Inhaltsverzeichnis

1 Einführung in Empirische Prozesse **2**

2 Vorbereitung **6**
2.1 Rademacher-Komplexität 6
2.2 Lévi-Abstand 8

3 Aussage des Artikels **10**

4 Anwendung **14**

5 Vergleich **16**

1 Einführung in Empirische Prozesse

Wir betrachten zu einer Familie von Klassifizierungsfunktionen $(f_\alpha : [0,1] \to \{-1,1\} \text{ mb.})_{\alpha\in\Lambda}$ die Familie

$$\mathcal{F} := \{g_\alpha : [0,1] \times \{-1,1\} \to \{0,1\}; (x,y) \mapsto 1_{f_\alpha(x)\neq y} \mid \alpha \in \Lambda\}.$$

Es folgt sofort $R_P(f_\alpha) = \mathbb{E}_P[g_\alpha] = P(g_\alpha)$, sowie $R_{\hat{P}_n}(f_\alpha) = \mathbb{E}_{\hat{P}_n}[g_\alpha] = \hat{P}_n(g_\alpha)$.

Die Frage nach der Konvergenz $R_{\hat{P}_n}(f_{T_n}) - R_P(f_{T_n}) \xrightarrow{n\to\infty} 0$ (Konsistenz der ERM.) verwandelt sich somit in die Frage der uniformen Konvergenz

$$\sup_{f\in\mathcal{F}} |\hat{P}_n(f) - P(f)| \xrightarrow{n\to\infty} 0.$$

Definition 1.1 (Glivenko-Cantelli-Klassen).
*Sei P ein W-Maß auf $(E,\mathcal{A})$ und $\mathcal{F} \subset L^1(P)$, dann heißt $\mathcal{F}$ eine P-**Glivenko-Cantelli-Klasse** (P-GC.), falls*

$$\sup_{f\in\mathcal{F}} |\hat{P}_n(f) - P(f)| \xrightarrow{n\to\infty}_{fs.} 0.$$

Satz 1.2 (Charakterisierung von P-Glivenko-Cantelli-Klassen).
Sei P ein W-Maß auf $[0,1]\times\{-1,1\}$ und $\mathcal{F} \subset L^1(P)$, dann sind äquivalent:

a) $||\hat{P}_n - P||_{\mathcal{F}} \xrightarrow{n\to\infty}_{fs.} 0$

b) $\mathbb{P}^*\left(||\hat{P}_n - P||_{\mathcal{F}} > \varepsilon\right) \xrightarrow{n\to\infty} 0\ , \quad \forall\varepsilon > 0$
und $\mathcal{F}_{0,P} := \{f - P(f) \mid f \in \mathcal{F}\}$ besitzt eine P-integrierbare Majorante.

Beweis.

$a \Leftrightarrow b$: vgl. [2] Sektion 6.6 Thm. A

□

Definition 1.3 (Donsker-Klassen).
*Eine Funktionenklasse $\mathcal{F} \subset L^2(P)$ heißt P-**Donsker-Klasse**, falls der durch $\mathcal{F}$ indizierte Prozess $\left(\sqrt{n}(\hat{P}_n(f) - P(f))\right)_{f\in\mathcal{F}}$ in Verteilung gegen einen Prozess G_P konvergiert. D.h.*

$$\forall H \in C_b(\ell^\infty(\mathcal{F})) : \mathbb{E}^*\left[H\left(\left(\sqrt{n}(\hat{P}_n(f) - P(f))\right)_{f\in\mathcal{F}}\right)\right] \xrightarrow{n\to\infty} \mathbb{E}[H(G_P)]$$

(Die Verteilung des Prozesses G_P ist nach dem Zentralen Grenzwertsatz eindeutig.)

Falls $\mathcal{F}$ eine Donsker-Klasse ist, gilt nach [2] Thm. 9.4.2, dass

$$\exists C > 0 \forall n \in \mathbb{N} : \ \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - P(f)\right|\right] \leq \frac{C}{\sqrt{n}}.$$

Damit ist jede P-Donsker-Klasse auch eine P-Glivenko-Cantelli-Klasse.

Lemma 1.4.
Falls für $y \in \{-1, 1\}$ die Funktionenklassen

$$\mathcal{F}_y := \{x \mapsto L(f_\alpha(x), y) \mid \alpha \in \Lambda\}$$

jeweils $P_{X|Y=y} =: P_y$-Donsker-Klassen und gleichmäßig beschränkt sind, dann existiert ein $C > 0$, sodass

$$\mathbb{E}\left[R_{L,P}(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha)\right] \leq \frac{C}{\sqrt{n}}.$$

Beweis. [eigenständig bewiesen]
Vorüberlegungen:
Sei $Z \sim B(n.p)$, dann berechnen wir

$$\mathbb{E}\left[\left|\frac{Z}{n} - p\right|\right]^2 \leq \frac{1}{n^2}\mathbb{E}\left[(Z - np)^2\right] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n},$$

woraus

$$\mathbb{E}\left[\left|\frac{Z}{n} - p\right|\right] \leq \sqrt{\frac{p(1-p)}{n}} \leq n^{-\frac{1}{2}} \qquad (*_1)$$

folgt.

Die $\mathcal{F}_y$ sid gleichmäßig beschränkt, also existiert ein $M > 0$ mit

$$\forall f \in \mathcal{F}_{-1} \cup \mathcal{F}_1 : ||f||_\infty \leq M. \qquad (*_2)$$

Da die F_y jeweils P_y-Donsker-Klassen sind, existieren $C_{-1}, C_1 > 0$, sodass

$$\forall n \in \mathbb{N} : \ \mathbb{E}_{X\sim P_y}\left[\sup_{f\in\mathcal{F}_y}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - P_{X|Y=y}(f)\right|\right] \leq \frac{C_y}{\sqrt{n}}. \qquad (*_3)$$

Seien $y_1, ..., y_n \in \{-1, 1\}$ beliebig und $n_{-1} := \#\{i \leq n \mid y_i = -1\}$ und analog $n_1 := \#\{i \leq n \mid y_i = 1\}$, dann gilt

$$
\begin{aligned}
&\mathbb{E}_{P^{\otimes n}}\left[\sup_{f\in\mathcal{F}}\left|\hat{P}_n(f) - P(f)\right| \mid Y_1, ..., Y_n = y_1, ..., y_n\right] \\
&= \mathbb{E}_{P^{\otimes n}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i) - P(f)\right| \mid Y_1, ..., Y_n = y_1, ..., y_n\right] \\
&= \mathbb{E}_{X_i\sim P_{y_i}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i, y_i) - P(f)\right|\right] \\
&= \mathbb{E}_{X_i\sim P_{-1}\,;\,\widetilde{X}_i\sim P_1}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n_{-1}} f(X_i, -1) + \frac{1}{n}\sum_{i=1}^{n_1} f(\widetilde{X}_i, 1)\right.\right. \\
&\qquad\qquad \left.\left. - P(Y=-1)\mathbb{E}_{P_{-1}}[f(\cdot,-1)] - P(Y=1)\mathbb{E}_{P_1}[f(\cdot,1)]\right|\right] \\
&\leq \mathbb{E}_{X_i\sim P_{-1}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n_{-1}} f(X_i, -1) - P(Y=-1)\mathbb{E}_{P_{-1}}[f(\cdot,-1)]\right|\right] \\
&\quad + \mathbb{E}_{\widetilde{X}_i\sim P_1}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n_1} f(\widetilde{X}_i, 1) - P(Y=1)\mathbb{E}_{P_1}[f(\cdot,1)]\right|\right] \\
&= \mathbb{E}_{X_i\sim P_{-1}}\left[\sup_{f\in\mathcal{F}_{-1}}\left|\frac{1}{n}\sum_{i=1}^{n_{-1}} f(X_i) - P(Y=-1)P_{-1}(f)\right|\right] \\
&\quad + \mathbb{E}_{\widetilde{X}_i\sim P_1}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{1}{n}\sum_{i=1}^{n_1} f(\widetilde{X}_i) - P(Y=1)P_1(f)\right|\right] \\
&\leq \mathbb{E}_{X_i\sim P_{-1}}\left[\sup_{f\in\mathcal{F}_{-1}}\left|\frac{1}{n}\sum_{i=1}^{n_{-1}} f(X_i) - \frac{n_{-1}}{n}P_{-1}(f)\right|\right] + \sup_{f\in\mathcal{F}_{-1}}|P_{-1}(f)|\left|P(Y=-1) - \frac{n_{-1}}{n}\right| \\
&\quad + \mathbb{E}_{\widetilde{X}_i\sim P_1}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{1}{n}\sum_{i=1}^{n_1} f(\widetilde{X}_i) - \frac{n_1}{n}P_1(f)\right|\right] + \sup_{f\in\mathcal{F}_1}|P_1(f)|\left|P(Y=1) - \frac{n_1}{n}\right| \\
&\overset{*_2,*_3}{\leq} \frac{n_{-1}}{n}\frac{C_{-1}}{\sqrt{n_{-1}}} + M\left|P(Y=-1) - \frac{n_{-1}}{n}\right| + \frac{n_1}{n}\frac{C_1}{\sqrt{n_1}} + M\left|P(Y=1) - \frac{n_1}{n}\right| \\
&= \sqrt{\frac{n_{-1}}{n}}\frac{C_{-1}}{\sqrt{n}} + M\left|P(Y=-1) - \frac{n_{-1}}{n}\right| + \sqrt{\frac{n_1}{n}}\frac{C_1}{\sqrt{n}} + M\left|P(Y=1) - \frac{n_1}{n}\right| \\
&\leq \frac{C_{-1} + C_1}{\sqrt{n}} + M\left|P(Y=-1) - \frac{n_{-1}}{n}\right| + M\left|P(Y=1) - \frac{n_1}{n}\right| \qquad (*_4)
\end{aligned}
$$

Damit können wir nun berechnen:

$$
\begin{aligned}
&\mathbb{E}_{P^{\otimes n}}\left[R_{L,P}(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha)\right] \\
&\leq \mathbb{E}_{P^{\otimes n}}\left[|R_{L,P}(\Phi(T_n)) - R_{L,\hat{P}_n}(\Phi(T_n))|\right] \\
&+ \quad \underbrace{\mathbb{E}_{P^{\otimes n}}\left[|R_{L,\hat{P}_n}(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_{L,\hat{P}_n}(f_\alpha)|\right]}_{=0,\text{ bei ERM.}} \\
&+ \quad \mathbb{E}_{P^{\otimes n}}\left[|\inf_{\alpha\in\Lambda} R_{L,\hat{P}_n}(f_\alpha) - \inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha)|\right] \\
&\leq 2\mathbb{E}_{P^{\otimes n}}\left[\sup_{\alpha\in\Lambda}\left|R_{L,\hat{P}_n}(f_\alpha) - R_{L,P}(f_\alpha)\right|\right] \\
&= 2\mathbb{E}_{P^{\otimes n}}\left[\sup_{f\in\mathcal{F}}\left|\hat{P}_n(f) - P(f)\right|\right] \\
&= 2\mathbb{E}_{P^{\otimes n}}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\hat{P}_n(f) - P(f)\right| \mid Y_1, ..., Y_n\right]\right] \\
&\overset{*_4}{\leq} 2\mathbb{E}_{P^{\otimes n}}\left[\frac{C_{-1}+C_1}{\sqrt{n}} + M\left|P(Y=-1) - \frac{n_{-1}}{n}\right| + M\left|P(Y=1) - \frac{n_1}{n}\right|\right] \\
&= 2\frac{C_{-1}+C_1}{\sqrt{n}} + 2M\mathbb{E}_{P^{\otimes n}}\left[\left|P(Y=-1) - \frac{n_{-1}}{n}\right| + \left|P(Y=1) - \frac{n_1}{n}\right|\right] \\
&\overset{*_1}{\leq} 2\frac{C_{-1}+C_1}{\sqrt{n}} + 2M\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}}\left(2C_{-1} + 2C_1 + 4M\right)
\end{aligned}
$$

□

Lemma 1.5.
Sei $\mathcal{F} \subset L^1(P_X)$, sodass die l^1-Variation

$$||f||_{TV} := \lim_{m\to\infty} \sup_{x_1<\ldots<x_m} \sum_{i=1}^{m-1} |f(x_i) - f(x_{i+1})|$$

auf $\mathcal{F}$ gleichmäßig beschränkt ist, dann gilt für jede im 1. Argument lipschitzstetige Verlustfunktion $L : \mathbb{R} \times \{-1, 1\} \to [0, \infty]$, dass die Funktionenklassen

$$\mathcal{F}_y := \{x \mapsto L(f_\alpha(x), y) \mid \alpha \in \Lambda\} \quad , y \in \{-1, 1\}$$

universelle Donsker-Klassen sind.

Beweis. [eigenständig bewiesen]
Aus der Lipschitzstetigkeit folgt leicht, dass $\mathcal{F}_y$ auch gleichmäßig beschränkte l^1-Variation besitzen. Nun folgt aus [2] Seite 329 die Behauptung. □

2 Vorbereitung

2.1 Rademacher-Komplexität

Definition 2.1 ($\mathcal{RK}_n(\mathcal{F})$).
Sei $\mathcal{F} \subset L^1(P)$ und $n \in \mathbb{N}$, dann ist die ***Rademacher-Komplexität*** *$\mathcal{RK}_n(\mathcal{F})$ von $\mathcal{F}$ definiert durch*

$$\mathcal{RK}_n(\mathcal{F}) := \mathbb{E}^* \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right],$$

für

$$\varepsilon_1, \varepsilon_2, ...uiv. \sim U(\{-1, 1\}) \text{ und } X_1, X_2, ...uiv. \sim P.$$

Bemerkung 2.2 (Symmetrisierung). *Für jedes $\mathcal{F} \subset L^1(P)$ gilt nach* [4] *Lemma 2.3.1*

$$\mathbb{E}^* \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - P(f) \right| \right] \leq 2\mathcal{RK}_n(\mathcal{F}).$$

Lemma 2.3 (Rademacher-Komplexität von Donsker-Klassen).
Sei $\mathcal{F} \subset L^2(P)$ eine P-Donsker-Klasse, dann exisitert ein $C > 0$, sodass

$$\forall n \in \mathbb{N} : \ \mathcal{RK}_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}}.$$

Beweis. [eigenständig bewiesen]
Nach [2] Thm. 9.4.2 gilt für jede $\widetilde{P}$-Donsker-Klasse $\widetilde{\mathcal{F}}$, dass

$$\exists C > 0 \forall n \in \mathbb{N} : \ \mathbb{E} \left[\sup_{\widetilde{f} \in \widetilde{\mathcal{F}}} \left| \frac{1}{n} \sum_{i=1}^{n} (\widetilde{f}(\widetilde{X_i}) - \widetilde{P}(\widetilde{f})) \right| \right] \leq \frac{C}{\sqrt{n}}.$$

Nun konstruieren wir zu einer P-Donsker-Klasse $\mathcal{F}$ ein W-Maß $\widetilde{P}$ und eine $\widetilde{P}$-Donsker-Klasse $\widetilde{\mathcal{F}}$, sodass die obere Aussage sich zur gesuchten Form umschreiben lässt.
Sei

$$\widetilde{E} := \underbrace{E}_{=:E_1} \sqcup \underbrace{E}_{=:E_2} \ ; \ \ \widetilde{P} := \frac{1}{2} P|_{E_1} + \frac{1}{2} P|_{E_2}$$
$$\widetilde{\mathcal{F}_1} := \{\widetilde{f_1}(x) := 1_{x \in I_1} 2f(x) \mid f \in \mathcal{F}\}$$
$$\widetilde{\mathcal{F}_2} := \{\widetilde{f_2}(x) := -1_{x \in I_2} 2f(x) \mid f \in \mathcal{F}\},$$

dann sieht man leicht, dass hier $\widetilde{\mathcal{F}}_1$ und $\widetilde{\mathcal{F}}_2$ auch $\widetilde{P}$-Donsker-Klassen sind. Damit ist auch $\widetilde{\mathcal{F}}_1 \cup \widetilde{\mathcal{F}}_2$ eine $\widetilde{P}$-Donsker-Klasse. Nach [4] Thm. 2.10.3 ist dann auch die Konvexe Hülle $\text{conv}(\widetilde{\mathcal{F}}_1 \cup \widetilde{\mathcal{F}}_2)$ eine $\widetilde{P}$-Donsker-Klasse, womit zuletzt auch

$$\widetilde{\mathcal{F}} := \left\{ \tilde{f}(x) := 1_{x \in E_1} f(x) - 1_{x \in E_2} f(x) \mid f \in \mathcal{F} \right\} \subset \text{conv}(\widetilde{\mathcal{F}}_1 \cup \widetilde{\mathcal{F}}_2)$$

eine $\widetilde{P}$-Donsker-Klasse ist. Es folgt die Eigenschaft

$$\begin{aligned}
\frac{C}{\sqrt{n}} &\geq \mathbb{E}\left[\sup_{\tilde{f} \in \widetilde{\mathcal{F}}} \left| \frac{1}{n} \sum_{i=1}^{n} (\tilde{f}(\widetilde{X_i}) - \underbrace{\widetilde{P}(\tilde{f})}_{=0}) \right| \right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(1_{\widetilde{X_i} \in E_1} f(\widetilde{X_i}) - 1_{\widetilde{X_i} \in E_2} f(\widetilde{X_i}) \right) \right| \right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(1_{\widetilde{X_i} \in E_1} - 1_{\widetilde{X_i} \in E_2} \right) f(\widetilde{X_i})}_{\text{beide Faktoren sind unabhängig}} \right| \right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]
\end{aligned}$$

für

$$\widetilde{X_1}, \widetilde{X_1}, ...uiv. \sim \widetilde{P} \; ; \; X_1, X_2, ...uiv. \sim P.$$

□

Lemma 2.4 (Rademacher-Komplexität von GC.-Klassen)**.**
Sei $\mathcal{F} \subset L^1(P)$ eine P-Glivenko-Cantelli-Klasse, dann gilt

$$\mathcal{RK}_n(\mathcal{F}) \xrightarrow{n \to \infty} 0.$$

Beweis. [eigenständig bewiesen]
Die definierende Eigenschaft von P-GC.-Klassen ist

$$\sup_{f \in \mathcal{F}} |\hat{P}_n(f) - P(f)| \xrightarrow{n \to \infty}_{fs.} 0.$$

Mit analoger Konstruktion zum oberen Lemma erhalten wir, dass $\widetilde{\mathcal{F}}$ ebenfalls $\widetilde{P}$-GC. ist. Die gleichmäßige Beschränktheit von $\mathcal{F}$ vererbt sich auf $\widetilde{\mathcal{F}}$, womit $\left(\sup_{\tilde{f} \in \widetilde{\mathcal{F}}} |\hat{\widetilde{P}}_n(\tilde{f}) - \widetilde{P}(\tilde{f})| \right)_{n \in \mathbb{N}}$ ggI. ist und die fs.-Konvergenz aus der GC.-Eigenschaft die L^1-Konvergenz

$$\mathbb{E}\left[\sup_{\tilde{f} \in \widetilde{\mathcal{F}}} |\hat{\widetilde{P}}_n(\tilde{f}) - \widetilde{P}(\tilde{f})| \right] \xrightarrow{n \to \infty} 0$$

impliziert. Wie schon im oberen Lemma lässt sich dies in die gewünschte Aussage umformen. □

2.2 Lévi-Abstand

Definition 2.5 (Lévi-Abstand).
*Seien P_1 und P_2 zwei W-Maße auf $(E, \mathcal{A})$ und sei $f : E \to \mathbb{R}$ messbar, dann ist der von f abhängige **Lévi-Abstand** der Maße definiert durch*

$$\mathbb{L}_f(P_1, P_2) := \inf \{\delta > 0 \mid \forall t \in \mathbb{R} : P_i(f \leq t) \leq P_j(f \leq t + \delta) + \delta \ ; \ i, j \in \{1, 2\}\}.$$

Anschaulich beschreibt also der Lévi-Abstand wie weit die Graphen der Verteilungsfunktionen $F_{(f)_*P_1}$ und $F_{(f)_*P_2}$ von einander entfernt sind.

Lemma 2.6.
Seien P_1 und P_2 zwei W-Maße auf $(E, \mathcal{A})$ und sei $f : E \to [-M, M]$ messbar, dann gilt

$$|P_1(f) - P_2(f)| \leq 8M \, \mathbb{L}_f(P_1, P_2).$$

Beweis. [eigenständig bewiesen]

$$\begin{aligned}
&P_1(f) - P_2(f) = P_1(f + M) - P_2(f + M) \\
&= \int_0^{2M} x \, d((f + M)_* P_1)(x) - \int_0^{2M} x \, d((f + M)_* P_2)(x) \\
&= \int_0^{2M} \left(\int_0^{2M} 1_{y<x} \, dy \right) d((f + M)_* P_1)(x) - \int_0^{2M} \left(\int_0^{2M} 1_{y<x} \, dy \right) d((f + M)_* P_2)(x) \\
&= \int_0^{2M} \left(\int_0^{2M} 1_{y<x} \, d((f + M)_* P_1)(x) - \int_0^{2M} 1_{y<x} \, d((f + M)_* P_2)(x) \right) dy \\
&= \int_0^{2M} \underbrace{P_1(f + M > y)}_{=:1-F_1(y-M)} - \underbrace{P_2(f + M > y)}_{=:1-F_2(y-M)} \, dy = \int_{-M}^{M} F_2 - F_1 \, d\lambda
\end{aligned}$$

Für jedes $\delta > \mathbb{L}_f(P_1, P_2)$ gilt

$$F_1 \leq F_2(\cdot + \delta) + \delta \quad ; \quad F_2 \leq F_1(\cdot + \delta) + \delta.$$

und somit

$$F_1(\cdot - \delta) - \delta \leq min(F_1, F_2) \leq max(F_1, F_2) \leq F_1(\cdot + \delta) + \delta.$$

Es folgt

$$|P_1(f) - P_2(f)| \leq \int_{-M}^{M} |F_1 - F_2| \, d\lambda \leq \int_{-M}^{M} F_1(\cdot + \delta) + \delta - F_1(\cdot - \delta) + \delta \, d\lambda$$

$$
\begin{aligned}
&= 4\delta M + \int_{-M}^{M} P_1\left(f \in (y-\delta, y+\delta]\right)\, dy \\
&= 4\delta M + \int_{-M}^{M} \int_{\mathbb{R}} 1_{x\in(y-\delta,y+\delta]}\, d((f)_* P_1)(x)\, dy \\
&\leq 4\delta M + \int_{\mathbb{R}} 4\delta M\, d((f)_* P_1)(x) = 8\delta M.
\end{aligned}
$$

□

Lemma 2.7.
Sei $f : [0,1] \to \mathbb{R}$ messbar, und P ein W-Maß auf $[0,1] \times \{-1,1\}$, dann ist

$$\mathbb{P}\left(sign(f(X)) \neq Y\right) \leq TODO$$

Beweis. [eigenständig bewiesen]
Für jedes $\delta > \mathbb{L}_f(P, \hat{P}_n)$ ist

$$F_{(f)_* P}(\cdot - \delta) - \delta \leq min(F_{(f)_* P}, F_{(f)_* \hat{P}_n}) \leq max(F_{(f)_* P}, F_{(f)_* \hat{P}_n}) \leq F_{(f)_* P}(\cdot + \delta) + \delta.$$

TODO □

Lemma 2.8.
Seien P_1 und P_2 zwei W-Maße auf $(E, \mathcal{A})$ und $f : E \to [-M, M]$ messbar. Für jedes $\delta \in (0,1]$ sei $\varphi_\delta : \mathbb{R} \to [0,1]$ die Funktion, die auf $(-\infty, 0]$ gleich 1, auf $[\delta, \infty)$ gleich 0 und dazwischen linear ist, dann gilt

$$\mathbb{L}_f(P_1, P_2) \leq \max\left(\delta, \sup_{y\in[-M,M]} |P_1(\varphi_\delta(f-y)) - P_2(\varphi_\delta(f-y))|\right) =: C_{P_1,P_2}(\delta).$$

Beweis. [nach [1] Beweis von Thm. 10]
Für jedes $y \in [-M, M]$ gilt

$$
\begin{aligned}
&P_1(f \leq y) \leq P_1(\varphi_\delta(f-y)) \\
&\leq P_2(\varphi_\delta(f-y)) + \sup_{y\in[-M,M]} |P_1(\varphi_\delta(f-y)) - P_2(\varphi_\delta(f-y))| \\
&\leq P_2(f \leq y+\delta) + \sup_{y\in[-M,M]} |P_1(\varphi_\delta(f-y)) - P_2(\varphi_\delta(f-y))| \\
&\leq P_2(f \leq y + C_{P_1,P_2}(\delta)) + C_{P_1,P_2}(\delta).
\end{aligned}
$$

Für vertauschte P_1 und P_2 gilt die Rechnung analog. □

3 Aussage des Artikels

Satz 3.1 (Allgemeine Abschätzungen mit der Rademacher-Komplexität)**.** *Sei $\mathcal{F} \subset \{f : E \to \mathbb{R} \ mb.\}$ gleichmäßig beschränkt und sei (φ_k) eine Familie von Lipschitzstetigen Funktionen mit $|\varphi_k(x) - \varphi_k(y)| \leq c_k |x - y|$ und $\varphi_k \geq 1_{(-\infty,0]}$, dann gilt für jedes $t > 0$, dass*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \ P(f \leq 0) > \frac{t}{\sqrt{n}} + \inf_{k\in\mathbb{N}} \left[\hat{P}_n(\varphi_k \circ f) + 4c_k \mathcal{RK}_n(\mathcal{F}) + \left(\frac{\log(k)}{n}\right)^{\frac{1}{2}}\right]\right) \leq 2e^{-2t^2}$$

Beweis. [nach [1] Thm. 1]
Sei OE. $1 \geq \varphi_k$ und damit auch $\varphi_k|_{(-\infty,0]} = 1$, dann definieren wir

$$\mathcal{G}_{\varphi_k} := \{\varphi_k \circ f - 1 \mid f \in \mathcal{F}\}.$$

Aus [6] Thm. 9.2 folgt

$$\mathbb{P}\left(||\hat{P}_n - P||_{\mathcal{G}_{\varphi_k}} - \mathbb{E}\left[||\hat{P}_n - P||_{\mathcal{G}_{\varphi_k}}\right] \geq \frac{t}{\sqrt{n}}\right) \leq e^{-2t^2}.$$

Der Erwartungswert lässt sich nach [4] Lemma 2.3.1 und [3] Thm. 4.12 durch

$$\mathbb{E}\left[||\hat{P}_n - P||_{\mathcal{G}_{\varphi_k}}\right] \leq 2\mathcal{RK}_n(\mathcal{G}_{\varphi_k}) \leq 4c_k \, \mathcal{RK}_n(\mathcal{F})$$

abschätzen und wir erhalten

$$\mathbb{P}\left(||\hat{P}_n - P||_{\mathcal{G}_{\varphi_k}} \geq \frac{t}{\sqrt{n}} + 4c_k \, \mathcal{RK}_n(\mathcal{F})\right) \leq e^{-2t^2}.$$

Nun gilt für jedes $f \in \mathcal{F}$, dass

$$\begin{aligned} ||\hat{P}_n - P||_{\mathcal{G}_{\varphi_k}} &= \sup_{f\in\mathcal{F}} |\hat{P}_n(\varphi_k \circ f) - P(\varphi_k \circ f)| \\ &\geq P(\varphi_k \circ f) - \hat{P}_n(\varphi_k \circ f) \geq P(f \leq 0) - \hat{P}_n(\varphi_k \circ f), \end{aligned}$$

woraus folgt:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \ P(f \leq 0) \geq \frac{t}{\sqrt{n}} + 4c_k \, \mathcal{RK}_n(\mathcal{F}) + \hat{P}_n(\varphi_k \circ f)\right) \leq e^{-2t^2}.$$

einsetzen von $t = \widetilde{t} + \sqrt{\log(k)}$ liefert

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \ P(f \leq 0) \geq \frac{\widetilde{t}}{\sqrt{n}} + \inf_{k\in\mathbb{N}} \left[4c_k \, \mathcal{RK}_n(\mathcal{F}) + \hat{P}_n(\varphi_k \circ f) + \sqrt{\frac{\log(k)}{n}}\right]\right)$$
$$\leq \sum_{k=1}^{\infty} e^{-2\left(\widetilde{t}+\sqrt{\log(k)}\right)^2} \leq 2e^{-2\widetilde{t}^2}.$$

□

Satz 3.2 (Abschätzung des Levi-Abstandes).
Sei $\mathcal{F} \subset \{f : E \to [-M, M] \ \ mb.\}$, dann gilt für jedes $t > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \mathbb{L}_f(P, \hat{P}_n) \geq \frac{t}{\sqrt{n}} + 2\left(\frac{M}{\sqrt{n}} + \mathcal{RK}_n(\mathcal{F})\right)^{\frac{1}{2}}\right) \leq e^{-2t^2}.$$

Beweis. [nach [1] Thm. 10]
Sei zu beliebigem $\delta > 0$ die Funktion φ_δ wie in Lemma 2.8 beschrieben, dann ist φ_δ Lipschitzstetig zur Konstante $\frac{1}{\delta}$.
Wir definieren

$$\mathcal{F}' := \{f(\cdot) - a \mid f \in \mathcal{F}\ ,\ a \in [-M, M]\}.$$

Analog zur ersten Hälfte des Beweises von Satz 3.1 definieren wir

$$\mathcal{G}_{\varphi_\delta} :=:= \{\varphi_\delta \circ f - 1 \mid f \in \mathcal{F}'\}$$

und erhalten

$$\mathbb{P}\left(||\hat{P}_n - P||_{\mathcal{G}_{\varphi_\delta}} \geq \frac{t}{\sqrt{n}} + \frac{4}{\delta}\,\mathcal{RK}_n(\mathcal{F}')\right) \leq e^{-2t^2}.$$

Mit folgender Abschätzung der Rademacher-Komplexität

$$\begin{aligned}
\mathcal{RK}_n(\mathcal{F}') &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sup_{a\in[-M,M]}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i) - a\right)\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sup_{a\in[-M,M]}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) - \frac{1}{n}\sum_{i=1}^n \varepsilon_i a\right|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \sup_{a\in[-M,M]} a\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \\
&= \mathcal{RK}_n(\mathcal{F}) + M\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \leq \mathcal{RK}_n(\mathcal{F}) + \frac{M}{\sqrt{n}}
\end{aligned}$$

wird daraus

$$\mathbb{P}\left(||\hat{P}_n - P||_{\mathcal{G}_{\varphi_\delta}} \geq \frac{t}{\sqrt{n}} + \frac{4}{\delta}\left(\mathcal{RK}_n(\mathcal{F}) + \frac{M}{\sqrt{n}}\right)\right) \leq e^{-2t^2}.$$

Nun setzen wir $\delta = 2\sqrt{\mathcal{RK}_n(\mathcal{F}) + \frac{M}{\sqrt{n}}}$, dann ist mit Wahrscheinlichkeit $1 - e^{-2t^2}$

$$\max\left(\delta, ||\hat{P}_n - P||_{\mathcal{G}_{\varphi_\delta}}\right) \leq 2\sqrt{\mathcal{RK}_n(\mathcal{F}) + \frac{M}{\sqrt{n}}} + \frac{t}{\sqrt{n}}$$

und wir erhalten nach Lemma 2.8 die Aussage

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\mathbb{L}_f(P,\hat{P}_n)\geq\frac{t}{\sqrt{n}}+2\left(\frac{M}{\sqrt{n}}+\mathcal{RK}_n(\mathcal{F})\right)^{\frac{1}{2}}\right)\leq e^{-2t^2}.$$

□

Korollar 3.3 (GC.-Klassen und Levi-Abstand)**.**
Sei $\mathcal{F}\subset\{f:E\to[-M,M]\ \ mb.\}$, dann ist äquivalent

a) $\mathcal{F}$ ist P-GC.

b) $\sup_{f\in\mathcal{F}}\mathbb{L}_f(P,\hat{P}_n)\xrightarrow{n\to\infty}_{fs.}0$.

Beweis. [eigenständig bewiesen]

$a\Rightarrow b$:

Nach Lemma 2.4 gilt

$$\mathcal{RK}_n(\mathcal{F})\xrightarrow{n\to\infty}0.$$

Wir wenden Satz 3.2 mit $t=\sqrt{\log(n)}$ an und erhalten für jedes $n\in\mathbb{N}$

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\mathbb{L}_f(P,\hat{P}_n)\geq\frac{\sqrt{\log(n)}}{\sqrt{n}}+2\left(\frac{M}{\sqrt{n}}+\mathcal{RK}_n(\mathcal{F})\right)^{\frac{1}{2}}\right)\\&\leq e^{-2\sqrt{\log(n)}^2}=\frac{1}{n^2}.\end{aligned}$$

Mit $\sum_{n=1}^{\infty}\frac{1}{n^2}<\infty$ und Borel-Cantelli folgt

$$\mathbb{P}\left(\limsup_{n\to\infty}\left\{\sup_{f\in\mathcal{F}}\mathbb{L}_f(P,\hat{P}_n)\geq\underbrace{\frac{\sqrt{\log(n)}}{\sqrt{n}}+2\left(\frac{M}{\sqrt{n}}+\mathcal{RK}_n(\mathcal{F})\right)^{\frac{1}{2}}}_{\to0}\right\}\right)=0.$$

$b\Rightarrow a$:

Folgt direkt aus Lemma 2.6.

□

Korollar 3.4 (Donsker-Klassen und Levi-Abstand)**.**
Sei $\mathcal{F}\subset\{f:E\to[-M,M]\ \ mb.\}$ eine P-Donsker-Klasse, dann existiert ein $C>0$, sodass

$$\forall n\in\mathbb{N}:\ \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{L}_f(P,\hat{P}_n)\right]\leq C\,n^{-\frac{1}{4}}.$$

Beweis. [nach [1] Thm. 9]
Folgt direkt aus Lemma 2.3 und Satz 3.2. □

Satz 3.5 (Margin-Abschätzung)**.**
Sei P eine Verteilung auf $E \times \{-1,1\}$ und $\mathcal{F} \subset \{f : E \times \{-1,1\} \to \mathbb{R}\, mb.\}$.
Wir definieren für ein $f \in \mathcal{F}$:

$$m_f(x,y) := f(x,y) - f(x,-y)$$

(Anschaulich soll f ein Klassifizierer sein, der x der Kategorie y zuordnet, falls $m_f(x,y) > 0$. Der Wert von m_f beschreibt dann den Abstand 'Margin' zu der Entscheidungsgrenze.)
Sei

$$RK_n := \mathcal{RK}_n\left(\{f(\cdot,1) \mid f \in \mathcal{F}\} \cup \{f(\cdot,-1) \mid f \in \mathcal{F}\}\right),$$

dann gilt für jedes $t > 0$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} :\ P(m_f \le 0) > \inf_{\delta \in (0,1)} \left[\hat{P}_n(m_f \le \delta) + \frac{48}{\delta} RK_n + \sqrt{\frac{\log\log_2(\frac{2}{\delta})}{n}}\right] + \frac{t}{\sqrt{n}}\right) \le 2e^{-2t^2}$$

(vgl. [1] Thm. 11)

4 Anwendung

Bemerkung 4.1.
Wir erinnern uns an den Zusammenhang

$$\left|R_{L,\hat{P}_n}(f_\alpha) - R_{L,P}(f_\alpha)\right| = \left|\hat{P}_n(L(f_\alpha(X),Y)) - P(L(f_\alpha(X),Y))\right|.$$

Wenn wir

$$\mathcal{F} := \{L(f_\alpha(X),Y) \mid \alpha \in \Lambda\}$$

definieren, erhalten wir mit Hilfe von Lemma 2.6 die Aussage

$$\sup_{\alpha\in\Lambda}\left|R_{\hat{P}_n}(f_\alpha) - R_P(f_\alpha)\right| \leq 4\sup_{f\in\mathcal{F}}\mathbb{L}_f(\hat{P}_n, P).$$

Hiermit sieht man leicht den Nutzen für ERM.-Algorithmen, da wir nun das Algorithmus-Risiko folgendermaßen abschätzen können. Sei dazu $f^* := \arg\min_{\alpha\in\Lambda} R_{L,P}(f_\alpha)$.

$$\begin{aligned}
&R_P(\Phi(T_n)) - R_P(f^*)\\
&= \left(R_P(\Phi(T_n)) - R_{\hat{P}_n}(\Phi(T_n))\right) + \underbrace{\left(R_{\hat{P}_n}(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_{\hat{P}_n}(f_\alpha)\right)}_{=0,\ \textit{bei ERM.}}\\
&\quad + \underbrace{\left(\inf_{\alpha\in\Lambda} R_{\hat{P}_n}(f_\alpha) - R_{\hat{P}_n}(f^*)\right)}_{\leq 0} + \left(R_{\hat{P}_n}(f^*) - R_P(f^*)\right)\\
&\leq 2\sup_{\alpha\in\Lambda}\left|R_{\hat{P}_n}(f_\alpha) - R_P(f_\alpha)\right| \leq 8\sup_{f\in\mathcal{F}}\mathbb{L}_f(\hat{P}_n, P).
\end{aligned}$$

Mit Korollar 3.4 kann man z.B. für Donsker-Klassen $\mathcal{F}$ *das erwartete Risiko abschätzen:*

$$\begin{aligned}
&\mathbb{E}\left[R_P(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_P(f_\alpha)\right]\\
&\leq \mathbb{E}\left[8\sup_{f\in\mathcal{F}}\mathbb{L}_f(\hat{P}_n, P)\right] \leq 8C\, n^{-\frac{1}{4}}.
\end{aligned}$$

Das ist in diesem Fall leider eine schlechtere Abschätzung, als die aus Lemma 1.4, womit der direkte Nutzen für Donsker-Klassen eher begrenzt ist.
Auch würde sich für die Abschätzung des Risikos bei nicht-Donsker-Klassen in Abhängigkeit von $\mathcal{RK}_n(\mathcal{F})$ *die in Bemerkung 2.2 erwähnte Ungleichung besser eignen, denn wir erhalten*

$$\mathbb{E}\left[R_P(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_P(f_\alpha)\right] \leq 4\mathcal{RK}_n(\mathcal{F}).$$

Das ist wieder eine echt bessere Abschätzung des erwarteten Risikos, als wir mit Satz 3.2 erhalten würden.

Bemerkung 4.2.
Wir stellen leicht fest, dass für eine Funktionenklasse $\mathcal{F}$ die konvexe Hülle $co(\mathcal{F})$ dieselbe Rademacher-Komplexität besitzt.

$$\mathcal{RK}_n(co(\mathcal{F})) = \mathbb{E}\left[\sup_{m\in\mathbb{N}}\sup_{f_1,\dots,f_m\in\mathcal{F}}\sup_{\lambda_j>0;\sum_{j=1}^m\lambda_j=1}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\sum_{j=1}^m\lambda_j f_j(X_i)\right|\right]$$

$$\leq \mathbb{E}\left[\sup_{m\in\mathbb{N}}\sup_{f_1,\dots,f_m\in\mathcal{F}}\sup_{\lambda_j>0;\sum_{j=1}^m\lambda_j=1}\sum_{j=1}^m\lambda_j\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f_j(X_i)\right|\right]$$

$$= \mathbb{E}\left[\sup_{m\in\mathbb{N}}\sup_{\lambda_j>0;\sum_{j=1}^m\lambda_j=1}\sum_{j=1}^m\lambda_j\sup_{f_j\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f_j(X_i)\right|\right]$$

$$= \mathbb{E}\left[\underbrace{\sup_{m\in\mathbb{N}}\sup_{\lambda_j>0;\sum_{j=1}^m\lambda_j=1}\sum_{j=1}^m\lambda_j}_{=1}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] = \mathcal{RK}_n(\mathcal{F}) \leq \mathcal{RK}_n(co(\mathcal{F}))$$

Diese Eigenschaft macht das Ergebnis des Artikels interessant für Lern-Algorithmen, die aus gewichteten Summen von Entscheidungen simpler Algorithmen bestehen.

5 Vergleich

Der folgende Satz verwendet nur Lemma 1.4 und eine ähnliche Methode zu der die im Beweis von Satz 3.1 verwendet wird. Es werden jedoch keine Ergebnisse aus dem Artikel direkt benötigt.

Satz 5.1 (RKHS. Algorithmus-Risiko)**.**
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$. Wir definieren die Funktionsfamilien

$$(f_\alpha)_{\alpha \in \Lambda_{r,\sigma}} = \mathcal{F}_{r,\sigma} := \{f \in H_\sigma \mid ||f||_{H_\sigma} \leq r\}$$

und die ERM.

$$\Phi_{r,\sigma}(T_n) := \arg\min_{\alpha \in \Lambda_{r,\sigma}} \mathbb{R}_{L,\hat{P}_n}(f_\alpha)$$

für eine im 1. Argument Lipschitzstetige Verlusfunktion L.
Dann existieren (nur von P abhängige) $C_1, C_2 > 0$, sodass für beliebige $r, \sigma > 0$ und $n \in \mathbb{N}$ gilt

$$\mathbb{E}\left[R_{L,P}(\Phi_{r,\sigma}(T_n)) - \inf_{\alpha \in \Lambda_{r,\sigma}} R_{L,P}(f_\alpha)\right] \leq \frac{r}{\sigma}\frac{C_1}{\sqrt{n}} + \frac{C_2}{\sqrt{n}}.$$

Beweis. [eigenständig beiwesen]
In Lemma 6.1 im letzten Vortrag haben wir gezeigt, dass für jedes $f \in H_\sigma$ gilt:

$$||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}}||f||_{H_\sigma}$$

Damit ist

$$\mathcal{F}_{r,\sigma} = B_r^{H_\sigma}(0) \subset E_{\sqrt{2}\frac{r}{\sigma}} := \left\{f : [0,1] \to \mathbb{R}\, mb. \mid ||f||_{TV} \leq \sqrt{2}\frac{r}{\sigma}\right\}.$$

Aus der Lipschitzstetigkeit von $L(\cdot,0)$ und $L(\cdot,1)$ folgt, dass ein $M > 0$ existiert, sodass

$$\frac{\sigma}{r}L(\mathcal{F}_{r,\sigma},0) \cup \frac{\sigma}{r}L(\mathcal{F}_{r,\sigma},1) \subset E_M.$$

Nach [2] S. 329 ist E_M eine universelle Donsker-Klasse, womit $\frac{\sigma}{r}L(\mathcal{F}_{r,\sigma},0)$ und $\frac{\sigma}{r}L(\mathcal{F}_{r,\sigma},1)$ ebenfalls universelle Donsker-Klassen sind.
Nun erhalten wir mit der Eigenschaft von Donsker-Klassen (vgl. [2] Thm. 9.4.2), dass für $y \in \{-1,1\}$ gilt

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_{r,\sigma}} \left|P\left(\frac{\sigma}{r}L\left(f(X),y\right)\right) - \hat{P}_n\left(\frac{\sigma}{r}L\left(f(X),y\right)\right)\right|\right] \leq \frac{C}{\sqrt{n}}$$

für ein $C > 0$. Nun bringen wir die Vorfaktoren auf die andere Seite und erhalten

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_{r,\sigma}}\left|P\left(\frac{\sigma}{r}L\left(f(X),y\right)\right)-\hat{P}_n\left(\frac{\sigma}{r}L\left(f(X),y\right)\right)\right|\right]\leq\frac{r}{\sigma}\frac{C}{\sqrt{n}}.$$

Mit analoger Rechnung zu Lemma 1.4 erhalten wir damit

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_{r,\sigma}}\left|R_{L,P}(f)-\inf_{g\in\mathcal{F}_{r,\sigma}}R_{L,P}(g)\right|\right]\leq\frac{r}{\sigma}\frac{C_1}{\sqrt{n}}+\frac{C_2}{\sqrt{n}}$$

für $C_1, C_2 > 0$. □

Satz 5.2 (RKHS. Familien-Risiko)**.**
Unter denselben Voraussetzungen, wie im vorherigen Satz (falls P den GNE. $\alpha \in (0,\infty)$ hat), gilt mit Hinge-Loss l:

$$\inf_{f\in\mathcal{F}_{r,\sigma}}R_{l,P}(f)-\inf_{f:[0,1]\to\{-1,1\}}R_{l,P}(f)\leq c\left(\frac{1}{\sigma r^2}+C_\alpha 2^{\frac{\alpha}{2}}\sigma^\alpha\right)$$

Für ein von r, σ unabhängiges $c > 0$ und C_α wie in der Definition des GNE. aus dem letzten Vortrag.

Beweis. Dies folgt sofort aus [7] Thm. 2.7. □

Bemerkung 5.3.
Falls die Folgen r_n und σ_n so gewählt werden, dass

$$\frac{r_n}{\sigma_n}\xrightarrow{n\to\infty}0,$$

dann kann das Familien-Risiko

$$\inf_{f\in\mathcal{F}_{r(n),\sigma(n)}}R_{l,P}f-\inf_{f:[0,1]\to\{-1,1\}\,mb.}R_{l,P}(f)$$

für nicht-triviale P nicht mehr gegen 0 konvergieren.
Es ist also nicht möglich die erhaltene Konvergenzgeschwindigkeit von $n^{-\frac{1}{2}}$ durch geeignete Wahl von r_n und σ_n zu verbessern.
(Wir werden sie sogar verschlechtern müssen um Konvergenz von Algorithmus-Risiko und Familien-Risiko zu erhalten.)

Bemerkung 5.4.
Wir wollen nun die Folgen r_n und σ_n geeignet wählen um eine möglichst schnelle Konvergenzgeschwindigkeit von

$$\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}(T_n))-\inf_{f:[0,1]\to\{-1,1\}}R_{l,P}(f)\right]$$

zu erhalten. Nach den Sätzen 5.1 und 5.2 ist die Konvergenzgeschwindigkeit bis auf Konstanten bestimmt durch

$$\max\left(\sigma_n^{\alpha}, \frac{1}{\sigma_n r_n^2}, \frac{r_n}{\sigma_n\sqrt{n}}, \frac{1}{\sqrt{n}}\right).$$

Dabei macht der letzte Term keine Unterschied, denn wie in Bemerkung 5.3 erwähnt, darf $\frac{r_n}{\sigma_n}$ keine Nullfolge sein. Durch Gleichsetzen der drei übrigen Terme erhalten wir die (bis auf Konstanten) optimale Wahl

$$r_n := n^{\frac{1}{6}} \qquad \sigma_n := n^{-\frac{1}{3(\alpha+1)}}$$

mit Konvergenzgeschwindigkeit $n^{-\frac{\alpha}{3(\alpha+1)}}$.

Fazit:
Durch die primitivere Methode der ERM. mit schrittweiser Vergrößerung der Klassifikatoren-Familie $(f_\alpha)_{\alpha\in\Lambda_{r_n,\sigma_n}}$ erhalten wir (wie zu erwarten) nicht so gute Raten, wie mit der im letzten Vortrag beschriebenen ERMR. auf dem RKHS. H_{σ_n}.

Satz 5.5.
Sei P eine Verteilung auf $\mathbb{R} \times \{-1, 1\}$, sodass P_X eine beschränkte Dichte p hat und sei s^* der Bayes-Klassifikator.
Wir betrachten das Setting vom Ende des letzten Vortrags mit

$$\Phi_n^N(x) := \Phi^{Naiv}(T_n)(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma^2}}.$$

Definiere für $y \in \{-1, 1\}$:

$$P_y = P_{X|Y=y} \quad , \quad P_y = p_y \cdot \lambda$$

und

$$\hat{P}_{n,\sigma_n,y} = \hat{p}_{n,\sigma_n,y} \cdot \lambda \quad , \quad \hat{p}_{n,\sigma_n,y}(x) = \frac{1}{n} \sum_{i \le n, Y_i = y} \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma_n^2}}.$$

Unter der Voraussetzung, dass P den TNE $q \in [0, \infty]$ besitzt und dass gilt:

$$\exists \varepsilon > 0 \ \forall \sigma > 0, y \in \{-1, 1\} : \\ \left| \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(x + t\sigma)) \frac{e^{-t^2}}{\sqrt{2\pi}} \, dt \, dP_y(x) \right| \le c \, \sigma^{\varepsilon} \qquad (*_1)$$

Dann folgt für die Wahl $\sigma_n := n^{-\frac{1}{2(1+\varepsilon)}}$, dass

$$\mathbb{E}\left[R_P(sign \circ \Phi_n^N) - R_P(s^*) \right] \le c \, n^{-\frac{\varepsilon q}{2(q+1)(1+\varepsilon)}}.$$

Beweis. [eigenständig bewiesen]
Zuerst wollen wir für $n_y := \#\{i \le n \mid Y_i = y\}$ und $\lambda_y = P(Y = y)$ den Term

$$P\big(Y \Phi_n^N(X) \le 2\delta_n\big) - P\big(Y(\lambda_Y p_Y(X) - \lambda_{-Y} p_{-Y}(X)) \le 3\delta_n\big)$$

abschätzen. Dafür berechnen wir

$$\begin{aligned} &\mathbb{E}_{x \sim P_y} \left[p_y(x) - P_y \left(\frac{K_{\sigma_n}(x, X)}{\sigma_n \sqrt{2\pi}} \right) \right] \\ &= \int_{\mathbb{R}} p_y(x) - \int_{\mathbb{R}} p_y(z) \frac{K_{\sigma_n}(x, z)}{\sigma_n \sqrt{2\pi}} \, dz \, dP_y(x) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(z)) \frac{e^{-\frac{(x-z)^2}{\sigma_n^2}}}{\sigma_n \sqrt{2\pi}} \, dz \, dP_y(x) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(x + z)) \frac{e^{-\frac{z^2}{\sigma_n^2}}}{\sigma_n \sqrt{2\pi}} \, dz \, dP_y(x) \end{aligned}$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(x + t\sigma_n)) \frac{e^{-t^2}}{\sqrt{2\pi}} \, dt \, dP_y(x) \overset{*_1}{\leq} c\sigma_n^{\varepsilon} \qquad (*_2)$$

Daraus folgt

$$\begin{aligned}
&P\big(Y\Phi_n^N(\widetilde{X}) \leq 2\delta_n\big) - P\big(\lambda_Y p_Y(\widetilde{X}) - \lambda_{-Y} p_{-Y}(\widetilde{X}) \leq 3\delta_n\big) \\
&= \sum_{y \in \{-1,1\}} \lambda_y \left(\mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[1_{y\Phi_n^N(\widetilde{X}) \leq 2\delta_n} \right] - \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[1_{\lambda_y p_y(\widetilde{X}) - \lambda_{-y} p_{-y}(\widetilde{X}) \leq 3\delta_n} \right] \right) \\
&\leq \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[1_{y\Phi_n^N(\widetilde{X}) \leq 2\delta_n \wedge \lambda_y p_y(\widetilde{X}) - \lambda_{-y} p_{-y}(\widetilde{X}) \leq 3\delta_n} \right] \\
&\leq \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[1_{|y\Phi_n^N(\widetilde{X}) - \lambda_y p_y(\widetilde{X}) + \lambda_{-y} p_{-y}(\widetilde{X})| \geq \delta_n} \right] \\
&\leq \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[\left| y\Phi_n^N(\widetilde{X}) - \lambda_y p_y(\widetilde{X}) + \lambda_{-y} p_{-y}(\widetilde{X}) \right| \right] \\
&= \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[\left| y \frac{1}{n} \sum_{i=1}^{n} Y_i \frac{K_{\sigma_n}(\widetilde{X}, X_i)}{\sigma_n \sqrt{2\pi}} - \lambda_y p_y(\widetilde{X}) + \lambda_{-y} p_{-y}(\widetilde{X}) \right| \right] \\
&= \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[\left| \frac{1}{n} \sum_{i \leq n, Y_i = y}^{n} \frac{K_{\sigma_n}(\widetilde{X}, X_i)}{\sigma_n \sqrt{2\pi}} - \right. \right. \\
&\qquad \left. \left. \frac{1}{n} \sum_{i \leq n, Y_i = -y}^{n} \frac{K_{\sigma_n}(\widetilde{X}, X_i)}{\sigma_n \sqrt{2\pi}} - \lambda_y p_y(\widetilde{X}) + \lambda_{-y} p_{-y}(\widetilde{X}) \right| \right] \\
&= \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y, T_n \sim P^{\otimes n}} \left[\left| \frac{n_y}{n} (\widehat{P_y})_{n_y} \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - \right. \right. \\
&\qquad \left. \left. \frac{n_{-y}}{n} (\widehat{P_{-y}})_{n_{-y}} \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - \lambda_y p_y(\widetilde{X}) + \lambda_{-y} p_{-y}(\widetilde{X}) \right| \right] \\
&\leq \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y} \left[\left| \frac{n_y}{n} P_y \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - \frac{n_y}{n} p_y(\widetilde{X}) \right| + \right. \\
&\qquad \left. \left| \frac{n_{-y}}{n} P_{-y} \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - \frac{n_{-y}}{n} p_{-y}(\widetilde{X}) \right| \right] \\
&\quad + \frac{2}{\delta_n} \sum_{y \in \{-1,1\}} \mathbb{E} \left[\left| \frac{n_y}{n} - \lambda_y \right| \right] + \frac{c}{\delta_n \sigma_n \sqrt{n}} \\
&\leq \frac{1}{\delta_n} \sum_{y \in \{-1,1\}} \lambda_y \mathbb{E}_{\widetilde{X} \sim P_y} \left[\left| P_y \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - p_y(\widetilde{X}) \right| + \left| P_{-y} \left(\frac{K_{\sigma_n}(\widetilde{X}, X)}{\sigma_n \sqrt{2\pi}} \right) - p_{-y}(\widetilde{X}) \right| \right]
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{c}{\delta_n\sqrt{n}} + \frac{c}{\delta_n\sigma_n\sqrt{n}} \\
&\overset{*_2}{\leq} \frac{c\sigma_n^{\varepsilon}}{\delta_n} + \frac{c}{\delta_n\sqrt{n}} + \frac{c}{\delta_n\sigma_n\sqrt{n}} \qquad (*_3)
\end{aligned}
$$

Als weitere Vorüberlegung wollen wir $\mathbb{E}\left[\hat{P}_n(Y\Phi_n^N(X) \leq \delta_n) - P(Y\Phi_n^N(X) \leq 2\delta_n)\right]$ abschätzen. Dazu definieren wir φ_n als 1 auf $(-\infty, \delta_n]$, als 0 auf $[2\delta_n, \infty)$ und dazwischen linear. Wir sehen nach [2] Seite 329 leicht, dass die Menge

$$\mathcal{F} := \{\delta_n\varphi_n \circ y\Phi_n^N \mid n \in \mathbb{N},\ y \in \{-1, 1\},\ T_n \text{ Trainigsdatensatz}\}$$

eine Donsker-Klasse ist. Wir folgern

$$
\begin{aligned}
&\mathbb{E}\left[\hat{P}_n(Y\Phi_n^N(X) \leq \delta_n) - P(Y\Phi_n^N(X) \leq 2\delta_n)\right] \\
&\leq \mathbb{E}\left[\hat{P}_n(\varphi_n \circ Y\Phi_n^N(X) \leq \delta_n) - P(\varphi_n \circ Y\Phi_n^N(X) \leq 2\delta_n)\right] \\
&\leq \frac{1}{\delta_n}\mathbb{E}\left[\hat{P}_n(\delta_n\varphi_n \circ Y\Phi_n^N(X) \leq \delta_n) - P(\delta_n\varphi_n \circ Y\Phi_n^N(X) \leq 2\delta_n)\right] \\
&\leq \frac{c}{\delta_n\sqrt{n}} \qquad (*_4)
\end{aligned}
$$

Nun wenden wir Satz 3.5 für $f_n(x, y) := \hat{p}_{n,\sigma_n,y}(x)$ an und erhalten

$$
\begin{aligned}
2e^{-2t^2} &\geq \mathbb{P}\left(\exists f \in \mathcal{F}:\ P(m_f \leq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_f \leq \delta) + \frac{c}{\sigma_n\delta\sqrt{n}} + \sqrt{\frac{\log\log_2(\frac{2}{\delta})}{n}}\right] + \frac{t}{\sqrt{n}}\right) \\
&\geq \mathbb{P}\left(\exists f \in \mathcal{F}:\ P(m_f \leq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_f \leq \delta) + \frac{c}{\sigma_n\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right) \\
&\geq \mathbb{P}\left(\ P(m_{f_n} \leq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_{f_n} \leq \delta) + \frac{c}{\sigma_n\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\ R_P(\text{sign} \circ \Phi_n^N) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(Y\Phi_n^N(X) \leq \delta) + \frac{c}{\sigma_n\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right) \\
&\geq \mathbb{P}\left(\ R_P(\text{sign} \circ \Phi_n^N) > \hat{P}_n(Y\Phi_n^N(X) \leq \delta_n) + \frac{c}{\sigma_n\delta_n\sqrt{n}} + \frac{t}{\sqrt{n}}\right)
\end{aligned}
$$

Für $t_n := \frac{1}{\sigma_n\delta_n\sqrt{n}}$ folgt:

$$\mathbb{E}\left[R_P(\text{sign} \circ \Phi_n^N) - \hat{P}_n(Y\Phi_n^N(X) \leq \delta_n)\right] \leq \frac{c}{\sigma_n\delta_n\sqrt{n}}. \qquad (*_5)$$

Mit diesen Vorüberlegungen berechnen wir nun

$$\begin{aligned}
&\mathbb{E}\left[R_P(\operatorname{sign}\circ\Phi_n^N) - R_P(s^*)\right] \\
&= \underbrace{\mathbb{E}\left[R_P(\operatorname{sign}\circ\Phi_n^N) - \hat{P}_n(Y\Phi_n^N(X)\le\delta_n)\right]}_{\le\frac{c}{\sigma_n\delta_n\sqrt{n}}\quad(*_5)} \\
&\quad + \underbrace{\mathbb{E}\left[\hat{P}_n(Y\Phi_n^N(X)\le\delta_n) - P(Y\Phi_n^N(X)\le 2\delta_n)\right]}_{\le\frac{c}{\delta_n\sqrt{n}}\quad(*_4)} \\
&\quad + \underbrace{P\big(Y\Phi_n^N(X)\le 2\delta_n\big) - P\big(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 3\delta_n\big)}_{\le\frac{c\sigma_n^\varepsilon}{\delta_n}+\frac{c}{\delta_n\sqrt{n}}+\frac{c}{\delta_n\sigma_n\sqrt{n}}\quad(*_3)} \\
&\quad + P\big(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 3\delta_n\big) - R_P(s^*)
\end{aligned}$$

für $\lambda_y = P(Y=y)$. Der letzte Term lässt sich mit Hilfe des TNE $q\in[0,\infty]$ abschätzen.

$$\begin{aligned}
&P\big(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 3\delta_n\big) - R_P(s^*) \\
&= P\big(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 3\delta_n\big) - P(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 0) \\
&\le P\big(\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)\le 3\delta_n \wedge \lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X) > 0\big) \\
&\le P\big(|\lambda_Y p_Y(X) - \lambda_{-Y}p_{-Y}(X)|\le 3\delta_n\big) \\
&= P\left(|2\eta(X)-1|\le 3\delta_n p(X)\right) \le P\left(\left|\eta(X)-\frac{1}{2}\right|\le\frac{3}{2}\delta_n||p||_\infty\right) \\
&\le c\delta_n^q.
\end{aligned}$$

Mit den oberen beiden Rechnungen erhalten wir

$$\begin{aligned}
&\mathbb{E}\left[R_P(\operatorname{sign}\circ\Phi_n^N) - R_P(s^*)\right] \\
&\le c\max\left(\frac{1}{\sigma_n\delta_n\sqrt{n}}, \frac{\sigma_n^\varepsilon}{\delta_n}, \frac{1}{\delta_n\sqrt{n}}, \frac{1}{\delta_n\sigma_n\sqrt{n}}, \frac{1}{\delta_n n^\varepsilon}, \delta_n^q\right) \\
&= c\max\left(\frac{1}{\sigma_n\delta_n\sqrt{n}}, \frac{\sigma_n^\varepsilon}{\delta_n}, \delta_n^q\right)
\end{aligned}$$

Durch Gleichsetzten der drei Terme erhalten wir die (bis auf konstanten) optimale Wahl $\sigma_n = n^{-\frac{1}{2(1+\varepsilon)}}$ und $\delta_n = \sigma_n^{\frac{\varepsilon}{q+1}} = n^{-\frac{\varepsilon}{2(q+1)(1+\varepsilon)}}$. Daraus erhalten wir die Konvergenzgeschwindigkeit $n^{-\frac{\varepsilon q}{2(q+1)(1+\varepsilon)}}$. □

Wir wenden nun Satz 3.5 auf die im letzten Vortrag erarbeitete andere Betrachtungsweise des ERMR Klassifikators an.

Satz 5.6.
Sei P ein W-Maß auf $[0,1]\times\{-1,1\}$ mit GNE $\alpha\in(0,\infty)$. Wähle $\lambda_n = n^{-\frac{1}{3}}$ und $\sigma_n = n^{-\frac{1}{3(\alpha+1)}}$. Wir erinnern uns an die im letzten Vortrag definierte ERM mit Regularisierung:

$$\Phi_{\lambda_n,\sigma_n}(T_n) = \arg\min_{f\in H_{\sigma_n}} \Big(\lambda_n\cdot||f||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(f)\Big)$$

Wir hatten auch im letzten Vortrag gesehen, dass der Minimierer die Form

$$\Phi_{\lambda_n,\sigma_n}(T_n)(x) = \sum_{i=1}^{n}\alpha_i Y_i K_{\sigma_n}(x,X_i)$$

für $\alpha_1^n,...,\alpha_n^n\geq 0$ annimmt.
Unter der Voraussetzung, dass ein $\varepsilon\in[0,\frac{1}{2})$ existiert mit

$$\mathbb{E}\Big[\underbrace{\sum_{j=1}^{n}\alpha_j^n}_{=:S_n}\Big] = \mathcal{O}(n^{\varepsilon}) \qquad (*_1)$$

existiert nun ein $C>0$, sodass für jedes $n\in\mathbb{N}$ gilt:

$$\mathbb{E}\left[R_P(sign(\Phi_{\lambda_n,\sigma_n}(T_n))) - \inf_{f:[0,1]\to\{-1,1\}} R_P(f)\right] \leq C\, n^{-\min\left(\frac{1-2\varepsilon}{4},\frac{\alpha}{3(\alpha+1)}\right)}.$$

Beweis. [eigenständig bewiesen]
Im letzten Vortrag haben wir gesehen, dass

$$\operatorname{sign}(\Phi_{\lambda_n,\sigma_n}(T_n)(\cdot)) = 1 \Leftrightarrow \sum_{i\leq n;Y_i=1}\underbrace{\frac{\alpha_i^n}{S_n}}_{=:\widetilde{\alpha}_i^n} K_{\sigma_n}(\cdot,X_i) > \sum_{i\leq n;Y_i=-1}\underbrace{\frac{\alpha_i^n}{S_n}}_{=:\widetilde{\alpha}_i^n} K_{\sigma_n}(\cdot,X_i).$$

Wir definieren nun

$$f_n(x,y) := \sum_{i\leq n;Y_i=y}\widetilde{\alpha}_i^n K_{\sigma_n}(x,X_i)$$

und (wie in Satz 3.5)

$$m_{f_n} := f_n(x,y) - f_n(x,-y) = \frac{y\,\Phi_{\lambda_n,\sigma_n}(T_n)(x)}{S_n} \qquad (*_2)$$

Wir sehen leicht, dass so ein $f_n(\cdot, y)$ immer in der Menge

$$\mathcal{F} := co\left(\{0\} \cup \{K_\sigma(\cdot, X) \mid \sigma > 0,\ X \in [0,1]\}\right)$$

liegen wird. Nach [2] Seite 329 ist $\{K_\sigma(\cdot, X) \mid \sigma > 0,\ X \in [0,1]\}$ eine universelle Donsker-Klasse und wir können mit Lemma 1.4 und Lemma 2.3 folgern, dass $\mathcal{RK}_n(\mathcal{F}) \leq \frac{c}{\sqrt{n}}$ für ein $c > 0$. $(*_3)$
(In Zukunft wird jede Konstante, die nicht von n oder δ abhängt mit c bezeichnet.)

Nun wenden wir Satz 3.5 ([1] Thm. 11) an und erhalten für jedes $t > 0$:

$$2e^{-2t^2} \geq \mathbb{P}\left(\exists f \in \mathcal{F} :\ P(m_f \geq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_f \leq \delta) + \frac{48}{\delta}RK_n + \sqrt{\frac{\log\log_2(\frac{2}{\delta})}{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$\geq \mathbb{P}\left(P(m_{f_n} \geq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_{f_n} \leq \delta) + \frac{48}{\delta}RK_n + \sqrt{\frac{\log\log_2(\frac{2}{\delta})}{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$\overset{*_3}{\geq} \mathbb{P}\left(P(m_{f_n} \geq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_{f_n} \leq \delta) + \frac{c}{\delta\sqrt{n}} + \underbrace{\sqrt{\frac{\log\log_2(\frac{2}{\delta})}{n}}}_{\leq \frac{c}{\delta\sqrt{n}}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$\geq \mathbb{P}\left(P(m_{f_n} \geq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n(m_{f_n} \leq \delta) + \frac{c}{\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$\overset{*_2}{=} \mathbb{P}\left(P(Y\,\Phi_{\lambda_n,\sigma_n}(T_n)(X) \geq 0) > \inf_{\delta\in(0,1)}\left[\hat{P}_n\left(Y\,\Phi_{\lambda_n,\sigma_n}(T_n)(X) \leq \delta\, S_n\right) + \frac{c}{\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$\geq \mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n)) > \inf_{\delta\in(0,1)}\left[\hat{P}_n\left(l(\Phi_{\lambda_n,\sigma_n}(T_n)(X), Y) \geq (1-\delta\, S_n)^+\right) + \frac{c}{\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\Bigg)$$

$$\geq \mathbb{P}\left(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n)) > \inf_{\delta\in(0,\frac{1}{S_n})}\left[\frac{R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta S_n} + \frac{c}{\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$= \mathbb{P}\left(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n)) > \inf_{\delta\in(0,1)}\left[\frac{R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta} + \frac{c\,S_n}{\delta\sqrt{n}}\right] + \frac{t}{\sqrt{n}}\right)$$

$$= \mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n)) - R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n)) >$$

$$\inf_{\delta\in(0,1)}\left[\frac{R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta}-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))+\frac{c\,S_n}{\delta\sqrt{n}}\right]+\frac{t}{\sqrt{n}}\Bigg)$$

$$=\mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))>$$

$$\inf_{\delta\in(0,1)}\left[\frac{\delta\,R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta}+\frac{c\,S_n}{\delta\sqrt{n}}\right]+\frac{t}{\sqrt{n}}\Bigg)$$

Wir definieren $\delta_n := \frac{\sqrt{S_n}}{n^{\frac{1}{4}}}$ und erhalten

$$2e^{-2t^2}\geq\mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))>$$

$$\inf_{\delta\in(0,1)}\left[\frac{\delta\,R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta}+\frac{c\,S_n}{\delta\sqrt{n}}\right]+\frac{t}{\sqrt{n}}\Bigg)$$

$$\geq\mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))>$$

$$\frac{\delta_n\,R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta_n}+\frac{c\,S_n}{\delta_n\sqrt{n}}+\frac{t}{\sqrt{n}}\Bigg)$$

$$\geq\mathbb{P}\Bigg(R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))>$$

$$\frac{\delta_n\,R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}{1-\delta_n}+\frac{c\sqrt{S_n}}{n^{\frac{1}{4}}}+\frac{t}{\sqrt{n}}\Bigg)$$

$$\geq\mathbb{P}\Bigg(\underbrace{R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))}_{0\leq\cdot\leq 4}>\frac{c\sqrt{S_n}}{n^{\frac{1}{4}}}+\frac{t}{\sqrt{n}}\Bigg)$$

Durch Wahl von $t_n := n^{\frac{1+2\varepsilon}{4}}$ folgt leicht mit $*_1$:

$$\mathbb{E}\left[R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))\right]\leq c\,n^{-\frac{1-2\varepsilon}{4}}\qquad(*_4)$$

Nun wollen wir $\mathbb{E}\left[|R_{l,P}(\Phi_{\lambda_n,\sigma_n}(T_n))-R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))|\right]$ abschätzen. Dazu stellen wir wie im letzten Vortrag fest, dass $\Phi_{\lambda_n,\sigma_n}(T_n)$ in $\mathcal{F}_{r,\sigma} := B_{r_n}^{H_{\sigma_n}}(0)$ für $r_n = \sqrt{\frac{1}{\lambda_n}}$ liegt, und somit die $l^1-Variation$ durch $\frac{\sqrt{2}r_n}{\sigma_n}$ beschränkt ist. Aus der Lipschitzstetigkeit von l folgt, dass $l(\Phi_{\lambda_n,\sigma_n}(T_n),-1)$ und $l(\Phi_{\lambda_n,\sigma_n}(T_n),1)$

in l^1-Variation ebenfalls kleinergleich $\frac{\sqrt{2}r_n}{\sigma_n}$ sind. Damit sind $\frac{\sigma_n}{r_n} l(\Phi_{\lambda_n,\sigma_n}(T_n), -1)$ und $\frac{\sigma_n}{r_n} l(\Phi_{\lambda_n,\sigma_n}(T_n), 1)$ nach [2] Seite 329 in der universellen Donskerklasse $E_{\sqrt{2}}$ enthalten und es folgt

$$\begin{aligned}
&\mathbb{E}\left[|R_{l,P}(\Phi_{\lambda_n,\sigma_n}(T_n)) - R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in E_{\frac{\sqrt{2}r_n}{\sigma_n}}} |P(l(f(X),Y)) - \hat{P}_n(l(f(X),Y))|\right] \\
&= \mathbb{E}\left[\frac{r_n}{\sigma_n} \sup_{f\in E_{\frac{\sqrt{2}r_n}{\sigma_n}}} \left|P\left(\frac{\sigma_n}{r_n} l(f(X),Y)\right) - \hat{P}_n\left(\frac{\sigma_n}{r_n} l(f(X),Y)\right)\right|\right] \\
&\leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}}. \qquad (*_5)
\end{aligned}$$

Als nächstes wird der Term $\mathbb{E}\left[R_{l,P}(\Phi_{\lambda_n,\sigma_n}(T_n)) - \inf_{f\in H_{\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right)\right]$ abgeschätzt. Wir definieren die Verlustfunktion

$$\widetilde{l}_n(f,(x,y)) := \lambda_n||f||^2_{H_{\sigma_n}} + l(f(x),y).$$

Man sieht leicht, dass dann $||\widetilde{l}_n(f,(\cdot,1))||_{TV}, ||\widetilde{l}_n(f,(\cdot,-1))||_{TV} \leq ||f||_{TV}$ gilt. Damit gilt analog zu Satz 5.1, dass

$$\mathbb{E}\left[\lambda_n||\Phi_{r,\sigma}(T_n)||^2_{H_{\sigma_n}} + R_{l,P}(\Phi_{r,\sigma}(T_n)) - \inf_{f\in\mathcal{F}_{r_n,\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right)\right] \leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}} + \frac{c}{\sqrt{n}}.$$

Für $r_n = \sqrt{\frac{1}{\lambda_n}}$ ist das Infimum bereits das Infimum über ganz H_{σ_n} und wir folgern

$$\mathbb{E}\left[R_{l,P}(\Phi_{r,\sigma}(T_n)) - \inf_{f\in H_{\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right)\right] \leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}} + \frac{c}{\sqrt{n}}. \qquad (*_6)$$

Aus [7] Thm. 2.7 folgt sofort die Ungleichung

$$\inf_{f\in H_{\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f) \leq \frac{c\lambda_n}{\sigma_n} + c\sigma_n^{\alpha}. \qquad (*_7)$$

Mit diesen Vorüberlegungen berechnen wir:

$$\mathbb{E}\left[R_P(\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))) - \inf_{f:[0,1]\to\{-1,1\}} R_P(f)\right]$$

$$
\begin{aligned}
&= \underbrace{\mathbb{E}\left[R_P(\operatorname{sign}\circ\Phi_{\lambda_n,\sigma_n}(T_n)) - R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n))\right]}_{\leq c\, n^{-\frac{1-2\varepsilon}{4}} \quad (*_4)} \\
&\quad + \underbrace{\mathbb{E}\left[R_{l,\hat{P}_n}(\Phi_{\lambda_n,\sigma_n}(T_n)) - R_{l,P}(\Phi_{\lambda_n,\sigma_n}(T_n))\right]}_{\leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}} \quad (*_5)} \\
&\quad + \underbrace{\mathbb{E}\left[R_{l,P}(\Phi_{\lambda_n,\sigma_n}(T_n)) - \inf_{f\in H_{\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right)\right]}_{\leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}}+\frac{c}{\sqrt{n}} \quad (*_6)} \\
&\quad + \underbrace{\inf_{f\in H_{\sigma_n}}\left(\lambda_n||f||^2_{H_{\sigma_n}} + R_{l,P}(f)\right) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f)}_{\leq \frac{c\lambda_n}{\sigma_n}+c\sigma_n^\alpha \quad (*_7)} \\
&\leq c\,\max\left(n^{-\frac{1-2\varepsilon}{4}}, \frac{r_n}{\sigma_n\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{\lambda_n}{\sigma_n}, \sigma_n^\alpha\right) \quad , \text{ mit } \lambda_n = \frac{1}{r_n^2} \\
&\leq c\,\max\left(n^{-\frac{1-2\varepsilon}{4}}, \frac{r_n}{\sigma_n\sqrt{n}}, \frac{1}{\sigma_n r_n^2}, \sigma_n^\alpha\right)
\end{aligned}
$$

Wähle $r_n = n^{\frac{1}{6}}$ und $\sigma_n = n^{-\frac{1}{3(\alpha+1)}}$ wie in Bemerkung 5.4 und erhalte die Konvergenzgeschwindigkeit $n^{-\min\left(\frac{1-2\varepsilon}{4}, \frac{\alpha}{3(\alpha+1)}\right)}$. □

Literatur

[1] V. Koltchinskii and D. Panchenko, *Empirical margin distributions and bounding the generalization error of combined classifiers*, Ann. Statist. **30** (2002), no. 1, 1–50, DOI 10.1214/aos/1015362182.

[2] R. M. Dudley, *Uniform central limit theorems*, 2nd ed., Cambridge Studies in Advanced Mathematics, vol. 142, Cambridge University Press, New York, 2014.

[3] Michel Ledoux and Michel Talagrand, *Probability in Banach spaces*, Classics in Mathematics, Springer-Verlag, Berlin, 2011. Isoperimetry and processes; Reprint of the 1991 edition.

[4] Aad W. van der Vaart and Jon A. Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996. With applications to statistics.

[5] Evarist Giné and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series in Statistical and Probabilistic Mathematics, [40], Cambridge University Press, New York, 2016.

[6] Luc Devroye, László Györfi, and Gábor Lugosi, *A probabilistic theory of pattern recognition*, Applications of Mathematics (New York), vol. 31, Springer-Verlag, New York, 1996.

[7] Ingo Steinwart and Clint Scovel, *Fast rates for support vector machines using Gaussian kernels*, Ann. Statist. **35** (2007), no. 2, 575–607, DOI 10.1214/009053606000001226.