{1}{n}\sum_{i=1}^{n}(\beta^T x_i - y_i)^2 + \gamma \|\beta\|^2$.

Solution : $\beta = \frac{1}{n} QXY$ with $Q = \left(\frac{1}{n}XX^T + \gamma I_p\right)^{-1}$.

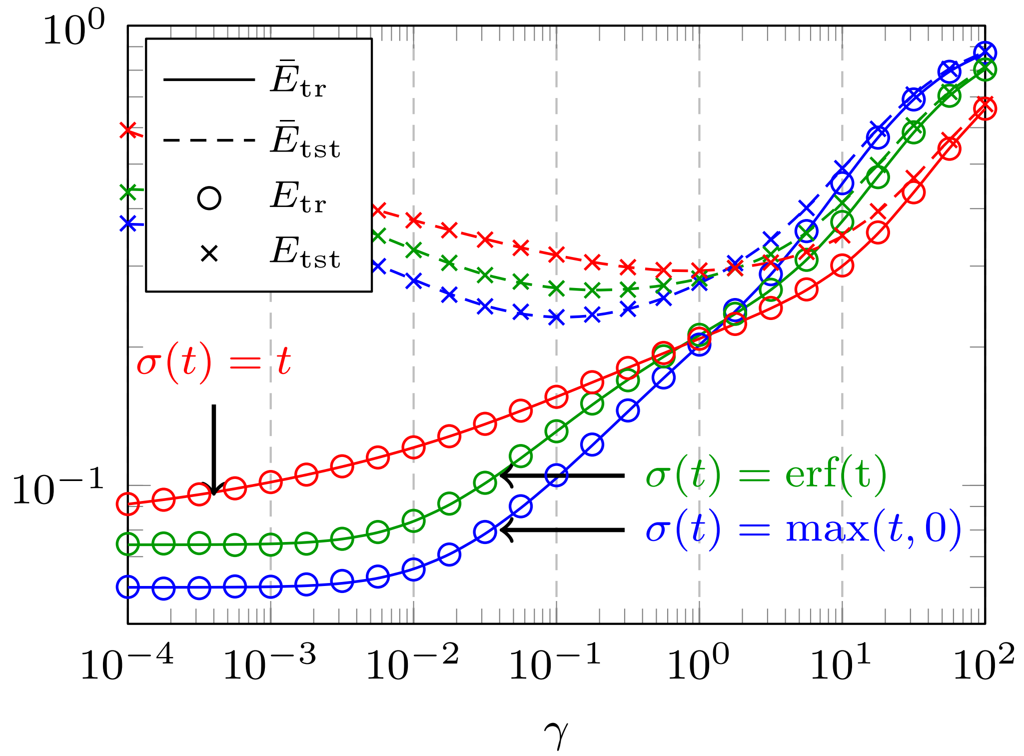
Example with One-Layer Neural Net $X = \sigma(WZ)$

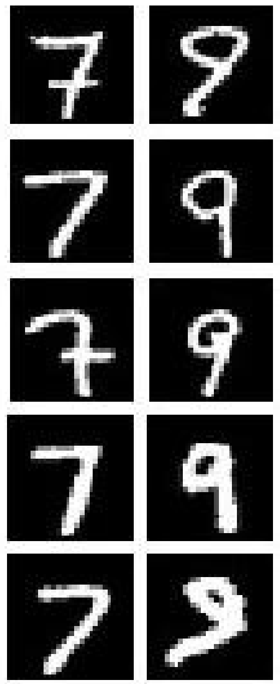
- $Z = (z_1, \dots z_n) \in \mathbb{R}^{q\times n}$, MNIST data
- $W \in \mathcal{M}_{p,q}$, fixed initial drawing
- $\sigma : \mathbb{R} \to \mathbb{R}$: Lipschitz activation function

Performance estimation:

Training error: $E_{\text{tr}} = \frac{1}{n}\|X^T\beta - Y\|^2$

$$\bar{E}_{\text{tr}} = \frac{1}{n}\mathbb{E}\left[\left\|\frac{1}{n}X^T QXY - Y\right\|^2\right] = f^\circ(\tilde{Q}) \approx f^\circ(\Sigma_\pm, \mu_\pm).$$





Test error: $E_{\text{tst}} = \frac{1}{n}\|X_t^T\beta - Y\|^2$, X_t, X i.i.d

$$\bar{E}_{\text{tst}} = \frac{1}{n}\mathbb{E}\left[\frac{1}{n}YXQX_tX_t^TQXY - 2Y^TX_t^TQXY + Y^TY\right] \approx f^\circ(\Sigma_\pm, \mu_\pm).$$

$\longrightarrow$ As if $x_1, \dots, x_n$ were Gaussian!

Robust Regression

Minimize $\frac{1}{n}\sum_{i=1}^{n} f(y_i x_i^T \beta) + \gamma \|\beta\|^2, \quad \beta \in \mathbb{R}^p$

Solution : $\beta = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i^T \beta) z_i$ with $z_i = y_i x_i$ and $\phi = -\frac{1}{2\gamma} f'$

Theorem Assume that:

- $X \propto \mathcal{E}_2$
- $\phi : \mathbb{R} \to \mathbb{R}$ is λ-Lipschitz bounded
- $\gamma > \frac{1}{\sqrt{n}} \lambda \|\mathbb{E}[X]\|^2$

↔

- $X \propto \mathcal{E}_2$
- $\phi : \mathbb{R} \to \mathbb{R}$ is convex

Then β is uniquely defined and:

$$\beta \propto \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \qquad \text{and} \qquad \mathbb{E}[\|\beta\|] = O(1).$$

$\to$ Estimation of $\mathbb{E}[\beta]$ and $\mathbb{E}[\beta\beta^T]$ to predict performances

Estimate $\mathbb{E}[\beta]$

New formulation: $\beta = \frac{1}{n}\sum_{i=1}^n \phi(z_i^T\beta)z_i \implies \mathbb{E}[\beta] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi(z_i^T\beta)z_i]$

Problem: Dependence between z_i and β

Solution: Leave-one-out : $\beta_{-i} = \frac{1}{n}\sum_{j\neq i}\phi(z_j^T\beta_{-i})z_j$

Leave-one-out method: Find a relation between β and β_{-i}.

Progressively remove contribution of z_i, $i \in [n]$, $\forall t \in [0,1]$:

$$\beta_{-i}(t) = \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^n \phi(z_j^T\beta_{-i}(t))z_j + \frac{t}{n}\phi(z_i^T\beta_{-i}(t)$$

$\implies \quad \beta = \beta_{-i}(1)$ and $\beta_{-i} \equiv \beta_{-i}(0)$ independent of z_i.

Strategy:

(a) Differentiation
(b) Approximation (thanks to concentration of measure tools)
(c) Integration

$$\forall t \in [0,1]: \qquad \beta_{-i}(t) = \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \phi(z_j^T \beta_{-i}(t)) z_j + \frac{1}{n} \underbrace{t\phi(z_i^T \beta_{-i}(t))}_{\chi_i(t)} z_i .$$

(a) Differentiation:

$$\beta'_{-i}(t) = \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \underbrace{\phi'(z_j^T \beta_{-i}(t))}_{D_j^{(i)}(t)} z_j z_j^T \beta'_{-i}(t) + \frac{1}{n}\chi'_i(t) z_i$$

$$= \frac{1}{n} Z_{-i} D_{-i}(t) Z_{-i}^T \beta'_{-i}(t) + \frac{1}{n}\chi'_i(t) z_i$$

$$= \frac{1}{n}\chi'_i(t) Q_{-i}(t) z_i$$

With: $D_{-i}(t) \equiv \text{Diag}(D_1^{(i)}(t), \ldots, D_n^{(i)}(t))$ $\qquad Q_{-i}(t) \equiv \left(I_p - \frac{1}{n} Z_{-i} D_{-i}(t) Z_{-i}^T\right)^{-1}$

$Z_{-i} \equiv (z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n)$

$$\forall t \in [0,1]: \qquad \beta_{-i}(t) = \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \phi(z_j^T \beta_{-i}(t)) z_j + \frac{1}{n} \underbrace{t\phi(z_i^T \beta_{-i}(t))}_{\chi_i(t)} z_i.$$

(b) Approximation:

$$\beta'_{-i}(t) = \frac{1}{n}\chi'_i(t) Q_{-i}(t) z_i$$

Proposition: $\|Q_{-i}(t)z_i - Q_{-i}(0)z_i\| \leq O(1)$

Consequence: $\frac{1}{n} z_i^T Q(t) z_i \in \delta \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) + \mathcal{E}_1\left(\frac{1}{n}\right)$

With: $D_{-i}(t) \equiv \mathrm{Diag}(D_1^{(i)}(t), \ldots, D_n^{(i)}(t))$

$$Z_{-i} \equiv (z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n)$$

$$Q_{-i}(t) \equiv \left(I_p - \frac{1}{n} Z_{-i} D_{-i}(t) Z_{-i}^T\right)^{-1}$$

$$\delta \equiv \mathbb{E}\left[\frac{1}{n} z_i^T Q(0) z_i\right].$$

$$\forall t \in [0,1]: \qquad \beta_{-i}(t) = \frac{1}{n}\sum_{\substack{j=1\\j\neq i}}^{n} \phi(z_j^T\beta_{-i}(t))z_j + \frac{1}{n}\underbrace{t\phi(z_i^T\beta_{-i}(t))}_{\chi_i(t)} z_i.$$

(c) Integration:

$$\beta'_{-i}(t) = \frac{1}{n}\chi'_i(t)Q_{-i}(t)z_i \qquad \text{and} \qquad \tfrac{1}{n}z_i^T Q(t)z_i \in \delta \pm \mathcal{E}_2\left(\tfrac{1}{\sqrt{n}}\right) + \mathcal{E}_1\left(\tfrac{1}{n}\right)$$

Link between β and β_{-i}!

$$\implies z_i^T\beta - z_i^T\beta_{-i} = \textstyle\int_0^1 z_i^T\beta'_{-i}(t)dt \approx \delta\int_0^1 \chi'_i(t)dt = \delta\phi(z_i^T\beta).$$

Proposition: $\forall u \in \mathbb{R}, \exists! \xi(u) \mid \xi(u) = u + \delta\phi(\xi(u))$.

Proposition: $z_i^T\beta \in \xi(z_i^T\beta_{-i}) \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right).$

With: $D_{-i}(t) \equiv \text{Diag}(D_1^{(i)}(t), \ldots, D_n^{(i)}(t))$

$Z_{-i} \equiv (z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n)$

$$Q_{-i}(t) \equiv \left(I_p - \frac{1}{n}Z_{-i}D_{-i}(t)Z_{-i}^T\right)^{-1}$$

$$\delta \equiv \mathbb{E}\left[\frac{1}{n}z_i^T Q(0)z_i\right].$$

Proposition: $\forall t \in \mathbb{R}, \exists! \xi(t) \mid \xi(t) = t + \delta\phi(\xi(t))$.

Proposition: $z_i^T \beta \in \xi(z_i^T \beta_{-i}) \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right)$.

Estimation of the statistics of β

$$\mathbb{E}[\beta] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi(z_i^T \beta) z_i] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi(\xi(z_i^T \beta_{-i})) z_i]$$

Conjecture: $\beta \sim \mathcal{N}(m_\beta, C_\beta)$

$$\Longrightarrow \quad z_i^T \beta_{-i} \sim \mathcal{N}(m_z^T m_\beta, \mathrm{Tr}(C_z C_\beta) + m_\beta^T C_z m_\beta)$$

$\Longrightarrow$ Can use Stein formulas to compute m_β and C_β.

Mai, Liao, Couillet - A Large Scale Analysis Of Logistic Regression: Asymptotic Performances And New Insights

- $p = 128,\ n = 512$
- $x_i \propto \mathcal{N}(y_i\mu, \Sigma)$
- $\Sigma = 2I_p$
- $\Sigma' = \mathrm{diag}[1,\ 5,\ \mathbf{1}_{p-2}]$

Mai, Liao, Couillet - A Large Scale Analysis Of Logistic Regression: Asymptotic Performances And New Insights