Groupe de lecture : Prédiction par minimisation de risque empirique

Mohamed Ametti, Giovanni Lebaron

ENS Rennes
14 janvier 2024

Cadre et hypothèses

On observe un jeu de données :

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

Pour simplifier, on suppose que $Y_1, \ldots, Y_n \in \{-1, 1\}$.

Cadre et hypothèses

On observe un jeu de données :

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

Pour simplifier, on suppose que $Y_1, \ldots, Y_n \in \{-1, 1\}$.

Objectif. Construire un bon *classifieur*, c'est-à-dire une application

$$h : \mathcal{X} \to \{-1, 1\}$$

dont la probabilité $L(h)$ de mauvaise classification est faible :

$$L(h) = \mathbb{P}(h(X) \neq Y).$$

Cadre et hypothèses

Par orthogonalité dans $\mathrm{L}^2(X)$:

$$\begin{aligned} L(h) &= \frac{1}{4}\mathbb{E}\left[(h(X)-Y)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(h(X)-\mathbb{E}[Y|X])^2\right]. \end{aligned}$$

Cadre et hypothèses

Par orthogonalité dans $\mathrm{L}^2(X)$:

$$\begin{aligned} L(h) &= \frac{1}{4}\mathbb{E}\left[(h(X) - Y)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(h(X) - \mathbb{E}[Y|X])^2\right]. \end{aligned}$$

C'est le *classifieur de Bayes* qui minimise $L(h)$:

$$h_*(X) = \operatorname{sgn} \mathbb{E}[Y|X]$$

Cadre et hypothèses

Par orthogonalité dans $\mathrm{L}^2(X)$:

$$\begin{aligned} L(h) &= \frac{1}{4}\mathbb{E}\left[(h(X) - Y)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(h(X) - \mathbb{E}[Y|X])^2\right]. \end{aligned}$$

C'est le *classifieur de Bayes* qui minimise $L(h)$:

$$h_*(X) = \operatorname{sgn} \mathbb{E}[Y|X]$$

. . . mais il est inconnu en général, car il dépend de la loi $\mathbb{P}$.

Cadre et hypothèses

Par orthogonalité dans $\mathrm{L}^2(X)$:

$$\begin{aligned} L(h) &= \frac{1}{4}\mathbb{E}\left[(h(X) - Y)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] + \frac{1}{4}\mathbb{E}\left[(h(X) - \mathbb{E}[Y|X])^2\right]. \end{aligned}$$

C'est le *classifieur de Bayes* qui minimise $L(h)$:

$$h_*(X) = \operatorname{sgn} \mathbb{E}[Y|X]$$

. . . mais il est inconnu en général, car il dépend de la loi $\mathbb{P}$.

Objectif. Construire $\widehat{h}$ tel que $L(\widehat{h}) - L(h_*)$ soit aussi faible que possible.

Minimisation du risque empirique

Définitions de base

Définition

Si h est un classifieur, la *probabilité empirique de mauvaise classification* de h est définie par :

$$\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{h(X_i)\neq Y_i\}}$$

c'est-à-dire :

$$\widehat{L}_n(h) = \widehat{\mathbb{P}}_n(h(X) \neq Y) \quad \text{avec} \quad \widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}.$$

Définition

Un *dictionnaire* est un ensemble de classifieurs.

Définition

Un *dictionnaire* est un ensemble de classifieurs.

Si $\mathcal{H}$ est un dictionnaire, on définit le *classifieur de minimisation du risque empirique* par :

$$\widehat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L_n}(h).$$

Définition

Un *dictionnaire* est un ensemble de classifieurs.

Si $\mathcal{H}$ est un dictionnaire, on définit le *classifieur de minimisation du risque empirique* par :

$$\widehat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L_n}(h).$$

$\rightarrow$ choix de $\mathcal{H}$?

Définition

Un *dictionnaire* est un ensemble de classifieurs.

Si $\mathcal{H}$ est un dictionnaire, on définit le *classifieur de minimisation du risque empirique* par :

$$\widehat{h}_{\mathcal{H}} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L_n}(h).$$

$\rightarrow$ choix de $\mathcal{H}$?

$\rightarrow$ $L(\widehat{h}_{\mathcal{H}}) - L(h_*)$?

Minimisation du risque empirique

Exemples de dictionnaires

Dictionnaire des classifieurs linéaires :

$$\mathcal{H}_{\text{lin}} = \{x \longmapsto \text{sgn}(\langle w, x \rangle) \mid w \in \mathbb{S}^n\}.$$

Dictionnaire des classifieurs linéaires :

$$\mathcal{H}_{\text{lin}} = \{x \longmapsto \text{sgn}(\langle w, x \rangle) \mid w \in \mathbb{S}^n\}.$$

Dictionnaire des classifieurs affines :

$$\mathcal{H}_{\text{aff}} = \{x \longmapsto \text{sgn}(\langle w, x \rangle + b) \mid w \in \mathbb{S}^n,\ b \in \mathbb{R}\}.$$

Dictionnaire des classifieurs linéaires :

$$\mathcal{H}_{\text{lin}} = \{x \longmapsto \text{sgn}(\langle w, x \rangle) \mid w \in \mathbb{S}^n\}.$$

Dictionnaire des classifieurs affines :

$$\mathcal{H}_{\text{aff}} = \{x \longmapsto \text{sgn}(\langle w, x \rangle + b) \mid w \in \mathbb{S}^n,\ b \in \mathbb{R}\}.$$

Dictionnaire des classifieurs polygonaux :

$$\mathcal{H}_{\text{poly}} = \{2\,\mathbb{1}_A - 1 \mid A \text{ polygone dans } \mathcal{X}\}.$$

On dispose de la décomposition :

$$L(\widehat{h}_{\mathcal{H}}) - L(h_*) = \underbrace{\min_{h \in \mathcal{H}} L(h) - L(h_*)}_{\text{erreur d'approximation}} + \underbrace{L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h)}_{\text{erreur stochastique}}.$$

On dispose de la décomposition :

$$L(\widehat{h}_{\mathcal{H}}) - L(h_*) = \underbrace{\min_{h \in \mathcal{H}} L(h) - L(h_*)}_{\text{erreur d'approximation}} + \underbrace{L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h)}_{\text{erreur stochastique}}.$$

$\rightarrow$ l'erreur d'approximation provient de la qualité de l'approximation de h_* par un classifieur de $\mathcal{H}$

Minimisation du risque empirique

Décomposition « biais-variance »

On dispose de la décomposition :

$$L(\widehat{h}_{\mathcal{H}}) - L(h_*) = \underbrace{\min_{h \in \mathcal{H}} L(h) - L(h_*)}_{\text{erreur d'approximation}} + \underbrace{L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h)}_{\text{erreur stochastique}}.$$

$\rightarrow$ l'erreur d'approximation provient de la qualité de l'approximation de h_* par un classifieur de $\mathcal{H}$

$\rightarrow$ l'erreur stochastique provient du choix de $\widehat{\mathbb{P}}_n$

Dimension de Vapnik-Chervonenkis

Coefficients d'éclatement

Définition

Soient $\mathcal{H}$ un dictionnaire et $n \geq 1$. Le $n^{ème}$ *coefficient d'éclatement* de $\mathcal{H}$ est défini par :

$$\mathbb{S}_n(\mathcal{H}) = \max_{x \in \mathcal{X}^n} \text{card} \left\{ (h(x_1), \ldots, h(x_n)) \mid h \in \mathcal{H} \right\}.$$

Définition

Soient $\mathcal{H}$ un dictionnaire et $n \geq 1$. Le $n^{ème}$ *coefficient d'éclatement* de $\mathcal{H}$ est défini par :

$$\mathbb{S}_n(\mathcal{H}) = \max_{x \in \mathcal{X}^n} \text{card} \left\{ (h(x_1), \ldots, h(x_n)) \mid h \in \mathcal{H} \right\}.$$

C'est le nombre maximal d'étiquetages différents de n points que les classifieurs de $\mathcal{H}$ peuvent produire.

Définition

Soient $\mathcal{H}$ un dictionnaire et $n \geq 1$. Le $n^{ème}$ *coefficient d'éclatement* de $\mathcal{H}$ est défini par :

$$\mathbb{S}_n(\mathcal{H}) = \max_{x \in \mathcal{X}^n} \text{card} \left\{ (h(x_1), \ldots, h(x_n)) \mid h \in \mathcal{H} \right\}.$$

C'est le nombre maximal d'étiquetages différents de n points que les classifieurs de $\mathcal{H}$ peuvent produire.

Par exemple, on constate facilement que $\mathbb{S}_n(\mathcal{H}_{\text{poly}}) = 2^n$.

Dimension de Vapnik-Chervonenkis

Bornes de risque

Théorème 1

Soient $\mathcal{H}$ un dictionnaire, $n \geq 1$ et $t \in (0, \infty)$. Alors, avec probabilité au moins $1 - \mathrm{e}^{-t}$:

(i) $$L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h) \leq 4\sqrt{\frac{2}{n} \log(2\mathbb{S}_n(\mathcal{H}))} + \sqrt{\frac{2t}{n}} \; ;$$

(ii) $$\left| L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) \right| \leq 2\sqrt{\frac{2}{n} \log(2\mathbb{S}_n(\mathcal{H}))} + \sqrt{\frac{t}{2n}}.$$

Théorème 1

Soient $\mathcal{H}$ un dictionnaire, $n \geq 1$ et $t \in (0, \infty)$. Alors, avec probabilité au moins $1 - \mathrm{e}^{-t}$:

(i) $$L(\widehat{h}_{\mathcal{H}}) - \min_{h \in \mathcal{H}} L(h) \leq 4\sqrt{\frac{2}{n} \log(2\mathbb{S}_n(\mathcal{H}))} + \sqrt{\frac{2t}{n}} \; ;$$

(ii) $$\left| L(\widehat{h}_{\mathcal{H}}) - \widehat{L}_n(\widehat{h}_{\mathcal{H}}) \right| \leq 2\sqrt{\frac{2}{n} \log(2\mathbb{S}_n(\mathcal{H}))} + \sqrt{\frac{t}{2n}}.$$

La démonstration consiste à prouver trois lemmes intermédiaires, et fait appel à une inégalité de concentration sophistiquée.

Théorème (McDiarmid, 1989)

Soit $F : \mathcal{X}^n \to \mathbb{R}$. On suppose qu'il existe $\delta_1, \ldots, \delta_n > 0$ tels que, pour tout $i \in \{1, \ldots, n\}$:

$$\left|F(x_1, \ldots, x_i, \ldots, x_n) - F(x_1, \ldots, x_i', \ldots, x_n)\right| \leq \delta_i.$$

Alors, pour $s > 0$, et pour toutes variables aléatoires indépendantes $X_1, \ldots, X_n$:

$$\mathbb{P}(F(X_1, \ldots, X_n) > \mathbb{E}[F(X_1, \ldots, X_n)] + s) \leq \exp\left(-\frac{2s^2}{\delta_1^2 + \ldots \delta_n^2}\right).$$

On se donne un ensemble fini de dictionnaires $\mathcal{D} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$.

Dimension de Vapnik-Chervonenkis

Pénalisation du risque empirique

On se donne un ensemble fini de dictionnaires $\mathcal{D} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$.

$\rightarrow$ quel est le dictionnaire $\mathcal{H}_*$ tel que $L(h_{\mathcal{H}_*})$ est minimal ?

Dimension de Vapnik-Chervonenkis

Pénalisation du risque empirique

On se donne un ensemble fini de dictionnaires $\mathcal{D} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$.

$\rightarrow$ quel est le dictionnaire $\mathcal{H}_*$ tel que $L(h_{\mathcal{H}_*})$ est minimal ?

$\rightarrow$ sélection d'un dictionnaire $\mathcal{H}_{\hat{m}}$ comparable à $\mathcal{H}_*$

On se donne un ensemble fini de dictionnaires $\mathcal{D} = \{\mathcal{H}_1, \dots, \mathcal{H}_M\}$.

$\rightarrow$ quel est le dictionnaire $\mathcal{H}_*$ tel que $L(h_{\mathcal{H}_*})$ est minimal ?

$\rightarrow$ sélection d'un dictionnaire $\mathcal{H}_{\hat{m}}$ comparable à $\mathcal{H}_*$

Théorème 2

On pose, pour $n \geq 1$:

$$\hat{m} = \underset{1 \leq m \leq M}{\operatorname{argmin}} \widehat{L}_n(\widehat{h}_{\mathcal{H}_m}) + 4\sqrt{\frac{2}{n} \log(2\mathbb{S}_n(\mathcal{H}))}.$$

Alors, avec probabilité au moins $1 - \mathrm{e}^{-t}$:

$$L(\widehat{h}_{\mathcal{H}_{\hat{m}}}) \leq \min_{1 \leq m \leq M} \left(\inf_{h \in \mathcal{H}_m} L(h) + \text{pénalité} \right) + \sqrt{\frac{2}{n}(t + \log M)}.$$

Dimension de Vapnik-Chervonenkis

Définition et exemples

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup \{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup \{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

C'est le nombre maximal de points de $\mathcal{X}$ qui peuvent être classifiés de n'importe quelle façon par les éléments de $\mathcal{H}$.

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

C'est le nombre maximal de points de $\mathcal{X}$ qui peuvent être classifiés de n'importe quelle façon par les éléments de $\mathcal{H}$.

Quelques exemples pour $\mathcal{X} = \mathbb{R}^d$:

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup \{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

C'est le nombre maximal de points de $\mathcal{X}$ qui peuvent être classifiés de n'importe quelle façon par les éléments de $\mathcal{H}$.

Quelques exemples pour $\mathcal{X} = \mathbb{R}^d$:

- $d_{VC}(\mathcal{H}_{\text{lin}}) = d$;

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup \{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

C'est le nombre maximal de points de $\mathcal{X}$ qui peuvent être classifiés de n'importe quelle façon par les éléments de $\mathcal{H}$.

Quelques exemples pour $\mathcal{X} = \mathbb{R}^d$:

- $d_{VC}(\mathcal{H}_{\text{lin}}) = d$;
- $d_{VC}(\mathcal{H}_{\text{aff}}) = d + 1$;

Définition

Soit $\mathcal{H}$ un dictionnaire. La *dimension de Vapnik-Chervonenkis* (aussi appelée *VC-dimension*) de $\mathcal{H}$ est définie par :

$$d_{VC}(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid \mathbb{S}_n(\mathcal{H}) = 2^n\} \in \mathbb{N} \cup \{\infty\}$$

où l'on a posé $\mathbb{S}_0(\mathcal{H}) = 1$.

C'est le nombre maximal de points de $\mathcal{X}$ qui peuvent être classifiés de n'importe quelle façon par les éléments de $\mathcal{H}$.

Quelques exemples pour $\mathcal{X} = \mathbb{R}^d$:

- $d_{VC}(\mathcal{H}_{\text{lin}}) = d$;
- $d_{VC}(\mathcal{H}_{\text{aff}}) = d + 1$;
- $d_{VC}(\mathcal{H}_{\text{poly}}) = \infty$.

Calculer la VC-dimension : difficile en général.

Calculer la VC-dimension : difficile en général.

Majorer la VC-dimension : facile avec des dimensions de référence.

Calculer la VC-dimension : difficile en général.

Majorer la VC-dimension : facile avec des dimensions de référence.

Lemme (Sauer)

Soient $\mathcal{H}$ un dictionnaire de VC-dimension $d < \infty$ et $n \in \mathbb{N}$. Alors :

$$\mathbb{S}_n(\mathcal{H}) \leq \sum_{i=0}^{d} \binom{n}{i} \leq (n+1)^d.$$

Minimisation coûteuse en général car $\mathcal{H}$ et $\widehat{L}_n$ ne sont pas convexes.

Convexification du risque empirique

Principe de la convexification

Minimisation coûteuse en général car $\mathcal{H}$ et $\widehat{L}_n$ ne sont pas convexes.

Idée. Remplacer $\mathcal{H}$ par un ensemble convexe, et $\widehat{L}_n$ par une fonction convexe.

Minimisation coûteuse en général car $\mathcal{H}$ et $\widehat{L}_n$ ne sont pas convexes.

Idée. Remplacer $\mathcal{H}$ par un ensemble convexe, et $\widehat{L}_n$ par une fonction convexe.

Si $\mathcal{F}$ est un convexe de $\mathbb{R}^{\mathcal{X}}$ et $f \in \mathcal{F}$, alors $\operatorname{sgn} f$ est un classifieur :

$$\begin{aligned}\widehat{L}_n(\operatorname{sgn} f) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\operatorname{sgn}(f(X_i))\, Y_i < 0\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{f(X_i)\, Y_i < 0\}} \cdot\end{aligned}$$

Idée. Remplacer $\ell(z) = \mathbb{1}_{\{z<0\}}$ par une fonction convexe, et considérer :

$$\widehat{f}_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}_n^{\ell}(f) \quad \text{avec} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i) Y_i)$$

puis prendre le classifieur $\widehat{h}_{\mathcal{F}} = \operatorname{sgn}(\widehat{f}_{\mathcal{F}})$.

Convexification du risque empirique

Principe de la convexification

Idée. Remplacer $\ell(z) = \mathbb{1}_{\{z<0\}}$ par une fonction convexe, et considérer :

$$\widehat{f}_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}_n^{\ell}(f) \quad \text{avec} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i) Y_i)$$

puis prendre le classifieur $\widehat{h}_{\mathcal{F}} = \operatorname{sgn}(\widehat{f}_{\mathcal{F}})$.

Question. Quel ensemble convexe $\mathcal{F}$ choisir ?

Idée. Remplacer $\ell(z) = \mathbb{1}_{\{z<0\}}$ par une fonction convexe, et considérer :

$$\widehat{f}_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}_n^{\ell}(f) \quad \text{avec} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i) Y_i)$$

puis prendre le classifieur $\widehat{h}_{\mathcal{F}} = \operatorname{sgn}(\widehat{f}_{\mathcal{F}})$.

Question. Quel ensemble convexe $\mathcal{F}$ choisir ?

Si $\mathcal{H} = \{h_1, \dots, h_d\}$ et $\mathscr{C}$ est un convexe de $\mathbb{R}^d$, on peut par exemple prendre :

$$\mathcal{F}_{\mathscr{C}} = \left\{ \sum_{i=1}^{d} \alpha_i h_i \;\middle|\; \alpha_1, \dots, \alpha_d \in \mathscr{C} \right\}.$$

Théorème 3

Soient $R \in (0, \infty)$ ainsi que ℓ une fonction convexe, décroissante, positive, α-lipschitzienne sur $[-R, R]$ et vérifiant $\ell(z) \geq \mathbb{1}_{\{z<0\}}$ pour tout $z \in \mathbb{R}$. Alors, en prenant :

$$\mathscr{C} = \overline{B}_1(0, R) = \{v \in \mathbb{R}^d \mid \|v\|_1 \leq R\}$$

et $\mathcal{F}_{\mathscr{C}}$ comme précédemment, alors avec probabilité au moins $1 - \mathrm{e}^{-t}$:

$$L(\widehat{h}_{\mathcal{F}}) \leq \widehat{L}_n^{\ell}(f) + 2\alpha R\sqrt{\frac{2}{n}\log(2d)} + \Delta\ell(R)\sqrt{\frac{t}{2n}}$$

où l'on a posé :

$$\Delta\ell(R) = |\ell(R) - \ell(-R)|\,.$$

Principe de contraction

Soient $B \subseteq \mathbb{R}^d$ borné, et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction α-lipschitzienne telle que $\varphi(0) = 0$. Alors, pour des variables de Rademacher i.i.d. $\sigma_1, \ldots, \sigma_d$:

$$\mathbb{E}_\sigma \left[\sup_{z \in B} \left| \sum_{i=1}^{d} \sigma_i \varphi(z_i) \right| \right] \leq \alpha \mathbb{E}_\sigma \left[\sup_{z \in B} \left| \sum_{i=1}^{d} \sigma_i z_i \right| \right].$$

Merci !

Merci pour votre attention !