Grenoble | images | parole | signal | automatique | laboratoire

The asymptotic performance of DNN's seen from the softmax layer: a random matrix and concentration-of-measure approach

Cosme Louart, Romain Couillet, Mohamed El Amine Seddik

GIPSA-lab, INP Grenoble; List-CEA

21/01/2020

UMR 5216

Robust Regression algorithm

- Data matrix $X = (x_1, \ldots, x_n) \in \mathcal{M}_{p,n}$,
- labels : $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- Robust regression problem[1,2] with regularizing parameter:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) + \lambda \|\beta\|^2$$

with $\rho : \mathbb{R} \to \mathbb{R}$ convex, $\lambda > 0$.

- Score of a new data $x \in \mathbb{R}^p$: $\beta^T x$

Performance: $\mathbb{E}_{X,x}[\rho(\beta^T x - y_x)]$
Goal: Understand the statistics of $\beta = f(X)$.

---

[1] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. Proceed- ings of the National Academy of Sciences, 2013.

[2] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logis- tic regression: Asymptotic performance and new insights. In ICASSP'19

Setting and conclusion

Concentration hypotheses on the data X

- For all 1-Lipschitz maps $f : \mathcal{M}_{p,n} \to \mathbb{R}$:

$$\forall t > 0 : \quad \mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq Ce^{-ct^2}$$

- $x_1, \ldots, x_n$ are independent

Assets

- **(Representativity)** True if the columns are Lipschitz transformation of a Gaussian vector $Z \sim \mathcal{N}(0, I_p)$.
  $\longrightarrow$ dependence between entries of a column possibly complex
- **(Flexibility)** the inequality can be extended to the weight vector $\beta = \beta(X)$

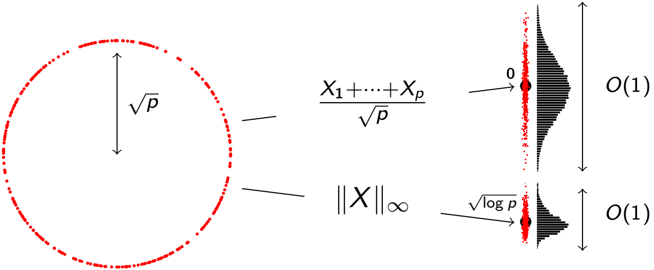
Table of Contents

Sommaire

Concentration of Measure Phenomenon[3]

$$X = (X_1, \dots, X_p) \sim \mathrm{Unif}(\mathbb{S}^{p-1})$$

Observations



Distribution diameter $= \mathbb{E}[\|Z - \mathbb{E}Z\|] \underset{p\to\infty}{=} O(\sqrt{p})$

Observable diameter $\underset{p\to\infty}{=} O(1)$

[3] Ledoux - 2001 : The concentration of measure phenomenon

Fundamental example of the Theory

Theorem

$Z \in \mathbb{R}^p$, if Z uniformly distributed on $\sqrt{p}\mathcal{S}^{p-1}$ or $Z \sim \mathcal{N}(0, I_p)$:
$\forall f : E \to \mathbb{R}$ 1-Lipschitz :

$$\forall t > 0 \ : \ \mathbb{P}\left(\left|f(Z) - \mathbb{E}[f(Z')]\right| \geq t\right) \leq 2e^{-t^2/2},$$

we note (since $2 \underset{p\to\infty}{=} O(1)$):

$$Z \propto \mathcal{E}_2(1) \qquad \text{or, more simply,} \qquad Z \propto \mathcal{E}_2$$

= Standard hypothesis

Notations

$(E, \|\cdot\|)$, normed vector space, $Z \in E$, random vector

- $\mathbb{R}^p$ endowed with: $\|x\| = \sqrt{\sum_{i=1}^p x_i^2}$ or $\|x\|_\infty = \sup_{1\leq i\leq p} |x_i|$
- $\mathcal{M}_{p,n}$ endowed with: $\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)} = \sqrt{\sum_{\substack{1\leq i\leq p\\ 1\leq j\leq n}} M_{i,j}^2}$
  or $\|M\| = \sup_{\|x\|\leq 1} \|Mx\|$

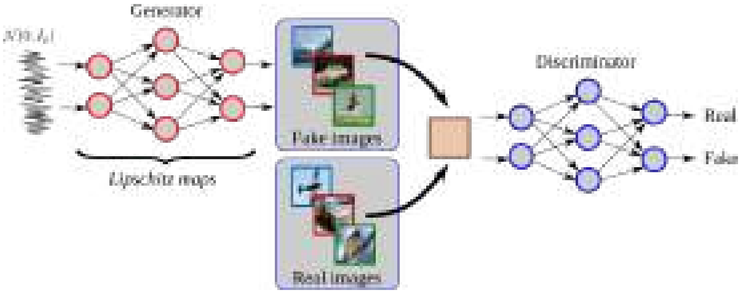
Lipschitz concentration and linear concentration

- "$Z \propto \mathcal{E}_2(\sigma)$"
  $\exists C, c > 0 \mid \forall p, n \in \mathbb{N}, \forall f : E \to \mathbb{R}$ 1-Lipschitz, :
  $$\forall t > 0 \ : \ \mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq t\right) \leq Ce^{-c(t/\sigma)^2},$$
  $\sigma = \sigma_{p,n}$: Observable Diameter of Z.
- "$Z \in \tilde{Z} \pm \mathcal{E}_2(\sigma)$"
  If $\forall p, n \in \mathbb{N}, \forall u : E \to \mathbb{R}$ 1-Lipschitz and linear :
  $$\forall t > 0 \ : \ \mathbb{P}\left(\left|u(Z - \tilde{Z})\right| \geq t\right) \leq Ce^{-c(t/\sigma)^2},$$
  $\tilde{Z}$: Deterministic equivalent of Z. $(Z \propto \mathcal{E}_2(\sigma) \implies Z \in \mathbb{E}[Z] \pm \mathcal{E}_2(\sigma))$

How to build new concentrated random vectors ?

- If $Z \propto \mathcal{E}_2(\sigma)$ and $f : E \to E$ λ-Lipschitz, $f(Z) \propto \mathcal{E}_2(\lambda\sigma)$
- No simple way to set the concentration of $(Z_1, \ldots, Z_p)$ if $Z_1 \propto \mathcal{E}_2(\sigma), \ldots, Z_p \propto \mathcal{E}_2(\sigma)$ non independent
- $Z_1, Z_2 \propto C\mathcal{E}_q(\sigma)$, independent $(Z_1, Z_2) \propto \mathcal{E}_q(\sigma)$
- $(Z_1, Z_2) = f(Z)$ where $Z \propto \mathcal{E}_q(\sigma)$, and f 1-Lipschitz $(Z_1, Z_2) \in \mathcal{E}_q(\sigma)$

Realistic images built with GANS are concentrated



$$\color{red}{\text{IMAGE} = f(Z)}, \quad \text{with } f \text{ } 1-\text{Lipschitz and } Z \sim \mathcal{N}(0, I_p)$$

Sommaire

Characterization with the centred moments

Proposition

$Z \propto \mathcal{E}_q(\sigma)$ *iif.* $\exists c > 0 | \forall p, n \in \mathbb{N}, \forall r \geq 0$, $\forall f : E \to \mathbb{R}$, 1-*Lipschitz:*

$$\mathbb{E}\left[|f(Z) - \mathbb{E}[f(Z)]|^r\right] \leq \left(\frac{r}{q}\right)^{\frac{r}{q}} c\sigma^r$$

Proof :

1. Fubini: $$\mathbb{E}\left[|Z - a|^r\right] = \int_Z \left(\int_0^\infty \mathbb{1}_{t \leq |Z-a|^r} dt\right) dZ = \int_0^\infty \mathbb{P}\left(|Z - a|^r \geq t\right) dt \leq \int_0^\infty Ce^{-t^{\frac{q}{r}}/\sigma^q} dt \ldots \leq C \left(\frac{r}{q}\right)^{\frac{r}{q}} \sigma^r$$

2. Markov inequality:

$$\mathbb{P}\left(|Z - a| \geq t\right) \leq \frac{\mathbb{E}[|Z-a|^r]}{t^r} \leq C\left(\frac{r}{q}\right)^{\frac{r}{q}} \left(\frac{\sigma}{t}\right)^r,$$

with $r = \frac{qt^q}{e\sigma^q} \geq q : \mathbb{P}\left(|Z - a| \geq t\right) \leq Ce^{-(t/\sigma)^q/e}$.

Sommaire

Control of the norm

Lemma

Given $(E, \|\cdot\|)$, *if* $Z \in \mathbb{E}[Z] \pm \mathcal{E}_2(\sigma)$:

$$\mathbb{E}[\|Z\|] \leq \|\mathbb{E}[Z]\| + O(\sigma\sqrt{\eta_{\|\cdot\|}})$$

- $\eta\left(\mathbb{R}^p, \|\cdot\|_\infty\right) = \log(p)$
- $\eta\left(\mathbb{R}^p, \|\cdot\|_2\right) = p$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|\right) = n + p$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|_F\right) = np.$

Example $Z \in \mathbb{R}^p$, $X \in \mathcal{M}_{p,n}$

- if $Z \in \tilde{Z} \pm \mathcal{E}_2$: $\mathbb{E}\,\|Z\|_\infty \leq \|\tilde{Z}\| + C\sqrt{\log p}$
- if $Z \in \tilde{Z} \pm \mathcal{E}_2$: $\mathbb{E}\,\|Z\| \leq \|\tilde{Z}\| + C\sqrt{p}$
- if $X \in \tilde{X} \pm \mathcal{E}_2$: $\mathbb{E}\,\|X\| \leq \|\tilde{X}\| + C\sqrt{p+n}$,

Sommaire

Concentration of the sum and the product

Proposition

If $(X, Y) \in \mathcal{E}_2(\sigma)$ *:* $X + Y \propto \mathcal{E}_q(\sigma)$

If $\|\mathbb{E}[X]\|', \|\mathbb{E}[Y]\|' \leq \sigma\sqrt{\eta_{\|\cdot\|'}}$ *where* $\forall x, y \in E$ $\|xy\| \leq \|x\|'\|y\|$*:*

$$XY \propto \mathcal{E}_2\left(\sigma^2\sqrt{\eta_{\|\cdot\|'}}\right) + \mathcal{E}_1\left(\sigma^2\right) \quad in \quad (E, \|\cdot\|)^4$$

Principal idea: $\|XY\| \leq \begin{cases} \|X\|\|Y\|' \\ \|X\|'\|Y\| \end{cases}$

Example

$X \in \mathcal{M}_{p,n}$, $Z \in \mathbb{R}^p$, $Z, X \in \mathcal{E}_2$, $\|\mathbb{E}[X]\| \leq O(1)$, $\|\mathbb{E}[Z]\|_\infty \leq O(1)$:

- $\frac{XX^T}{n} \in \mathcal{E}_2\left(\frac{\sqrt{p+n}}{n}\right) + \mathcal{E}_1\left(\frac{1}{n}\right)$ in $(\mathcal{M}_{p,n}, \|\cdot\|_F)$
- $Z \odot Z \in \mathcal{E}_2(\sqrt{\log p}) + \mathcal{E}_1$ in $(\mathbb{R}^p, \|\cdot\|)$

[4] $\iff \exists C, c > 0, \forall p, n, \forall f : E \to \mathbb{R}$, 1-Lipschitz, $\forall t > 0$:

$$\mathbb{P}\left(|f(XY) - \mathbb{E}[f(XY)]| \geq t\right) \leq Ce^{-c(t/\sigma^2)^2/\eta_{\|\cdot\|'}} + Ce^{-ct/\sigma^2}$$

Practical example: Hanson-Wright Theorem

Theorem

Given random $X, Y \in \mathbb{R}^p$, and $A \in \mathcal{M}_p$ deterministic, if $(X, Y) \propto \mathcal{E}_2$ and $\|\mathbb{E}[X]\|, \|\mathbb{E}[Y]\| \leq O(1)$:

$$X^T AY \propto \mathcal{E}_2(\sqrt{\log p}\|A\|_F) + \mathcal{E}_1(\|A\|_F)$$

Proof:

- Decompose $A = P\Lambda Q$, $P, Q \in \mathcal{O}_p$, $\Lambda \in \mathcal{D}_n$
- Note $\check{X} \equiv PX$, $\check{Y} \equiv QY$, $\check{X}, \check{Y} \propto \mathcal{E}_2$
- $X^T AY = \check{X}^T \Lambda \check{Y} = \lambda^T (\check{X} \odot \check{Y})$ where $\Lambda = \mathrm{Diag}(\lambda)$
- $\mathbb{E}[\|\check{X}\|_\infty] \leq \|\mathbb{E}[\check{X}]\|_\infty + O(\sqrt{\log p}) \leq \|\mathbb{E}[X]\| + O(\sqrt{\log p}) \leq O(\sqrt{\log p})$
- $\check{X} \odot \check{Y} \propto \mathcal{E}_2(\sqrt{\log n}) + \mathcal{E}_1$
- $\lambda^T (\check{X} \odot \check{Y}) \propto \mathcal{E}_2(\|\lambda\|\sqrt{\log n}) + \mathcal{E}_1(\|\lambda\|)$

Sommaire

When $E = \mathbb{R}$

- $\sigma > 0$, (changes with p, n),
- $X \sim \mathcal{N}(0, \sigma^2)$ (then $X \propto \mathcal{E}_2(\sigma)$),
- Y solution to $Y = 1 + XY$, $Y \equiv \frac{1}{1-X}$
- $f_Y(y) = \frac{e^{-(1-\frac{1}{y})^2/\sigma^2}}{\sqrt{2\pi}\sigma y^2}$,
- Clearly : $Y \not\propto \mathcal{E}_2(\sigma')$ because $f_Y(y) \underset{y\to\infty}{\sim} \frac{e^{-1/\sigma^2}}{y^2}$,
- Note $\mathcal{A}_Y \equiv \{X \leq \frac{1}{2}\}$, $\mathbb{P}(\mathcal{A}_Y^c) \leq 2e^{-1/8\sigma^2}$,
- $t \mapsto \frac{1}{1-t}$ 4-Lipschitz on $\mathcal{A}_Y$,

$\Longrightarrow$ $(Y|\mathcal{A}_Y) \propto \mathcal{E}_2(\sigma)$and we note $Y \overset{\mathcal{A}_Y}{\propto} \mathcal{E}_2(\sigma) \mid e^{-1/\sigma^2}$

(because $\mathbb{P}(\mathcal{A}_Y^c) \leq Ce^{-c/\sigma^2}$)

Concentration of solution to Concentrated equation

- $\mathcal{F}(E)$: set of mapping $E \to E$,
- $\|\phi\|_{\mathcal{B}(y_0,K)} = \sup_{\|x-y_0\|\leq K} \|\phi(x)\|$,
- $\|\phi\|_{\mathcal{L}} = \sup_{x,y\in E} \frac{\|\phi(x)-\phi(y)\|}{\|x-y\|}$.

Theorem

Given random $\phi : \mathbb{R}^n \to \mathbb{R}^n$, we note $\mathcal{A}_\phi = \{\|\phi\|_{\mathcal{L}} \leq 1-\varepsilon\}$ if:

- $\mathbb{P}(\mathcal{A}_\phi^c) \leq Ce^{-cn}$ *(for $C, c > 0$)*
- $\exists! y_0 \in \mathbb{R}^n$ / $y_0 = \mathbb{E}_{\mathcal{A}_\phi}[\phi(y_0)]$. $\forall K > 0$, *($K \leq O(1)$):*

$$\phi \overset{\mathcal{A}_\phi}{\propto} \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n} \qquad in \quad (\mathcal{F}(\mathbb{R}^n), \|\cdot\|_{\mathcal{B}(y_0,K)})$$

Then, under $\mathcal{A}_\phi$ the equation $Y = \phi(Y)$ admits a unique solution $Y \in \mathcal{M}_{p,n}$ that satisfies:

$$Y \overset{\mathcal{A}_\phi}{\propto} \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$$

Heuristic of the proof

Hypotheses

- $\mathbb{P}(\mathcal{A}_\phi^c) \leq Ce^{-cn}$ (for $C, c > 0$) with $\mathcal{A}_\phi = \{\|\phi\|_\mathcal{L} \leq 1 - \varepsilon\}$
- $\exists! y_0 \in \mathbb{R}^n \mid y_0 = \mathbb{E}_{\mathcal{A}_\phi}[\phi(y_0)]. \forall K > 0$, $(K \leq O(1))$:

$$\phi \overset{\mathcal{A}_\phi}{\propto} \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n} \qquad \text{in} \quad (\mathcal{F}(\mathbb{R}^n), \|\cdot\|_{\mathcal{B}(y_0,K)})$$

"Proof:"

- $Y \approx \phi^j(y_0)$ for j sufficiently big
- Under $\mathcal{A}_\phi$, for $K \leq O(1)$, sufficiently big

$$\forall j \in \mathbb{N},\ \phi^j(y_0) \in \mathcal{B}(y_0, K)$$

- Since ϕ concentrated in $(\mathcal{F}(\mathbb{R}^n), \|\cdot\|_{\mathcal{B}(y_0,K)})$,

$$\forall j \in \mathbb{N},\ \phi^j \overset{\mathcal{A}_\phi}{\propto} \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$$

$\Longrightarrow$ for j sufficiently big, $Y \approx \phi^j(y_0) \overset{\mathcal{A}_\phi}{\propto} \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$

Sommaire

Position of the problem

- Data matrix $X = (x_1, \dots, x_n) \in \mathcal{M}_{p,n}$,
- labels : $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$

Robust regression problem with regularizing parameter:

$$(P): \quad \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) + \lambda \|\beta\|^2$$

with $\rho : \mathbb{R} \to \mathbb{R}$ convex, $\lambda > 0$.

Differentiation:

$$(P) \iff \beta = \frac{1}{n\lambda} \sum_{i=1}^{n} \rho'(y_i - x_i^T \beta) x_i \iff \beta = \frac{1}{n} X f(X^T \beta)$$

- $f_i \equiv \frac{1}{\lambda} \rho'(y_i - \cdot)$
- $f : \mathbb{R}^n \to \mathbb{R}^n$, $f((z_i)_{1 \leq i \leq n}) = (f_i(z_i))_{1 \leq i \leq n}$

Hypotheses

On X, $\forall i \in [n]$, $\mu_i \equiv \mathbb{E}[x_i]$, $\Sigma_i \equiv \mathbb{E}[x_i x_i^T]$, $C_i \equiv \Sigma_i - \mu_i \mu_i^T$:

- $p = O(n)$
- $x_1, \ldots, x_n$ independent (with possibly different distributions)
- $X \propto \mathcal{E}_2$ (as if $X \sim \mathcal{N}(0, I_{pn})$) $\implies \|C_i\| \leq O(1)$
- $\|\mu_i\| = O(1)$ $\implies \mathbb{E}[\frac{1}{n}\|XX^T\|] \leq O(1)$

On f:

- $\|f\|_\infty \leq \infty$ ($\leq O(1)$) (unnecessary)
- $\|f'\|_\infty, \|f''\|_\infty \leq \infty$

Contractivity of $\beta = \frac{1}{n} X f(X^T \beta)$

- $\|f'\|_\infty \mathbb{E}[\|\frac{1}{n}XX^T\|] \leq 1 - 2\varepsilon$ with $\varepsilon \geq O(1)$

Goal

"Concentration of β and Estimation of first statistics"

$$\mu_\beta \equiv \mathbb{E}_{\mathcal{A}_\beta}[\beta] \qquad\qquad C_\beta \equiv \mathbb{E}_{\mathcal{A}_\beta}[\beta\beta^T] - \mu_\beta\mu_\beta^T$$

First approach: $\mu_\beta = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f(x_i^T\beta)x_i\right]$

If we admit x_i behaves like a Gaussian vector,

$\longrightarrow$ **Issue:** dependence between x_i and β

$\longrightarrow$ **Solution:** "Leave-one-out":

- introduce β_{-i}:

$$\beta_{-i} = \frac{1}{n}\sum_{\substack{1\leq j\leq n \\ j\neq i}} f(x_j^T\beta_{-i})x_j$$

- Construct $\zeta_i : \mathbb{R} \to \mathbb{R}$ deterministic $\mid x_i^T\beta \approx \zeta_i(x_i^T\beta_{-i})$

Strategy of the study

1. Introduce event $\mathcal{A}_\beta \equiv \{\|f'\|_\infty \|\frac{1}{n}XX^T\| \leq 1-\varepsilon\}$ where β concentrates.

2. Disentangle β and x_i :
$$\beta_{-i}(t) = \frac{1}{n}X_{-i}f(X_{-i}^T\beta_{-i}(t)) + \frac{t}{n}f(x_i^T\beta_{-i}(t))x_i$$
where $X_{-i} = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$.
$$\beta_{-i} = \beta_{-i}(0) \qquad \text{and} \qquad \beta = \beta_{-i}(1).$$
   2.1 Differentiate $\beta_{-i}(\cdot)$.
   2.2 Approximate $\beta'_{-i}(t)$.
   2.3 Integrate the approximation to obtain approximation of $\int_0^1 \beta'_{-i}(t)dt = \beta - \beta_{-i}$.

3. Construct deterministic $\zeta_i : \mathbb{R} \to \mathbb{R}$ st. $\beta^T x_i \approx \zeta_i(\beta_{-i}{}^T x_i)$.

4. Estimate μ_β, C_β with Gaussian Hypotheses on $x_1, \ldots, x_n$.

Sommaire

High probability of $\mathcal{A}_\beta \equiv \{\|f'\|_\infty \|\frac{1}{n}XX^T\| \leq 1 - \varepsilon\}$

Lemma

$\|\frac{1}{n}XX^T\| \propto \mathcal{E}_2(1/\sqrt{n})$

Contractivity of $\beta = \frac{1}{n}Xf(X^T\beta)$

- $\|f'\|_\infty \mathbb{E}[\|\frac{1}{n}XX^T\|] \leq 1 - 2\varepsilon$ with $\varepsilon \geq O(1)$

Lemma

$\exists C, c > 0,\ constant \mid \mathbb{P}(\mathcal{A}_\beta^c) \leq Ce^{-cn}$

Proof : $\mathbb{P}(\mathcal{A}_\beta^c) \leq \mathbb{P}(|\|\frac{1}{n}XX^T\| - \mathbb{E}[\|\frac{1}{n}XX^T\|]| \geq \frac{\varepsilon}{\|f'\|_\infty})$

$$\leq Ce^{-cn\varepsilon^2/\|f'\|_\infty^2}$$

Concentration of β

Lemma

Under $\mathcal{A}_\beta$, $\|\beta\| \leq O(1)$

Proof: $\|\beta\| = \|\frac{1}{n} X f(X^T \beta)\| \leq \frac{\|f\|_\infty}{n} \|X\| \|\mathbb{1}\| \leq O(1)$.

Note Ψ such that $\beta = \Psi(X)(\beta)$

Hypothesis for concentration of β:

1. $\mathbb{P}(\mathcal{A}_\beta^c) \leq Ce^{-cn}$ (recall that $A_\beta \equiv \{\|\Psi(X)\| \leq 1 - \varepsilon\}$)
2. $\forall K > 0$, $K \leq O(1)$[5]:
   $(\Psi(X) \mid \mathcal{A}_\beta) \propto \mathcal{E}_2(1/\sqrt{n})$ in $(\mathcal{F}(\mathbb{R}^p), \|\cdot\|_{\mathcal{B}(0,K)})$

[5] if $y_0 = \mathbb{E}_{\mathcal{A}_\beta}[\Psi(X)(y_0)]$, $\|y_0\| \leq O(1)$

Concentration of β

Proposition

$\beta \mid \mathcal{A}_\beta \propto \mathcal{E}_2(1/\sqrt{n})$

Proof: Recall that $\Psi(A)(y) = Af(A^T y)$, $(\beta = \Psi(X)(\beta))$
$\Psi : \mathcal{M}_{p,n} \to (\mathcal{F}(\mathbb{R}^p), \|\cdot\|_{\mathcal{B}(0,K)})$ is $O(1/\sqrt{n})$-Lipschitz on $\mathcal{A}_\beta$.
$\forall \|y\| \leq K$, $A, B \in \mathcal{A}_\beta$ $(\|A\|, \|B\| \leq O(1))$:

$$\|\Psi(A)(y) - \Psi(B)(y)\| \leq \frac{1}{n} \left\|(A - B)f(A^T y)\right\| + \frac{1}{n} \left\|B\left(f(A^T y) - f(B^T y)\right)\right\|$$
$$\leq O\left(\frac{1}{\sqrt{n}}\right) \|A - B\|,$$

$\Longrightarrow \Psi(X)(y) \propto \mathcal{E}_2(1/\sqrt{n})$ in $(\mathcal{F}(\mathbb{R}^p), \|\cdot\|_\infty)$.

Sommaire

Differentiation of $\beta_{-i}(\cdot)$

- $\beta_{-i}(t) = \frac{1}{n}X_{-i}f(X_{-i}^T\beta_{-i}(t)) + \frac{t}{n}f(x_i^T\beta_{-i}(t))x_i$
- $X_{-i} = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$

Proposition

$\beta_{-i}(\cdot)$ is differentiable and:

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

where:

- $Q_{-i}(t) = \left(I_p - \frac{1}{n}X_{-i}D(t)X_{-i}^T\right)^{-1} \in \mathcal{M}_p$
- $D(t) = \mathrm{Diag}(f'(x_j^T\beta_{-i}(t))_{1\leq j\leq n}$
- $\chi(t) = tf(x_i^T\beta_{-i}(t))$

$\longrightarrow$ Show that $t \mapsto \frac{1}{n}Q_{-i}(t)x_i$ is almost constant

Link between β and β_{-i}

Noting $Q_{-i} = Q_{-i}(0)$:

- $\beta'_{-i}(t) = \frac{1}{n} Q_{-i}(t) x_i \ \chi'(t)$
- $\chi(t) = tf(tx_i^T \beta_{-i}(t))$
- $\left\| \frac{1}{n} Q_{-i}(\cdot) x_i - \frac{1}{n} Q_{-i} x_i \right\| \in 0 \pm \mathcal{E}_2(1/n) \mid e^{-n}$
- $\chi'(t) \in O(1) \pm \mathcal{E}_2 \mid e^{-n}$

Proposition

$\left\| \beta - \beta_{-i} - \frac{1}{n} f(x_i^T \beta) Q_{-i} x_i \right\| \in 0 \pm \mathcal{E}_2\left(\frac{1}{n}\right) \ \mid e^{-n}$

Proof: $\beta_{-i}(1) = \beta$, $\chi(0) = 0$, $\chi(1) = \frac{1}{n} f(x_i^T \beta)$ so:

$$\beta - \beta_{-i} = \frac{1}{n} f(x_i^T \beta) Q_{-i} x_i + \frac{1}{n} \int_0^1 \chi'(t)(Q_{-i}(t) - Q_{-i}(0)) x_i dt.$$

$\Longrightarrow x_i^T \beta \approx x_i^T \beta_{-i} + \frac{1}{n} x_i Q_{-i} x_i f(x_i^T \beta).$

Sommaire

Deterministic mapping between β and β_{-i}

From $\left\|\beta - \beta_{-i} - \frac{1}{n} f(x_i^T \beta) Q_{-i} x_i \right\| \in 0 \pm \mathcal{E}_2\left(\frac{1}{n}\right) \mid e^{-n}$, we deduce:

- $\left\| x_i^T \beta - x_i^T \beta_{-i} - \Delta_i f(x_i^T \beta) \right\| \in 0 \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$ where:

$$\Delta_i = \mathbb{E}\left[\frac{1}{n} x_i^T Q_{-i} x_i\right] \quad \text{because} \quad \frac{1}{n} x_i^T Q_{-i} x_i \in \Delta_i \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$$

Lemma (Definition of ζ_i)

Given $i \in [n]$, $\exists! \zeta_i(t) \in \mathbb{R}$ / $\zeta_i(t) = t + \Delta_i f(\zeta_i(t))$

Proof: $\|f'\|_\infty \Delta_i = \mathbb{E}_{\mathcal{A}_Q}\left[\frac{\|f'\|_\infty}{n} x_i^T Q_{-i} x_i\right] \leq \mathbb{E}_{\mathcal{A}_Q}\left[\frac{\|f'\|_\infty}{n} x_i^T Q_{-i}^{\|f'\|_\infty} x_i\right]$

$$= \mathbb{E}_{\mathcal{A}_Q}\left[\frac{\|f'\|_\infty}{n} \frac{x_i^T Q_{-i}^{\|f'\|_\infty} x_i}{1 + \frac{\|f'\|_\infty}{n} x_i^T Q^{\|f'\|_\infty} x_i}\right] < 1$$

with $Q_{-i}^{\|f'\|_\infty} = \left(I_n - \frac{\|f'\|_\infty}{n} X_{-i} X_{-i}^T\right)^{-1}$

From $x_i^T\beta$ to $x_i^T\beta_{-i}$

Proposition

$$x_i^T\beta \in \zeta_i(x_i^T\beta_{-i}) \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \mid e^{-n}$$

Proof:
$$\begin{aligned}
&\left|x_i^T\beta - \zeta_i(x_i^T\beta_{-i})\right| \\
&\leq \left|x_i^T\beta - x_i^T\beta_{-i} - \Delta_i f(\zeta_i(x_i^T\beta_{-i}))\right| \\
&\leq \left|x_i^T\beta - x_i^T\beta_{-i} - \Delta_i f(x_i^T\beta)\right| + \Delta_i \left|f(x_i^T\beta) - f(\zeta_i(x_i^T\beta_{-i}))\right| \\
&\leq O\left(\frac{1}{\sqrt{n}}\right) + \|f'\|_\infty \Delta_i \left|x_i^T\beta - \zeta_i(x_i^T\beta_{-i})\right| \quad \leq \; O\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

Sommaire

Integration on x_i then on β_{-i}

Recall that $\mu_\beta = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f(x_i^T\beta)x_i\right]$

- noting $\xi_i = f \circ \zeta_i, \forall u \in \mathbb{R}^p,\ \|u\| \leq 1$:

$$\left|\mathbb{E}\left[f(x_i^T\beta)u^Tx_i\right] - \mathbb{E}\left[\xi_i(x_i^T\beta_{-i})u^Tx_i\right]\right| \leq O\left(\frac{1}{\sqrt{n}}\right)$$

- $\mu_\beta \approx \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\xi_i(x_i^T\beta_i)x_i\right]$,

Assumption

$x_i \sim \mathcal{N}(\mu_i, C_i)$

(i) Stein formula ($\int x_i$), (ii) Concentration of β_{-i} ($\int \beta_{-i}$)

$$\mathbb{E}_{-i,x_i}\left[\xi_i(x_i^T\beta_{-i})u^Tx_i\right] \overset{(i)}{=} \mathbb{E}_{-i,z}\left[\xi_i(z_{-i})\right]u^T\mu_i + \mathbb{E}_{-i}\left[\mathbb{E}_z[\xi_i'(z_{-i})]u^TC_i\beta_{-i}\right]$$

$$\overset{(ii)}{=} \mathbb{E}\left[\xi_i(z)\right]u^T\mu_i + \mathbb{E}_z[\xi_i'(z)]u^TC_i\mu_\beta + O\left(\frac{1}{\sqrt{n}}\right)$$

with $z_{-i} \sim \mathcal{N}(\beta_{-i}^T\mu_i, \beta_{-i}^TC_i\beta_{-i})$, and $z \sim \mathcal{N}\left(\mu_i^T\mu_\beta, \mathrm{Tr}(\Sigma_\beta C_i)\right)$

Fixed point equation for μ_β and C_β

Noting:

- $z \sim \mathcal{N}(\mu_i^T \mu_\beta, \operatorname{Tr}(C_\beta C_i))$
- $\tilde{\mu} \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i(z)]\mu_i$
- $\tilde{K} \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i'(z)]C_i$
- $\tilde{C} \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i^2(z)]C_i$

$$\mu_\beta = \tilde{\mu} + \tilde{K}\mu_\beta + O_{\|\cdot\|}\left(\frac{1}{\sqrt{n}}\right); \quad C_\beta = \tilde{C} + \tilde{K} C_\beta \tilde{K} + O_{\|\cdot\|_*}\left(\frac{1}{\sqrt{n}}\right)$$

Lemma

$|\Delta_i - \frac{1}{n}\operatorname{Tr}(\Sigma_i(1-\tilde{K})^{-1})| \leq O(\frac{1}{\sqrt{n}})$ *and* $\|\tilde{K}\| \leq 1-\varepsilon$.

"Proof:" $\xi_i'(t) = \frac{f'(t+\Delta_i\xi_i(t))}{1-\Delta_i f'(t+\Delta_i\xi_i(t))}$ and:

$$\Delta_i = \mathbb{E}\left[\frac{1}{n}\operatorname{Tr}\left(\Sigma_i\left(I_p - \frac{1}{n}Xf_d'(X^T\beta)X\right)^{-1}\right)\right]$$

$$= \frac{1}{n}\operatorname{Tr}\left(\Sigma_i\left(I_p - \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\frac{f'(x_j^T\beta)}{1-\Delta_j f'(x_j^T\beta)}\right]C_j\right)^{-1}\right) + O\left(\frac{1}{\sqrt{n}}\right)$$

Fixed point equation for μ_β, C_β, Δ

Proposition (Unproven)

$\exists!(\Delta, m, \sigma) \in (\mathbb{R}^n)^3$ *satisfying:*

- $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i(z_i)]\mu_i$
- $\tilde{C} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i(z_i)^2]C_i$
- $\tilde{K} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_i'(z_i)]C_i$
- $\tilde{Q} = (I_p - \tilde{K})^{-1}$
- $\tilde{\mathcal{Q}} : \mathcal{M}_p \to \mathcal{M}_p$, $\forall M$: $\tilde{\mathcal{Q}}(M) = M + \tilde{K}\tilde{\mathcal{Q}}(M)\tilde{K}$
- $m_i = \mu_i^T \tilde{Q}\tilde{\mu}$
- $\sigma_i^2 = \frac{1}{n}\operatorname{Tr}(C_i\tilde{\mathcal{Q}}(\tilde{C})) + \tilde{\mu}^T\tilde{Q}C_i\tilde{Q}\tilde{\mu}.$
- $z_i \sim \mathcal{N}(m_i, \sigma_i^2)$
- $\Delta_i = \frac{1}{n}\operatorname{Tr}\left(C_i\left(I_p - \tilde{K}\right)^{-1}\right)$
- $\xi_i(z) = f(z + \Delta_i\xi_i(z))$

With these definitions,

$$\left\|\mu_\beta - \tilde{Q}\tilde{\mu}\right\| \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right) \qquad \left\|C_\beta - \frac{1}{n}\tilde{\mathcal{Q}}(\tilde{C})\right\|_* \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right),$$

Sommaire

Softmax classification

- $(x_i)_{1\leq i\leq n}$ belong to k possible classes $\mathcal{C}_1,\ldots,\mathcal{C}_k$,
- Labels $y_1,\ldots,y_n \in \mathbb{R}^k$, if $x_i \in \mathcal{C}_\ell$, $y_i = e_\ell$
- Knowing $(x_1,y_1),\ldots,(x_n,y_n)$:
  Learning Procedure = Attribute a weight w_ℓ to each class $\mathcal{C}_\ell$,
- Given $x \in \mathbb{R}^p$, score to be in $\mathcal{C}_\ell$: $p_\ell(x) = \frac{\exp(w_\ell^T x)}{\sum_{j=1}^k \exp(w_j^T x)}$
- Chose the weights $w_1,\ldots,w_k \in \mathbb{R}^p$ that minimize:

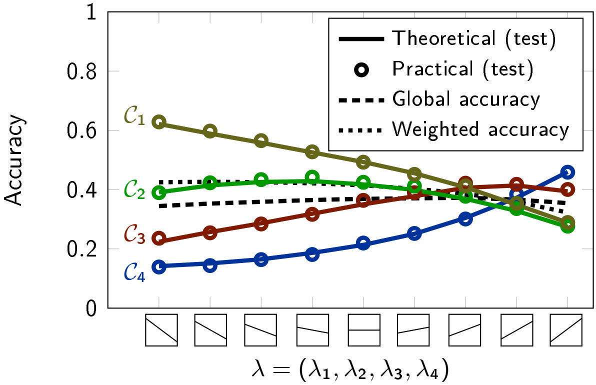
$$\mathcal{L}(w_1,\ldots,w_k) = -\frac{1}{n}\sum_{i=1}^n \sum_{\ell=1}^k y_{i,\ell}\log(p_\ell(x_i)) + \sum_{\ell=1}^k \lambda_\ell \|w_\ell\|^2$$

$$= \frac{1}{n}\sum_{i=1}^n y_i^T \log\left(\text{Softmax}(W^T x_i)\right) + \|W\Lambda\|_F^2$$

$\Longrightarrow$ **If λ is big enough, the weights concentrate and we can estimate their statistics.**

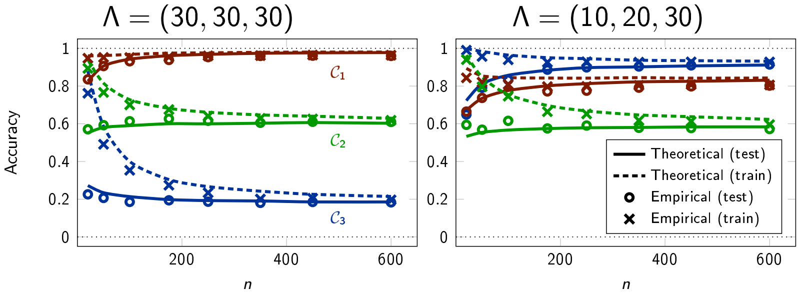
Prediction of performances on Gaussian data

With Gaussian data, $n = p = 200$,
4 classes $\#\mathcal{C}_1 > \#\mathcal{C}_2 > \#\mathcal{C}_3 > \#\mathcal{C}_4$



Prediction with GAN-generated MNIST data

With GAN- generated data, $p = 784$, 3 classes
$\#\mathcal{C}_1 > \#\mathcal{C}_2 > \#\mathcal{C}_3$.[6]



[6] Mohamed El Amine Seddik, Cosme Louart, Romain COUILLET, Mohamed Tamaazousti, "The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers", AISTATS 2021

Conclusion

Problem:[7,8]

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) + \lambda \|\beta\|^2 \iff \quad \beta = \frac{1}{n} X f(X^T \beta)$$

Main ingredients ?

- Concentration of measure hypothesis,
- Scalar product,
- Contractivity in fixed point equation

THANK YOU!

[7] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. Proceed- ings of the National Academy of Sciences, 2013.

[8] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logis- tic regression: Asymptotic performance and new insights. In ICASSP'19

Integration on β

Lemma

For any $\phi : \mathbb{R} \to \mathbb{R}$ such that $\|\phi'\|_\infty \leq O(1)$:

$$\mathbb{E}_{\beta,z}\left[\phi\left(\mu_i^T\beta + \sqrt{\beta^T C_i \beta} z\right)\right] = \mathbb{E}_z\left[\phi\left(\mu_i^T\mu_\beta + \sqrt{\mathrm{Tr}(\Sigma_\beta C_i)} z\right)\right] + O\left(\frac{1}{\sqrt{n}}\right)$$

where $z \sim \mathcal{N}(0,1)$ independent with β and $\Sigma_\beta = \mu_\beta \mu_\beta^T + C_\beta$

Proof: $\mathbb{E}_z\left[\phi\left(\mu_i^T\beta + \sqrt{\beta^T C_i \beta} z\right)\right] = \psi(\mu_i^T\beta, \beta^T C_i \beta)$ where $\psi : \mathbb{R}^2 \to \mathbb{R}$ $O(1)$-Lipschitz, thus:

$$\mathbb{E}_z\left[\phi\left(\mu_i^T\beta + \sqrt{\beta^T C_i \beta} z\right)\right] \in \psi\left(\mathbb{E}_\beta[\mu_i^T\beta], \mathbb{E}_\beta[\beta^T C_i \beta]\right) \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right)$$

Control of the norm

- Infinite norm ($Z \in \mathbb{R}^p$, $Z \propto \mathcal{E}_2(\sigma)$) :

$$\mathbb{P}\left(\|Z - \tilde{Z}\|_\infty \geq t\right) = \mathbb{P}\left(\sup_{1 \leq i \leq p} e_i^T(Z - \tilde{Z}) \geq t\right)$$

$$\leq p \sup_{1 \leq i \leq p} \mathbb{P}\left(e_i^T(Z - \tilde{Z}) \geq t\right)$$

$$\leq pCe^{-(t/\sigma)^q} \leq C'e^{-(t/\sigma\sqrt{\log(p)})^q},$$

- For the general case, use of "ε-nets".

$$\text{If } \exists H \subset (E^*, \|\cdot\|_*) \mid \forall z \in E: \ \|z\| = \sup_{f \in \mathcal{H}} f(z).^9$$

$$Z \in \tilde{Z} \pm C\mathcal{E}_2(\sigma) \implies \left\|Z - \tilde{Z}\right\| \in 0 \pm \mathcal{E}_2(\sigma\sqrt{\dim(\text{Vect}(H))})$$

[9] on $(\mathbb{R}^p, \|\cdot\|)$, $H = \mathbb{R}^p$, and $\dim(\text{Vect}(H)) = p$

Norm degree

Degree of a subset $H \subset E^*$ and of a norm

- $\eta_H = \log(\#H)$ if H is finite
- $\eta_H = \dim(\mathrm{Vect}(H))$ if H is infinite

Degree of a norm

- $\eta_{\|\cdot\|} = \inf \left\{ \eta_H, H \subset E^* \mid \forall x \in E, \|x\| = \sup_{f \in H} f(x) \right\}$

Example

- $\eta\left(\mathbb{R}^p, \|\cdot\|_\infty\right) = \log(p)$
- $\eta\left(\mathbb{R}^p, \|\cdot\|_r\right) = p$ for $r \geq 1$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|\right) = n + p$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|_F\right) = np.$

Concentration of the norm

If $Z \in \tilde{Z} \pm C\mathcal{E}_2(\sigma)$:

$$\left\| Z - \tilde{Z} \right\| \in 0 \pm C'\mathcal{E}_2(c'\sigma\eta_{\|\cdot\|}^{1/q}) \quad \text{and} \quad \mathbb{E}\left\| Z - \tilde{Z} \right\| \leq C'\sigma\eta_{\|\cdot\|}^{1/q}$$

Example $Z \in \mathbb{R}^p$, $X \in \mathcal{M}_{p,n}$

- if $Z \in \tilde{Z} \pm 2\mathcal{E}_2(\sqrt{2})$: $\mathbb{E}\|Z\| \leq \|\tilde{Z}\| + C\sqrt{p}$
- if $X \in \tilde{X} \pm 2\mathcal{E}_2(\sqrt{2})$: $\mathbb{E}\|X\| \leq \|\tilde{X}\| + C\sqrt{p+n}$,
- if $X \in \tilde{X} \pm 2\mathcal{E}_2(\sqrt{2})$: $\mathbb{E}\|X\|_F \leq \|\tilde{X}\|_F + C\sqrt{pn}$.

$\frac{1}{n}Q_{-i}(\cdot)x_i$ constant : Preliminary Lemmas

Lemma

$\|Q_{-i}(t)\| \leq \frac{1}{\varepsilon}$

We note $\beta_{-i} = \beta_{-i}(0)$, $X_{-i} = X_{-i}(0)$ and $Q_{-i} = Q_{-i}(0)$.

Lemma

$x_i^T \beta_{-i}(t) \propto \mathcal{E}_2(1) \mid e^{-n}$

$\|x_i\| \leq O(\sqrt{n})$ / $x_i \propto \mathcal{E}_2(\sqrt{n})$ $\quad$ $\|\beta_{-i}(t)\| \leq O(1)$ / $\beta_{-i}(t) \propto \mathcal{E}_2(1/\sqrt{n})$

Lemma

$\frac{1}{\sqrt{n}} X_{-i}^T Q_{-i} x_i \propto \mathcal{E}_2(1) \mid e^{-n}$ *and* $\mathbb{E}\left[\frac{1}{\sqrt{n}} \|X_{-i}^T Q_{-i} x_i\|_\infty\right] \leq O(1)$.

Proof: $\|\frac{1}{\sqrt{n}}\mathbb{E}[X_{-i}^T Q_{-i} x_i]\|_\infty \leq \|\frac{1}{\sqrt{n}}\mathbb{E}[X_{-i}^T Q_{-i}]\mu_i\| \leq O(1)$

$\mathbb{E}\left[\frac{1}{\sqrt{n}} \|X_{-i}^T Q_{-i} x_i\|_\infty\right]$

$\frac{1}{n}Q_{-i}(\cdot)x_i$ constant

Proposition

$\|Q_{-i}(t)x_i - Q_{-i}x_i\| \in O(1) \pm \mathcal{E}_2 \mid e^{-n}.$

Proof:

$$\|(Q_{-i}(t) - Q_{-i})x_i\| \leq \frac{1}{n}\left\|Q_{-i}(t)X_{-i}(D_{-i} - D(t))X_{-i}^TQ_{-i}x_i\right\|$$
$$\leq O\left(\frac{1}{\sqrt{n}}\right)\|X_{-i}^TQ_{-i}x_i\|_\infty\|D_{-i} - D_{-i}(t)\|_F.$$

Besides, $D_{-i}(t) = \mathrm{Diag}(f'(X^T\beta_{-i}(t)))$ and:

$$X^T\beta_{-i}(t) = \frac{1}{n}X^TX_{-i}f(X^T\beta_{-i}(t)) + \frac{t}{n}X^Tx_if(x_i^T\beta_{-i}(t)),$$

$$\|D_{-i} - D_{-i}(t)\|_F \leq \|f''\|_\infty\|X^T\beta_{-i}(t) - X^T\beta_{-i}(0)\|$$
$$\leq \frac{\|f''\|_\infty}{\varepsilon}\frac{t}{n}\left\|f(x_i^T\beta_{-i}(t))X^Tx_i\right\|$$
$$\leq O(\|f\|_\infty)$$