Diplomarbeit

Konversion von on-line Algorithmen in off-line Algorithmen beim Lernen von Funktionen

Norbert Klasner

14. März 1995

Betreuer: Prof. Dr. Hans Ulrich Simon

Inhaltsverzeichnis

Symbol	Bedeutung	s. Abschnitt
$\mathbb{R}$	reelle Zahlen	
$\mathbb{N}$	natürliche Zahlen	
$\lvert \cdot \rvert$	Absolutbetrag	
$\lVert \vec{x} \rVert_p$	p-Norm von $\vec{x}$	2
$(a.s.)$	fast überall	2
$\mathcal{A}, \mathcal{G}, \mathcal{G}_i$	σ-Algebren	
$\sigma(\mathcal{D})$	von $\mathcal{D}$ erzeugte σ-Algebra	2
$\sigma(\xi_1, \ldots, \xi_k)$	von $\xi_1, \ldots, \xi_k$ erzeugte σ-Algebra	5
P	Wahrscheinlichkeitsverteilung, -maß	2 und 3.2
$P(B)$	Wahrscheinlichkeit von B	
$P(B \mid C)$	bedingte Wahrscheinlichkeit von B bzgl. Ereignis C	2
$P(B \mid \mathcal{D})$	bedingte Wahrscheinlichkeit von B bzgl. endlicher oder abzählbarer Zerlegung $\mathcal{D}$	2
$P(B \mid \mathcal{G})$	bedingte Wahrscheinlichkeit von B bzgl. σ-Algebra $\mathcal{G}$	2
$E(\xi)$	Erwartungswert von ξ	
$E(\xi \mid \mathcal{G})$	bedingter Erwartungswert von ξ bzgl. $\mathcal{G}$	2
$E(\xi \mid \xi_1, \ldots, \xi_k)$	bedingter Erwartungswert von ξ bzgl. $\sigma(\xi_1, \ldots, \xi_k)$	2
Ω	Stichprobenraum	2
$\mathcal{H}$	Hypothesenraum	3
h, h_i	Hypothesen	3
h_0	Starthypothese	3.1
F	Menge (Klasse) von Zielfunktionen	3
f	(Ziel-)funktion	3
X	Instanzenraum	3
$L, L_p, L_0, L_{0,\gamma}$	Abstandsfunktionen	3
$opt_L(F)$	optimale Schranke für den Gesamtabstand	3.1
EX	Orakel	3.2
$Error(h)$	Fehler von h	3.2
$e_L(h), e_1(h), e_2(h)$	erwartete Abstände	3.2
δ	Überzeugungsparameter	3.2
ϵ, γ	Genauigkeitsparameter	3.2

Tabelle 1: Die verwendeten Symbole.

Symbol	Bedeutung	s. Abschnitt
g	Schranke für die maximale Anzahl verschiedener Hypothesen	5.2
D	maximal möglicher Abstand zwischen Vorhersage und tatsächlichem Wert	5.2
$\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_\infty$	Klassen glatter Funktionen mit beschränkter Steigung	6.1
$F_n(M,W), G_n(M,W)$	Klassen linearer Funktionen	6.2
M	maximale Norm des Instanzenvektors	6.2
W	(maximale) Norm des Gewichtsvektors	6.2
$VCdim(C)$	Vapnik-Chervonenkis Dimension	8
$Pdim(F)$	Pseudo-Dimension	8
$fat_F(\gamma)$	„fat-shattering function“	8
$d_F(\gamma)$	Dimension aus [Sim93]	8
$\Delta(F)$	maximaler Abstand von Funktionswerten in Klasse F	9

Tabelle 2: Die verwendeten Symbole (Fortsetzung).

1 Einführung

Es gibt bereits mehrere Transformationen, die on-line Lernalgorithmen für Boolesche Funktionen auf PAC-Lernalgorithmen abbilden. So gibt es z. B. die Strategie, on-line Lernalgorithmen mit einer Transformation von Littlestone [Lit88] in Lernalgorithmen mit Equivalence Fragen zu transformieren und diese dann mit einer Transformation von Angluin [Ang88] in PAC-Lernalgorithmen umzuwandeln. Diese Strategie ist ebenso wie die „Longest-Survivor“ Strategie, subobtimal. Littlestone [Lit89] hat eine für einige Funktionenklassen bis auf konstante Faktoren optimale Transformation entwickelt.

Hier betrachten wir statt Boolescher Funktionen jedoch reelwertige (bzw. beliebige) Funktionen. Unser on-line Modell entspricht dem von Littlestone [Lit88] für Funktionen $f : \{0,1\}^n \to \{0,1\}$ eingeführten on-line Modell und unsere off-line Modelle sind einfache Erweiterungen der verteilungsfreien off-line Modelle aus [Sim93], die sich an die Lernmodelle von Kearns und Schapire [KS94] für p-Konzepte anlehnen.

Wir stellen Transformationen von obigem on-line Modell in obige verteilungsfreie off-line Modelle vor. Diese Transformationen sind Erweiterungen der oben genannten Transformation aus [Lit89]. Wir zeigen obere Schranken für die Anzahl der benötigten Trainingsbeispiele bei Anwendung der Transformationen, selbst dann, wenn die benutzten on-line Lernalgorithmen keine konstante Schranke für den Gesamtabstand haben, sondern deren Gesamtabstand nach t Runden nur durch $O(\ln t)$ oder gar nur durch $O(t^{1/q})$, für $q > 1$ beschränkt ist.

Wir wenden die Transformationen auf stückweise glatte Funktionen mit beschränkter Steigung und auf lineare Funktionen an und erhalten so Algorithmen zum verteilungsfreien off-line Lernen dieser Klassen. Dabei benutzen wir on-line Resultate von Kimber und Long [KL92], Kivinen und Warmuth [KW94], sowie Bernstein [Ber92], um obere Schranken für die Anzahl der benötigten Trainingsbeispiele zu erhalten.

Um die so erhaltenen oberen Schranken einordnen zu können berechnen wir untere Schranken für die Anzahl der benötigten Trainingsbeispiele beim verteilungsfreien off-line Lernen der obigen Funktionenklassen. Diese berechnen wir mit den allgemeinen unteren Schranken für die Anzahl der benötigten Trainingsbeispiele von Simon [Sim93].

Der abschließende Vergleich zwischen oberen und unteren Schranken für die Anzahl der benötigten Trainingsbeispiele zeigt dann, daß die Transformationen auf den obigen Funktionenklassen bis auf logarithmische oder sogar bis auf konstante Faktoren (bzgl. der Anzahl der Trainingsbeispiele) optimale Lernalgorithmen für zumindest einen Teil der verteilungsfreien off-line Modelle liefern.

2 Vorbemerkungen

Für $\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ und $p > 0$ sei die *p-Norm* von $\vec{x}$ durch

$$\|\vec{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

definiert.

Für $1 \leq i \leq n$ und $1 \leq j \leq n$ mit $i \neq j$ seien $\vec{e}_i^{(n)} = (e_{i,1}, \ldots, e_{i,n}) \in \mathbb{R}^n$ die Vektoren mit $e_{i,i} = 1$ und $e_{i,j} = 0$. Die Vektoren $\vec{e}_i^{(n)}$ bilden die kanonische Basis von $\mathbb{R}^n$. Falls n eindeutig ist, schreiben wir $\vec{e}_i$ statt $\vec{e}_i^{(n)}$.

Für Funktionen f, die auf $[0, 1]$ definiert sind sei die *p-Norm* für $1 \leq p < \infty$ definiert durch

$$\left(\int_0^1 |f(x)|^p \, dx \right)^{1/p}$$

und für $p = \infty$ definiert durch

$$\lim_{q\to\infty} \left(\int_0^1 |f(x)|^q \, dx \right)^{1/q} .$$

Wir benötigen noch einige Konstruktionen aus der Wahrscheinlichkeitslehre, die hier nur kurz angerissen werden können. Für eine ausführlichere Abhandlung sei auf [Shi84] verwiesen.

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. (Dabei ist Ω der Stichprobenraum, $\mathcal{A}$ eine σ-Algebra und P eine Wahrscheinlichkeitsverteilung (Wahrscheinlichkeitsmaß).) Eine Eigenschaft gilt *fast überall* (kurz (*a.s.*) für „*almost surely*", falls eine Menge $\mathcal{N} \in \mathcal{A}$ mit $P(\mathcal{N}) = 0$ existiert, so daß die Eigenschaft für alle $\omega \in \Omega \setminus \mathcal{N}$ gilt.

Sei $C \in \mathcal{A}$ ein Ereignis mit $P(C) > 0$. Die *bedingte Wahrscheinlichkeit von B bzgl. C* wird bezeichnet durch $P(B|C)$ und ist definiert durch $P(BC)/P(C)$. Sei $\mathcal{D}$ eine endliche oder abzählbare Zerlegung $\{D_1, D_2, \dots\}$ von Ω, d. h. $D_i \neq \emptyset$ für alle i, $D_i \cap D_j = \emptyset$ für alle $i \neq j$ und $\bigcup_i D_i = \Omega$. Die *bedingte Wahrscheinlichkeit bzgl. der endlichen oder abzählbaren Zerlegung $\mathcal{D} = \{D_1, D_2, \dots\}$* mit $P(D_i) > 0$ für alle i, bezeichnet durch $P(B|\mathcal{D})$, ist die Wahrscheinlichkeitsvariable, die für $\omega \in D_i$ gleich $P(B|D_i)$ ist. D. h.

$$P(B|\mathcal{D}) = \sum_i P(B|D_i) I_{D_i}(\omega),$$

wobei

$$I_{D_i}(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{falls } \omega \notin A \end{cases} .$$

Betrachten wir alle Vereinigungen (incl. unendliche) der Mengen in $\mathcal{D}$, dann ist die resultierende Menge von Mengen zusammen mit der leeren Menge eine σ-Algebra. Wir bezeichnen diese durch $\sigma(\mathcal{D})$. Umgekehrt gibt es zu jeder σ-Algebra $\mathcal{G} \subseteq \mathcal{A}$ genau eine Zerlegung $\mathcal{D}$ von Ω. Sei $\mathcal{D}$ die Zerlegung zu der σ-Algebra $\mathcal{G}$, dann sei die *bedingte Wahrscheinlichkeit von B bzgl. der σ-Algebra $\mathcal{G}$*, bezeichnet durch $P(B|\mathcal{G})$, gleich $P(B|\mathcal{D})$.

Sei $\mathcal{G} = \{\emptyset, \Omega\}$ (triviale σ-Algebra), dann ist $\mathcal{D} = \{\Omega\}$ (triviale Zerlegung) die zugehörige Zerlegung. Es gilt dann $P(B|\mathcal{G}) = P(B|\Omega) = P(B)$.

Die *bedingte Erwartung* einer nichtnegativen Zufallsvariablen ξ bzgl. der σ-Algebra $\mathcal{G}$ ist eine nichtnegative erweiterte[1] Zufallsvariable, bezeichnet durch $E(\xi|\mathcal{G})$ oder $E(\xi|\mathcal{G})(\omega)$, so daß gilt:

[1] D. h. sie darf auch den Wert ∞ annehmen.

1. $E(\xi|\mathcal{G})$ ist $\mathcal{G}$-meßbar.

2. Für jedes $A \in \mathcal{G}$ gilt
$$\int_A \xi \, dP = \int_A E(\xi|\mathcal{G}) \, dP.$$

Sei nun ξ irgendeine Zufallsvariable und $\xi^+ = \max\{\xi, 0\}$, $\xi^- = -\min\{\xi, 0\}$. Die bedingte Erwartung $E(\xi|\mathcal{G})$ oder $E(\xi|\mathcal{G})(\omega)$ einer Zufallsvariablen ξ bzgl. der σ-Algebra $\mathcal{G}$ ist, falls
$$\min\{E(\xi^+|\mathcal{G}), E(\xi^-|\mathcal{G})\} < \infty \quad (a.s.),$$
definiert durch
$$E(\xi|\mathcal{G}) = E(\xi^+|\mathcal{G}) - E(\xi^-|\mathcal{G}),$$
wobei auf der Menge (mit Wahrscheinlichkeit 0) der Punkte für die $E(\xi^+|\mathcal{G}) = E(\xi^-|\mathcal{G}) = \infty$ gilt, $E(\xi^+|\mathcal{G}) - E(\xi^-|\mathcal{G})$ auf einen beliebigen Wert, z. B. 0, gesetzt wird. Durch diese Eigenschaften wird $E(\xi|\mathcal{G})$ $(a.s.)$ eindeutig bestimmt.

Für Zufallsvariablen $\xi, \xi_1, \ldots, \xi_k$, sei $\mathcal{G}$ die durch $\xi_1, \ldots, \xi_k$ erzeugte σ-Algebra, dann bezeichne $E(\xi|\xi_1, \ldots, \xi_k)$ die bedingte Erwartung $E(\xi|\mathcal{G})$.

Beispiel: Sei $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ eine Zerlegung von Ω, so daß $\mathcal{G} = \sigma(\mathcal{D}) \subseteq \mathcal{A}$. Sei ξ eine nichtnegative Zufallsvariable. Wir setzen nun
$$E_i = \frac{1}{P(D_i)} \int_{D_i} \xi \, dP$$
für $i = 1, \ldots, n$. Also ist E_i der Erwartungswert, wenn $\omega \in D_i$. Nun ist
$$\zeta = \sum_{i=1}^{n} E_i I_{D_i}(\omega)$$
eine nichtnegative erweiterte Wahrscheinlichkeitsvariable. ζ ist offensichtlich $\mathcal{G}$-meßbar. Jedes $A \in \mathcal{G}$ läßt sich als (evtl. leere) disjunkte Vereinigung von Mengen aus $\mathcal{D}$ schreiben. Mit $\mathcal{D} = \{D_1, \ldots, D_n\}$ gilt für jedes $A \in \mathcal{G}$
$$\int_A \xi \, dP = \sum_{i:D_i \subseteq A} \int_{D_i} \xi \, dP = \sum_{i:D_i \subseteq A} E_i P(D_i) = \sum_{i:D_i \subseteq A} \int_{D_i} E_i \, dP$$
$$= \sum_{i:D_i \subseteq A} \int_A E_i I_{D_i} \, dP = \int_A \sum_{i:D_i \subseteq A} E_i I_{D_i} \, dP = \int_A \sum_{i=1}^{n} E_i I_{D_i} \, dP = \int_A \zeta \, dP.$$
Somit ist ζ die bedingte Erwartung von ξ bzgl. $\mathcal{G}$ oder kurz: $\zeta = E(\xi|\mathcal{G})$. $\diamond$

Für eine Wahrscheinlichkeitsvariable ξ, die $\mathcal{G}$-meßbar ist, gilt $E(\xi|\mathcal{G}) = \xi$ $(a.s.)$. (Beachte, daß ξ^+ und ξ^-, beide wie oben definiert, in diesem Fall ihre eigenen bedingten Erwartungen sind.)

Sei $(\Omega, \mathcal{A}, P)$ ein gegebener Wahrscheinlichkeitsraum und sei $\mathcal{G}_0 \subseteq \cdots \subseteq \mathcal{G}_n \subseteq \mathcal{A}$ für $n > 0$ eine Folge von σ-Algebren. Sei $T_0, \ldots, T_n$ eine Folge von auf $(\Omega, \mathcal{A}, P)$ definierten Zufallsvariablen und T_i weiterhin $\mathcal{G}_i$-meßbar für jedes $0 \leq i \leq n$, dann heißt $((T_0, \mathcal{G}_0), \ldots, (T_n, \mathcal{G}_n))$ *Martingale*, falls für alle $0 \leq i \leq n$

$$E(|T_i|) < \infty$$

und für alle $0 < i \leq n$

$$E(T_i|\mathcal{G}_{i-1}) = T_{i-1} \quad (a.s.)$$

gilt. Ersetzen wir die letzte Gleichung durch die Ungleichung

$$E(T_i|\mathcal{G}_{i-1}) \leq T_{i-1} \quad (a.s.),$$

so sprechen wir von einer *Supermartingale*. (Martingalen und Supermartingalen werden meistens als unendliche Folgen definiert, jedoch sind endliche Folge hier ausreichend.)

3 Die Lernmodelle

3.1 On-line Modell für Funktionen

Sei F die Menge der Zielfunktionen $f : X \rightarrow Y$. Dann heißt X Instanzenraum und $\mathcal{H} \subseteq \{h \mid h : X \rightarrow Y\}$ Hypothesenraum. Desweiteren existiere eine Repräsentationssprache R, so daß sich jedes $h \in \mathcal{H}$ in R repräsentieren läßt. Ein *on-line Lernalgorithmus* A für F mit einer festen Starthypothese $h_0 \in \mathcal{H}$ arbeitet in Runde $i \geq 1$ wie folgt.

1. A erhält eine Instanz $x_i \in X$.
2. A macht die Vorhersage $\hat{y}_i = h_{i-1}(x_i)$.
3. A erhält den *tatsächlichen Wert* $y_i = f(x_i)$ für ein beliebiges aber festes, dem Lernalgorithmus unbekanntes, $f \in F$.
4. A berechnet eine Hypothese $h_i \in \mathcal{H}$, die zur Vorhersage in der nächsten Runde benutzt wird.

Wir haben den Hypothesenraum $\mathcal{H}$ hier nicht weiter eingeschränkt, jedoch gilt z. B. für die on-line Algorithmen $LININT$ (s. Abschnitt 6.1) und die „EG-Algorithmen“ (s. Abschnitt 6.2) $\mathcal{H} = F$.

Formal läßt sich ein on-line Lernalgorithmus auch als Funktion A:$(X \times Y)^* \times X \to Y$ mit $A(S, x_i) = \hat{y}_i$ schreiben, wobei S die Folge der Instanz-Funktionswert-Paare der ersten $i-1$ Runden, x_i die Instanz der Runde i und $\hat{y}_i = h_{i-1}(x_i)$ die Vorhersage ist.

In jeder Runde läßt sich somit der *Abstand* zwischen Vorhersage und tatsächlichem Wert bzgl. einer Funktion $L : Y \times Y \to \mathbb{R}^{\geq 0}$ berechnen. Sinnvollerweise sollte $L(y, y) = 0$ für alle $y \in Y$ gelten. Der *Gesamtabstand* für eine Folge von Instanzen und Funktionswerten ist die Summe der Abstände der einzelnen Runden. Typische Abstandsfunktionen von $\mathbb{R} \times \mathbb{R}$ nach $\mathbb{R}^{\geq 0}$ sind die Funktionen

$$L_p(\hat{y}, y) = |\hat{y} - y|^p$$

für $p > 0$. L_1 wird auch *linearer Abstand* und L_2 *quadratischer Abstand* genannt.

Weitere in dieser Arbeit benutzte Abstandsfunktionen sind

$$L_0(\hat{y}, y) = \lim_{p \to 0} L_p(\hat{y}, y) = \begin{cases} 1 & \text{falls } \hat{y} \neq y \\ 0 & \text{falls } \hat{y} = y \end{cases}$$

und

$$L_{0,\gamma}(\hat{y}, y) = \begin{cases} 1 & \text{falls } |\hat{y} - y| \geq \gamma \\ 0 & \text{falls } |\hat{y} - y| < \gamma \end{cases} .$$

Das Ziel eines on-line Lernalgorithmus ist, das Supremum des Gesamtabstandes (bzgl. L) auf beliebigen endlichen Folgen (oder zumindest auf Folgen mit einer Maximallänge t) zu minimieren.

Beim on-line Lernen wird also nicht zwischen Arbeits- und Lernphase unterschieden. Insbesondere muß die Vorhersage in Runde 1 ohne gelernte Information über das Zielkonzept f auskommen.

Wir sagen *A lernt F im on-line Modell mit $l(t)$-beschränktem Abstand (bzgl. L)*, wenn für alle $f \in F$ und alle Beispielfolgen $S_t = ((x_i, f(x_i)))_{1 \leq i \leq t}$ der Länge t der Gesamtabstand (bzgl. L), den A verursacht, kleiner oder gleich $l(t)$ ist. Falls es eine Schranke l unabhängig von t gibt, so daß $l(t) \leq l$ für alle $t \in \mathbb{N}$, also der Gesamtabstand (bzgl. L) auf allen endlichen Beispielfolgen durch l beschränkt ist, so sagen wir *A lernt F im on-line Modell mit l-beschränktem Abstand (bzgl. L)*. Insbesondere sagen wir A

lernt F im on-line Modell mit $l(t)$-beschränktem (bzw. l-beschränktem) linearem Abstand, falls F von A im on-line Modell mit $l(t)$-beschränktem (bzw. l-beschränktem) Abstand (bzgl. L_1) gelernt wird und *A lernt F im on-line Modell mit $l(t)$-beschränktem (bzw. l-beschränktem) quadratischem Abstand*, falls F von A im on-line Modell mit $l(t)$-beschränktem (bzw. l-beschränktem) Abstand (bzgl. L_2) gelernt wird.

Die optimale Schranke (evtl. ∞) für die Größe des „worst case" Gesamtabstandes (bzgl. L) beim on-line Lernen von F ist durch

$$opt_L(F) = \inf_A \left\{ l_A \;\middle|\; \begin{array}{l} A \text{ lernt } F \text{ im on-line Modell mit } l_A\text{-} \\ \text{beschränktem Abstand (bzgl. } L) \end{array} \right\}$$

definiert.

Eine Vorhersage ist *falsch*, wenn sie nicht mit dem tatsächlichen Wert übereinstimmt.

Für Mengen F von Zielfunktionen $f : X \to \mathbb{R}$ gelten die nun folgenden Definitionen.

Ein on-line Lernalgorithmus A *irrt* sich (bzgl. $\gamma > 0$), falls der lineare Abstand zwischen seiner Vorhersage $\hat{y}$ und dem tatsächlichen Wert y mindestens γ beträgt, also $L_{0,\gamma}(\hat{y}, y) = 1$.

Analog, eine Hypothese h *irrt* sich auf x bzgl. einem Abstandsparameter $\gamma > 0$ und einem Zielkonzept f, falls $L_{0,\gamma}(h(x), f(x)) = 1$.

Die Anzahl der Irrtümer, die ein on-line Lernalgorithmus begeht, ist also der Gesamtabstand bzgl. $L_{0,\gamma}$.

Somit läßt sich, ähnlich wie bei on-line Lernalgorithmen für Boolesche Konzepte, Konservativität [Lit88] definieren. Ein on-line Lernalgorithmus A ist *konservativ* (bzgl. γ), falls er seine Hypothese jeweils nur dann ändert, wenn er sich (bzgl. γ) geirrt hat.

Der Abstandsparameter γ ist für die Konservativität im Allgemeinen notwendig, da im Gegensatz zum „Booleschen Fall" nicht für jede Funktionenklasse erwartet werden kann, daß die Vorhersage exakt mit dem tatsächlichen Wert übereinstimmt, so daß sich aus der Konservativität kein Nutzen ziehen lassen würde.

3.2 Verteilungsfreie off-line Modelle für Funktionen

Sei F eine Klasse von Funktionen von X nach Y. Sei $f \in F$ die (für den Lerner unbekannte) Zielfunktion und P eine unbekannte Wahrscheinlichkeitsverteilung auf X.[2]

[2]Formal: P ist ein Wahrscheinlichkeitsmaß für $(X, \mathcal{A})$, wobei $\mathcal{A}$ eine σ-Algebra auf X sei. Wir nehmen an, daß alle Funktionen aus F und alle benutzten Hypothesen h meßbare Funktionen von $(X, \mathcal{A})$

Auch hier sei der Hypothesenraum $\mathcal{H} \subseteq \{h \mid h : X \to Y\}$ und es existiere eine Repräsentationssprache R, so daß sich jedes $h \in \mathcal{H}$ in R repräsentieren läßt. Wir schränken wiederum den Hypothesenraum nicht weiter ein, jedoch werden unsere Transformationen (s. Abschnit 5) den Hypothesenraum des benutzten on-line Algorithmus auf den off-line Algorithmus übertragen.

EX heiße *Orakel* für f (bzgl. P), falls EX zufällig ein $x \in X$ (bzgl. der Wahrscheinlichkeitsverteilung P) zieht und dann das Trainingsbeispiel $(x, f(x))$ liefert. Bei mehrmaligem Aufruf des Orakels müssen außerdem die Instanzen unabhängig voneinander gezogen werden. Eine *Probe* I der Länge t (bzgl. P, f) ist eine Folge von t durch EX gelieferten Trainingsbeispielen. $Error(h)$ bezeichne den *Fehler* einer Hypothese $h : X \to \mathbb{R}$ bzgl. P, f und einem Abstandsparameter γ und ist definiert durch

$$Error(h) = P(\{x \mid |h(x) - f(x)| \geq \gamma\}).$$

Eine Funktion $h : X \to \mathbb{R}$ ist ein (ϵ, γ)-*gutes Modell* (bzgl. P) für f, falls $Error(h) < \epsilon$.

Der *erwartete Abstand* von h (bzgl. einer Funktion $L : Y^2 \to \mathbb{R}^{\geq 0}$ für den Abstand, sowie P, f) ist definiert durch

$$e_L(h) = E(L(h(x), f(x))) = \int_X L(h(x), f(x))\, dP.$$

Für $Y = \mathbb{R}$ sind der *erwartete absolute Abstand* und der *erwartete quadratische Abstand* von h (bzgl. P, f) durch

$$e_1(h) = E(|f(x) - h(x)|) = \int_X |f(x) - h(x)|\, dP$$

und

$$e_2(h) = E((f(x) - h(x))^2) = \int_X (f(x) - h(x))^2\, dP$$

definiert.

Ein *Lernalgorithmus* A für F mit einer Probenlänge $t = t(\epsilon, \gamma, \delta)$ erhält als Eingabe die Parameter ϵ, γ, δ und eine Probe I (bzgl. P, f) der Länge t bzw. die Möglichkeit t mal das Orakel EX aufzurufen. Er gibt eine Hypothese $h_I \in \mathcal{H}$ (genau genommen die Repräsentation bzgl. R für ein $h_I \in \mathcal{H}$) aus. Da die Ausgabe h_I von der zufällig gezogenen Probe I abhängt, ist es sinnvoll, von der Wahrscheinlichkeit (bzgl. P^t), daß h_I eine

nach $(Y, \mathcal{B})$ sind. Dabei sei $\mathcal{B}$ eine σ-Algebra auf Y und zwar für Y abzählbar: $\mathcal{B} = 2^Y$; $Y \subseteq \mathbb{R}$ (und überabzählbar): $\mathcal{B}$ ist die Standard-Algebra der Borel Mengen aus Y. Diese Annahmen garantieren, daß alle in dieser Arbeit auftretenden Wahrscheinlichkeiten und Erwartungen, wohldefiniert sind.

gewisse Eigenschaft hat, zu sprechen. Bei einigen der folgend definierten Lernmodelle fehlt der Parameter γ.

Wir sagen *A lernt F mit einem* (ϵ,γ)*-guten Modell*, wenn für alle $f \in F$, alle Verteilungen P auf X, alle $0 < \epsilon < 1$, $0 < \delta < 1$, $0 < \gamma$ und $t = t(\epsilon,\gamma,\delta)$ die Wahrscheinlichkeit, daß h_I ein (ϵ,γ)-gutes Modell für f ist, mindestens $1-\delta$ (bzgl. P^t) ist.

Ähnlich sagen wir, *A lernt F mit* ϵ*-beschränkter Differenz (bzgl. L)*, wenn für alle Verteilungen P auf X, alle $\epsilon > 0$, $0 < \delta < 1$ und $t = t(\epsilon,\delta)$ die Wahrscheinlichkeit, daß $e_L(h_I) < \epsilon$ ist, mindestens $1-\delta$ (bzgl. P^t) ist.

Analog, *A lernt F mit* ϵ*-beschränkter absoluter (bzw. quadratischer) Differenz*, wenn für alle Verteilungen P auf X, alle $\epsilon > 0$, $0 < \delta < 1$ und $t = t(\epsilon,\delta)$ die Wahrscheinlichkeit, daß $e_1(h_I) < \epsilon$ (bzw. $e_2(h_I) < \epsilon$) ist, mindestens $1-\delta$ (bzgl. P^t) ist.

Diese Lernalgorithmen haben also im Gegensatz zu den on-line Lernalgorithmen getrennte Lern- und Arbeitsphasen, d. h. sie nehmen sich zuerst t Trainingsbeispiele und erzeugen daraus eine Hypothese, die sich nicht mehr ändert.

Falls $F \subseteq \{f \mid f : X \to \{0,1\}\}$ und $\mathcal{H} = F$, dann sind alle Abstandsfunktionen L_p, $L_{0,\gamma}$ mit $p > 0$, $0 < \gamma \leq 1$ identisch zu L_0 und wir erhalten Valiants PAC-Lernmodell [Val84].

4 Einige Hilfsresultate

Für die Analyse unserer Transformationen benutzen wir die Chernoff Schranken und Verallgemeinerungen davon. Grob gesprochen liefern die Chernoff Schranken und deren Verallgemeinerungen eine Abschätzung für die Wahrscheinlichkeit, daß bei einer Folge von Zufallsversuchen die Summe der Ausgänge vom entsprechenden Erwartungswert stärker als eine Schranke abweicht. Bei den Chernoff Schranken selbst handelt es sich bei den Zufallsversuchen um unabhängige Bernoulli Versuche, die alle die gleiche Erfolgswahrscheinlichkeit haben[3]. Bei der Verallgemeinerung von Hoeffding (Hoeffding Ungleichungen) [Hoe63] muß es sich nicht mehr um Bernoulli Versuche, also Zufallsvariablen mit den Ausgängen 0 und 1, handeln, sondern es sind beliebige Zufallsvariablen $0 \leq X_i \leq 1$ erlaubt. Außerdem müssen die Zufallsvariablen der einzelnen Runden

[3] Betrachte z. B. eine Münze, die mit Wahrscheinlichkeit p Zahl (für Erfolg) und Wahrscheinlichkeit $1-p$ Kopf (für Mißerfolg) liefert. Es wird nun mehrere Male mit dieser Münze geworfen und die Anzahl der Erfolge gezählt.

nicht den gleichen Erwartungswert (und damit auch nicht die gleiche Verteilungsfunktion) haben. Allerdings müssen die einzelnen Wahrscheinlichkeitsvariablen immer noch unabhängig voneinander sein[4]. Littlestone [Lit89] hat nun darauf aufbauend Ungleichungen für den Fall gezeigt, daß statt der Unabhängigkeit die schwächere Bedingung, daß für Wahrscheinlichkeitsvariablen $0 \leq Y_i \leq 1$ die Folge $T_m = \sum_{i=1}^{m} X_i - \sum_{i=1}^{m} Y_i$ für $m = 1, \ldots, n$ eine Supermartingale ist, gilt.

Lemma 4.1 (Hoeffding Ungleichungen (additive Form) [Hoe63]) *Seien $X_1, \ldots, X_t$ unabhängige Wahrscheinlichkeitsvariablen mit $0 \leq X_i \leq 1$, $T = \sum_{i=1}^{t} X_i$, $p = E(T/t)$, dann seien $LE(p,t,r) = P(T \leq r)$ und $GE(p,t,r) = P(T \geq r)$. Nun gilt für $0 < c < p$*

$$
\begin{aligned}
LE(p,t,(p-c)t) &\leq \left(\left(\frac{1-p}{1-p+c} \right)^{1-p+c} \left(\frac{p}{p-c} \right)^{p-c} \right)^t \\
&\leq e^{-2tc^2}
\end{aligned}
$$

und für $0 < c < 1 - p$

$$
\begin{aligned}
GE(p,t,(p+c)t) &\leq \left(\left(\frac{p}{p+c} \right)^{p+c} \left(\frac{1-p}{1-p-c} \right)^{1-p-c} \right)^t \\
&\leq e^{-2tc^2}.
\end{aligned}
$$

Lemma 4.2 (Hoeffding Ungleichungen (multiplikative Form)) *Seien $X_1, \ldots, X_t$ unabhängige Wahrscheinlichkeitsvariablen mit $0 \leq X_i \leq 1$, $T = \sum_{i=1}^{t} X_i$, $p = E(T/t)$. Nun gilt für $0 < \alpha < 1$*

$$LE(p,t,(1-\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{2}}$$

und für $0 < \alpha < (1-p)/p$ und $\alpha \leq 1$

$$GE(p,t,(1+\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{3}}.$$

Beweis: s. Anhang A.1. □

[4]Hoeffding hat angegeben, daß seine Ungleichungen auch gelten, wenn die Unabhängigkeit durch die schwächere Bedingung, daß die Folge $T_m = \sum_{i=1}^{m} X_i - E(\sum_{i=1}^{m} X_i)$, $m = 1, \ldots, n$ eine Martingale ist, ersetzt wird.

Korollar 4.3 (Chernoff Schranken) *Für $0 \leq p \leq 1$ und $t \in \mathbb{N}$, bezeichne $LE(p,t,r)$ die Wahrscheinlichkeit, daß in t unabhängigen Bernoulli Versuchen, jeder mit Erfolgswahrscheinlichkeit p, höchstens r Erfolge eintreten; $GE(p,t,r)$ bezeichne entsprechend die Wahrscheinlichkeit für mindestens r Erfolge*[5]*. Nun gilt für $0 < \alpha < 1$*

$$LE(p,t,(1-\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{2}}$$

und für $0 < \alpha < (1-p)/p$ und $\alpha \leq 1$

$$GE(p,t,(1+\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{3}}.$$

Satz 4.4 ([Lit89]) *P sei ein Wahrscheinlichkeitsmaß für $(\Omega, \mathcal{A})$ und $\mathcal{G} \subseteq \mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \cdots \subseteq \mathcal{G}_t$ seien σ-Algebren enthalten in $\mathcal{A}$. $\xi_1, \ldots, \xi_t$ sei eine Folge von Zufallsvariablen auf diesem Wahrscheinlichkeitsmaß, so daß $0 \leq \xi_i \leq 1$ für $i = 1, \ldots, t$. $c_1, \ldots, c_t$ seien $\mathcal{G}$-meßbare Zufallsvariablen mit $0 \leq c_i \leq 1$ für $i = 1, \ldots, t$. Sei*

$$T_j = \sum_{i=1}^{j} (\xi_i - c_i)$$

für $1 \leq j \leq t$ und $T_0 = 0$. Sei $\mu = \frac{1}{t}\sum_{j=1}^{t} c_j$. Sei α eine $\mathcal{G}$-meßbare Zufallsvariable. Dann gelten, falls die Folge $((T_0, \mathcal{G}_0), \ldots, (T_t, \mathcal{G}_t))$ eine Supermartingale ist, die folgenden Aussagen an fast jedem Punkt[6] *aus Ω mit $0 < \mu < 1$.*

1. *Wenn $0 < \alpha < 1 - \mu$, dann*

$$P(T_t \geq \alpha t | \mathcal{G}) \leq \left(\left(\frac{\mu}{\mu + \alpha} \right)^{\mu + \alpha} \left(\frac{1 - \mu}{1 - \mu - \alpha} \right)^{1 - \mu - \alpha} \right)^t.$$

2. *Wenn $0 \leq \alpha \leq \min\{\mu, 1 - \mu\}$, dann*

$$P(T_t \geq \alpha t | \mathcal{G}) \leq e^{-\frac{\alpha^2 t}{3\mu}}.$$

3. *Wenn $0 \leq \alpha \leq 1 - \mu$, dann*

$$P(T_t \geq \alpha t | \mathcal{G}) \leq e^{-\frac{\alpha^2 t}{2(1-\mu)}}.$$

[5]Beachte, daß diese Definition von LE und GE konsistent zu der Definition in Lemma 4.1 ist.
[6]im Sinne von $(a.s.)$

Wir erhalten nun mit der 3. Aussage des obigen Satzes das folgende (für unsere Zwecke leichter anzuwendende) Korollar.

Korollar 4.5 *Seien $X_1, \ldots, X_t$ Wahrscheinlichkeitsvariablen mit $0 \leq X_i \leq 1$ und $c \in [0,1]$. Falls X_i für alle $1 \leq i \leq t$ nur von $X_1, \ldots, X_{i-1}$ abhängt und $E(X_1) \geq c$ (a.s.), sowie $E(X_i|X_1, \ldots, X_{i-1}) \geq c$ (a.s.) für $2 \leq i \leq t$ ist, dann gilt (a.s.) für $0 \leq \alpha \leq 1$*

$$P\left(\sum_{i=1}^{t} X_i \leq (1-\alpha)tc\right) \leq e^{-\frac{\alpha^2 tc}{2}}.$$

Beweis: Für $c = 0$ ist die Aussage trivial, so daß wir im Folgenden $c > 0$ annehmen können. Wir zeigen mit Hilfe von Satz 4.4 zuerst die Aussage:

> Seien $X'_1, \ldots, X'_t$ Wahrscheinlichkeitsvariablen mit $0 \leq X'_i \leq 1$ und $c' \in [0,1]$. Falls X'_i für alle $1 \leq i \leq t$ nur von $X'_1, \ldots, X'_{i-1}$ abhängt und $E(X'_1) \leq c'$ $(a.s.)$, sowie $E(X'_i|X'_1, \ldots, X'_{i-1}) \leq c'$ $(a.s.)$ für $2 \leq i \leq t$ ist, dann gilt $(a.s.)$ für $0 \leq \alpha' \leq 1 - c'$
> $$P\left(\sum_{i=1}^{t} X'_i \geq (c'+\alpha')t\right) \leq e^{-\frac{\alpha'^2 t}{2(1-c')}}.$$

Dann zeigen wir, daß aus dieser Aussage das Korollar folgt.

Für $1 \leq i \leq t$ sei $\xi_i = X'_i$, $c_i = c'$ und $\mathcal{G}_i = \sigma(X'_1, \ldots, X'_i)$. Weiterhin setzen wir $\alpha = \alpha'$ und $\mathcal{G} = \mathcal{G}_0 = \{\emptyset, \Omega\}$. Alle c_i und α sind offensichtlich $\mathcal{G}$-meßbar und alle ξ_i sind $\mathcal{G}_i$-meßbar. Wir setzen nun

$$T_j = \sum_{i=1}^{j}(\xi_i - c_i) = \xi_j - c' + T_{j-1}$$

für $1 \leq j \leq t$ und $T_0 = 0$. Weiterhin haben wir $\mu = \frac{1}{t}\sum_{i=1}^{t} c_i = c'$. Um Satz 4.4 anwenden zu können müssen wir nur noch zeigen, daß $((T_0, \mathcal{G}_0), \ldots, (T_t, \mathcal{G}_t))$ eine Supermartingale ist. Jedes T_j ist offensichtlich $\mathcal{G}_j$-meßbar. Wegen $0 \leq \xi_i \leq 1$ für alle $1 \leq i \leq j$ und $j \leq t < \infty$ ist $E(|T_j|) < \infty$ für alle $0 \leq j \leq t$. Wegen

$$\begin{aligned} E(T_j|\mathcal{G}_{j-1}) &= E(\xi_j|\mathcal{G}_{j-1}) - E(c'|\mathcal{G}_{j-1}) + E(T_{j-1}|\mathcal{G}_{j-1}) \\ &= E(X'_j|\mathcal{G}_{j-1}) - c' + T_{j-1} \end{aligned}$$

gilt

$$\begin{aligned} E(T_j|\mathcal{G}_{j-1}) &\leq T_{j-1} \quad (a.s.) \\ \Leftrightarrow E(X'_j|\mathcal{G}_{j-1}) &\leq c' \quad (a.s.) \end{aligned}$$

für $1 \leq j \leq t$. Da letztere Ungleichung lt. Voraussetzung erfüllt ist, ist $((T_0, \mathcal{G}_0), \ldots, (T_t, \mathcal{G}_t))$ eine Supermartingale. Satz 4.4 3. liefert uns nun für $0 \leq \alpha' \leq 1 - c'$

$$P\left(\sum_{i=1}^{t} X'_i \geq (c' + \alpha')t\right) = P(T_t \geq \alpha' t) = P(T_t \geq \alpha' t | \mathcal{G}) \leq e^{-\frac{\alpha'^2 t}{2(1-c')}} \quad (a.s.).$$

Wir nehmen nun an, daß die Voraussetzungen des Korollars erfüllt sind. Nun sei $X'_i = 1 - X_i$ für alle $1 \leq i \leq t$ und $\alpha' = c\alpha$. Dann ist $c' = 1 - c$ für fast jeden Punkt eine obere Schranke für die Erwartungswerte. Somit können wir die oben bewiesene Aussage anwenden und erhalten für $0 \leq \alpha \leq 1$

$$P\left(\sum_{i=1}^{t} X_i \leq (1 - \alpha)tc\right) \leq e^{-\frac{\alpha^2 tc}{2}} \quad (a.s.).$$

□

Das folgende Lemma liefert, in Abhängigkeit von γ und der Schranke für den Abstand bzgl. L_p für ein $p > 0$, eine Schranke für die Anzahl der Irrtümer (bzgl. γ) beim on-line Lernen. Das darauf folgende Lemma zeigt dann, daß sich ein on-line Lernalgorithmus leicht in einen bzgl. γ konservativen on-line Lernalgorithmus transformieren läßt, ohne daß die Anzahl der Irrtümer steigt. Da ein konservativer on-line Algorithmus seine Hypothese nur dann ändert, wenn er sich geirrt hat, haben wir somit auch eine Schranke für die Anzahl der verschiedenen Hypothesen, die der on-line Lernalgorithmus produziert[7]. Diese Schranke wird bei der Analyse unserer Transformationen benutzt.

Lemma 4.6 *Lernt A für $p > 0$ eine Klasse F im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p, so irrt sich A maximal $m_\gamma = \lfloor l_p/\gamma^p \rfloor$ mal. D. h. A lernt F im on-line Modell mit m_γ-beschränktem Abstand bzgl. $L_{0,\gamma}$.*

Beweis: Annahme: A irrt sich mehr als $m_\gamma = \lfloor l_p/\gamma^p \rfloor$ mal, also irrt sich, wegen der Ganzzahligkeit von m_γ, A mehr als l_p/γ^p mal. D. h., daß die Gesamtabweichung bzgl. L_p größer als $(l_p/\gamma^p)\gamma^p = l_p$ ist, was im Widerspruch zur Schranke l_p steht. □

Lemma 4.7 *Gibt es einen on-line Lernalgorithmus A, der sich maximal m mal irrt, so läßt sich daraus ein konservativer on-line Lernalgorithmus A^*, der sich ebenfalls maximal m mal irrt, konstruieren.*

[7] Die Anzahl der verschiedenen Hypothesen ergibt sich offensichtlich aus der Anzahl der Irrtümer +1 für die Starthypothese.

Beweis: Sei A ein on-line Lernalgorithmus, der sich maximal m mal irrt. Wir konstruieren einen konservativen on-line Lernalgorithmus A^* aus A, der sich auf jeder endlichen Folge $S = ((x_1, y_1), \ldots, (x_t, y_t))$ von Beispielen maximal m mal irrt.

A^* verarbeitet on-line eine Folge S, indem er A auf einer Unterfolge S' von S simuliert. Die aktuelle Hypothese von A^* nach Verarbeitung von S ist stets gleich der aktuellen Hypothese h von A nach Verarbeitung von S'. Anfangs gilt $S = S' = \emptyset$ (leere Folge) und $h = h_0$ (Starthypothese von A). Wenn S um (x, y) erweitert wird, geht A^* wie folgt vor:

Fall 1: A begeht auf (x, y) einen Irrtum.
Dann gehört (x, y) auch zu S'. A^* vollzieht den Hypothesenwechsel mit.

Fall 2: A begeht auf (x, y) keinen Irrtum.
Dann gehört (x, y) nicht zu S'. A^* behält h bei. □

Lemma 4.8 *Seien sowohl h als auch f Funktionen $X \to \mathbb{R}$. Es ist $e_{L_{0,\gamma}}(h) < \epsilon$ genau dann, wenn h ein (ϵ, γ)-gutes Modell für entsprechendes f ist.*

Beweis: Definiere

$$err_h(x) = \begin{cases} 1 & \text{falls } |f(x) - h(x)| \geq \gamma \\ 0 & \text{sonst} \end{cases} .$$

Dann gilt

$$e_{L_{0,\gamma}}(h) = \int_X err_h(x)\, dP = P(\{x \mid |f(x) - h(x)| \geq \gamma\}).$$

Daraus folgt sofort die Behauptung des Lemmas. □

5 Die Transformationen

Der folgende Algorithmus wird in der zweiten Phase unserer Transformationen vom online Modell in das (ϵ, γ)-gute Modell, sowie in die Modelle mit ϵ-beschränkter Differenz benutzt.

Algorithmus $MI(H, S, L)$

Parameter:
H ist eine endliche Menge von Hypothesen.
$S = ((x_1, y_1), \ldots, (x_t, y_t))$ ist eine Probe bestehend aus t Paaren aus $X \times \mathbb{R}$ für beliebiges

$t \in \mathbb{N}$.
L ist eine Funktion $Y^2 \to [0,1]$ für den Abstand.

Beschreibung:
$MI(H,S,L)$ liefert die Hypothese $h \in H$, die zu S den geringsten Gesamtabstand liefert. D. h. es wird das $h \in H$ zurückgegeben, für welches $\sum_{i=1}^{t} L(h(x_i), y_i)$ minimal ist. (Falls mehrere solche $h \in H$ existieren, kann jede beliebige Strategie zur Auswahl eines dieser entsprechenden $h \in H$ angewendet werden.) •

$MI(H,S,L)$ bezeichne die Hypothese, die MI auf den angegebenen Parametern ausgibt.

Beachte, daß $MI(H,S,L_{0,\gamma})$ die Hypothese $h \in H$ liefert, die auf S die wenigsten Irrtümer (bzgl. γ) begeht. D. h. es wird das $h \in H$ zurückgegeben, für welches die Kardinalität von $\{j \mid |y_j - h(x_j)| \geq \gamma, 1 \leq j \leq t\}$ minimal ist.

Lemma 5.1 *Für eine beliebige, aber feste Funktion $L : Y^2 \to [0,1]$ fixiere eine Zielfunktion f und sei P die Wahrscheinlichkeitsverteilung, bzgl. der der erwartete Abstand der Hypothese (also $e_L(h)$) definiert ist. Für $\epsilon > 0$ sei H eine endliche Menge von Hypothesen, wovon mindestens eine Hypothese einen erwarteten Abstand von höchstens $\epsilon/2$ hat. Sei k die Anzahl der Hypothesen in H. Wenn nun $x_1, \ldots, x_t$ unabhängige Zufallsinstanzen der Wahrscheinlichkeitsverteilung P sind, dann gilt mit Wahrscheinlichkeit von mindestens $1 - ke^{-\epsilon t/28}$*

$$e_L(MI(H, ((x_1, f(x_1)), \ldots, (x_t, f(x_t))), L)) < \epsilon.$$

Beweis: Für $\epsilon > 1$ ist die Aussage trivial. Somit reicht es, das Lemma für $0 < \epsilon \leq 1$ zu zeigen. Betrachte einen Test auf t Beispielen für alle Hypothesen $h \in H$. Für eine Konstante c akzeptiere der Test eine Hypothese h gdw. der Gesamtabstand auf den t Beispielen zwischen h und f höchstens $\epsilon t/c$ ist[8]. Wir wählen $c = (1+\sqrt{3})/2$, da so die beiden unten gemachten Abschätzungen gleich werden. Mit den Hoeffding Ungleichungen (multiplikative Form) ist die Wahrscheinlichkeit, daß eine Hypothese h

[8] D. h. $\sum_{i=1}^{t} L(h(x_i), f(x_i)) \leq \epsilon t/c$

mit $e_L(h) \leq \epsilon/2$ vom Test verworfen wird,

$$\begin{aligned}
P\left(\sum_{i=1}^{t} L(h(x_i), f(x_i)) > \frac{\epsilon t}{c}\right) &\leq P\left(\sum_{i=1}^{t} L(h(x_i), f(x_i)) \geq \frac{\epsilon t}{c}\right) \\
&= GE\left(e_L(h), t, \frac{\epsilon t}{c}\right) \\
&\leq GE\left(\frac{\epsilon}{2}, t, \frac{\epsilon t}{c}\right) \\
&= GE\left(\frac{\epsilon}{2}, t, \left(1 + \frac{2-c}{c}\right)\frac{\epsilon t}{2}\right) \\
&\leq e^{-\frac{(2-c)^2 \epsilon t}{6c^2}} \\
&\leq e^{-\frac{\epsilon t}{28}}.
\end{aligned}$$

Die Wahrscheinlichkeit, daß eine Hypothese h mit $e_L(h) \geq \epsilon$ vom Test akzeptiert wird, ist

$$\begin{aligned}
P\left(\sum_{i=1}^{t} L(h(x_i), f(x_i)) \leq \frac{\epsilon t}{c}\right) &= LE\left(e_L(h), t, \frac{\epsilon t}{c}\right) \\
&\leq LE\left(\epsilon, t, \frac{\epsilon t}{c}\right) \\
&= LE\left(\epsilon, t, \left(1 - \frac{c-1}{c}\right)\epsilon t\right) \\
&\leq e^{-\frac{(c-1)^2 \epsilon t}{2c^2}} \\
&\leq e^{-\frac{\epsilon t}{28}}.
\end{aligned}$$

Damit MI eine Hypothese h mit $e_L(h) \geq \epsilon$ liefert, muß jede Hypothese h mit $e_L(h) \leq \epsilon/2$ vom Test verworfen werden oder eine Hypothese h mit $e_L(h) \geq \epsilon$ muß vom Test akzeptiert werden. Die Wahrscheinlichkeit, daß eines der beiden Ereignisse eintritt, durch

$$ke^{-\frac{\epsilon t}{28}}$$

beschränkt. □

5.1 Transformationen vom on-line Modell in das (ϵ, γ)-gute Modell

Wir geben nun eine Konversion von konservativen on-line Lernalgorithmen, deren Anzahl von Irrtümern durch eine Konstante beschränkt ist, in off-line Lernalgorithmen, die mit einem (ϵ, γ)-guten Modell lernen, an. Um die Konversion anwenden zu können müssen wir eine Schranke für die Anzahl der Irrtümer kennen. Mit Lemma 4.6 erhalten wir eine Schranke für die Anzahl der Irrtümer aus einer Schranke bzgl. einer Abstandsfunktion L_p für ein $p > 0$. Wegen Lemma 4.7 läßt sich jeder on-line Lernalgorithmus mit einer Schranke für die Anzahl der Irrtümer in einen konservativen on-line Lernalgorithmus mit gleicher Schranke für die Anzahl der Irrtümer transformieren, so daß die Voraussetzung der Konservativität keine Einschränkung ist.

Algorithmus $MBPAC(A, EX, \gamma, t_1, t_2)$

Parameter:
A ist ein konservativer on-line Lernalgorithmus für Funktionen $X \to \mathbb{R}$.
EX ist ein Orakel für ein Zielkonzept $f : X \to \mathbb{R}$.
γ ist der Abstandsparameter.
t_1 ist die „Länge“ der ersten Phase.
t_2 ist die „Länge“ der zweiten Phase.

Beschreibung:
Der Transformationsalgorithmus arbeitet in zwei Phasen. In der ersten Phase simuliert er t_1 Runden lang den on-line Algorithmus A, indem er t_1 mal das Orakel EX aufruft und die erhaltenen Beispiele als Instanzen und tatsächliche Werte nutzt. Die Menge der dabei auftretenden Hypothesen $H = \{h_0, \ldots, h_{t_1}\}$ wird für die zweite Phase gespeichert. (Da A konservativ ist, treten viele Hypothesen mehrfach auf; H ist die Menge der verschiedenen Hypothesen.)
In der zweiten Phase ruft der Transformationsalgorithmus t_2 mal das Orakel EX auf und erhält somit eine Folge S von t_2 Beispielen. Dann gibt er die Hypothese $MI(H, S, L_{0,\gamma})$ aus. •

Satz 5.2 *Sei EX ein Orakel für eine Zielfunktion $f : X \to \mathbb{R}$ aus einer Funktionenklasse F. Die Anzahl der Irrtümer (bzgl. γ), die der konservative on-line Lernalgorithmus A beim Lernen der Zielfunktion f macht, sei durch m_γ beschränkt*[9]*. Falls nun*

[9] Also, der konservative Algorithmus A lernt F im on-line Modell mit m_γ-beschränktem Abstand bzgl. $L_{0,\gamma}$.

$0 < \epsilon < 1$, $0 < \delta < 1$, $0 < \gamma$, $m_\gamma \geq 1$,

$$t_1 = \left\lceil \max\left\{\frac{16}{\epsilon}\ln\frac{2}{\delta}, \frac{4(m_\gamma - 1)}{\epsilon}\right\}\right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon}\ln\frac{2(m_\gamma + 1)}{\delta}\right\rceil,$$

dann liefert $MBPAC(A, EX, \gamma, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese, die ein (ϵ, γ)-gutes Modell für f ist.
Die Anzahl der Beispiele, die $MBPAC(A, EX, \gamma, t_1, t_2)$ benutzt ist durch

$$t_1 + t_2 = \frac{1}{\epsilon}\left(44\ln\frac{2}{\delta} + 4m_\gamma + 28\ln(m_\gamma + 1) - 2\right) = O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + m_\gamma\right)\right)$$

beschränkt.

Wir benutzen die gleiche Beweisidee wie Littlestone [Lit89], der ein entsprechendes Resultat für den „Booleschen Fall“ ($f : X \to \{0, 1\}$) gezeigt hat. Sein Resultat folgt dann als Korollar aus diesem Satz.

Beweis: Zuerst berechnen wir die Schranke. Die Gesamtanzahl der benutzten Beispiele ist durch

$$\begin{aligned} t_1 + t_2 &\leq \frac{16}{\epsilon}\ln\frac{2}{\delta} + \frac{4(m_\gamma - 1)}{\epsilon} + \frac{28}{\epsilon}\ln\frac{2(m_\gamma + 1)}{\delta} + 2 \\ &\leq \frac{1}{\epsilon}\left(44\ln\frac{2}{\delta} + 4m_\gamma + 28\ln(m_\gamma + 1) - 2\right) \end{aligned}$$

beschränkt.

Das Ziel der ersten Phase ist, eine Menge von Hypothesen zu erzeugen, von denen mindestens eine einen Fehler kleiner als $\epsilon/2$ hat.

Wähle eine Zielfunktion $f \in F$ beliebig, aber fest. Die tatsächlichen Werte, die der Algorithmus A erhält, sind aus den Instanzen unter Berücksichtigung von f berechenbar. Somit sind die Vorhersagen des Algorithmus eine Funktion der Folge der Instanzen. Sei $\lambda_j(x_1, \ldots, x_j)$ die Vorhersage des Algorithmus A in Runde j, falls die Folge der Instanzen in den ersten j Runden $x_1, \ldots, x_j$ ist. Sei h_0 die Starthypothese von A und $h_j(x_1, \ldots, x_j)$ die Hypothese, die von A am Ende der j-ten Runde erzeugt wird. Somit ist der Wert von $h_j(x_1,\ldots,x_j)$ für ein bestimmtes $(x_1,\ldots,x_j)$ die Funktion $h : X \to Y$ mit $h(x) = \lambda_{j+1}(x_1,\ldots,x_j,x)$. Wir haben $Error(h_0) = P(\{x \mid |\lambda_1(x) - f(x)| \geq \gamma\})$ und für $j \geq 1$

$$\begin{aligned} Error(h_j(x_1,\ldots,x_j)) &= P(\{x \mid |\lambda_{j+1}(x_1,\ldots,x_j,x) - f(x)| \geq \gamma\}) \\ &= \int_X L_{0,\gamma}(\lambda_{j+1}(x_1,\ldots,x_j,x_{j+1}), f(x_{j+1}))\, dP(x_{j+1}). \end{aligned}$$

Für $j \geq 1$ sei $B_j = \{(x_1,\ldots,x_j) \mid Error(h_j(x_1,\ldots,x_j)) \geq \epsilon/2\}$ und $\tilde{r}_j : X^j \to \{0,1\}$ gegeben durch

$$\tilde{r}_1(x_1) = \begin{cases} L_{0,\gamma}(\lambda_1(x_1), f(x_1)) & \text{falls } Error(h_0) \geq \frac{\epsilon}{2} \\ 1 & \text{sonst} \end{cases}$$

und für $j > 1$ durch

$$\tilde{r}_j(x_1,\ldots,x_j) = \begin{cases} L_{0,\gamma}(\lambda_j(x_1,\ldots,x_j), f(x_j)) & \text{falls } (x_1,\ldots,x_{j-1}) \in B_{j-1} \\ 1 & \text{sonst} \end{cases}.$$

Grob gesprochen ist $\tilde{r}_j$ eine Funktion, die uns angibt, ob die Runde j günstig (bzgl. dem Erreichen des Ziels dieser Phase) verläuft. Ist $\tilde{r}_j = 1$, so gilt $Error(h) < \epsilon/2$ für die zur Vorhersage benutzte Hypothese h, was das Ziel dieser Phase erfüllt, oder die (schlechte) Hypothese, die zur Vorhersage benutzt wird, irrt sich auf der gegebenen Instanz (dadurch verkleinert sich der „Irrtumsvorrat“ des on-line Algorithmus um 1). Dagegen ist $\tilde{r}_j = 0$, wenn eine schlechte ($Error(h) \geq \epsilon/2$) Hypothese h sich auf der gegebenen Instanz nicht irrt.

Um die Wahrscheinlichkeit, daß die Summe der $\tilde{r}_j$ in den t_1 Runden klein ist, mit Korollar 4.5 nach oben abzuschätzen, müssen wir $E(\tilde{r}_1)$ und $E(\tilde{r}_j|\tilde{r}_1,\ldots,\tilde{r}_{j-1})$ für $j \geq 2$ nach unten abschätzen.

Falls $Error(h_0) < \epsilon/2$, dann ist $E(\tilde{r}_1) = 1$. Ansonsten ist $E(\tilde{r}_1) = E_{x\in X}(L_{0,\gamma}(\lambda_1(x), f(x))) = Error(h_0) \geq \epsilon/2$. Für $j \geq 2$ bezeichne $\sigma(\tilde{r}_1,\ldots,\tilde{r}_{j-1})$ die durch $\tilde{r}_1,\ldots,\tilde{r}_{j-1}$ erzeugte σ-Algebra. Da $\tilde{r}_1,\ldots,\tilde{r}_{j-1}$ nur von $x_1,\ldots,x_{j-1}$ und nicht von x_j (für $(x_1,\ldots,x_j) \in X^j$) abhängen, kann jede Menge aus $\sigma(\tilde{r}_1,\ldots,\tilde{r}_{j-1})$ in der Form $U \times X$ für ein $U \subseteq X^{j-1}$ dargestellt werden. Für jede solche Menge U gilt:

$$\int_{U\times X} E(\tilde{r}_j|\tilde{r}_1,\ldots,\tilde{r}_{j-1})\, dP^j = \int_{U\times X} \tilde{r}_j\, dP^j$$

$$= \int_{(U\cap B_{j-1})\times X} L_{0,\gamma}(\lambda_j(x_1,\ldots,x_j), f(x_j))\, dP^j + \int_{(U\setminus B_{j-1})\times X} 1\, dP^j$$

$$
\begin{aligned}
&= \int_{U \cap B_{j-1}} Error(h_{j-1})\, dP^{j-1} + \int_{U \setminus B_{j-1}} 1\, dP^{j-1} \\
&\geq \int_{U \cap B_{j-1}} \frac{\epsilon}{2}\, dP^{j-1} + \int_{U \setminus B_{j-1}} \frac{\epsilon}{2}\, dP^{j-1} \\
&= \int_{U} \frac{\epsilon}{2}\, dP^{j-1} = \int_{U \times X} \frac{\epsilon}{2}\, dP^{j}.
\end{aligned}
$$

Da dieses für jede Menge $U \times X \in \sigma(\tilde{r}_1, \ldots, \tilde{r}_{j-1})$ gilt und da $E(\tilde{r}_j | \tilde{r}_1, \ldots, \tilde{r}_{j-1})$ lt. Definition $\sigma(\tilde{r}_1, \ldots, \tilde{r}_{j-1})$-meßbar ist, gilt $E(\tilde{r}_j | \tilde{r}_1, \ldots, \tilde{r}_{j-1}) \geq \epsilon/2$ $(a.s.)$.

Wenden wir nun Korollar 4.5 mit $c = \epsilon/2$ und $\alpha = 1/2$ an, so erhalten wir dann $P(\sum_{i=1}^{t_1} \tilde{r}_i \leq \epsilon t_1/4) \leq e^{-\epsilon t_1/16}$ für fast jeden Punkt[10] aus X^{t_1}. Wegen $t_1 \geq (16/\epsilon)\ln(2/\delta)$ ist $e^{-\epsilon t_1/16} \leq \delta/2$.

Wenn $\tilde{r}_j = 1$, dann irrt sich der Algorithmus A in Runde j oder es gilt $Error(h_{j-1}) < \epsilon/2$. Gilt also $\sum_{j=1}^{t_1} \tilde{r}_j \geq m$, dann hat sich A mindestens m mal geirrt, oder es gilt $Error(h) < \epsilon/2$ für zumindest eine der Hypothesen $h \in \{h_0, \ldots, h_{t_1-1}\}$. Aus ersterem folgt, da m eine Schranke für die Anzahl der Irrtümer von A ist, daß sich A und damit die Hypothese h_{t_1} nicht mehr irren kann. Somit gilt dann $Error(h) < \epsilon/2$ für zumindest eine der Hypothesen $h \in H = \{h_0, \ldots, h_{t_1}\}$. Mit der Wahl $t_1 \geq 4(m-1)/\epsilon$ erhalten wir mit $\sum_{j=1}^{t_1} \tilde{r}_j > \epsilon t_1/4$ dann $\sum_{j=1}^{t_1} \tilde{r}_j > m - 1$. Wegen der Ganzzahligkeit von m gilt dann $\sum_{j=1}^{t_1} \tilde{r}_j \geq m$. Somit gilt mit Wahrscheinlichkeit von mindestens $1 - \delta/2$ zumindest für eine der Hypothesen $h \in H$ dann $Error(h) < \epsilon/2$.

Sei $H = \{h_0, \ldots, h_t\}$. Da A konservativ ist und sich höchstens m_γ mal irrt, ist die Anzahl der verschiedenen Hypothesen in $h_0, \ldots, h_t$ durch $m_\gamma + 1$ beschränkt, also $|H| \leq m_\gamma + 1$.

Mit Lemma 5.1 erhalten wir nun, daß die Wahrscheinlichkeit dafür, daß in der zweiten Phase eine Hypothese mit einem Fehler kleiner als ϵ ausgegeben wird, mindestens

$$
1 - |H| e^{-\frac{\epsilon}{28}\left\lceil \frac{28}{\epsilon} \ln \frac{2(m_\gamma+1)}{\delta} \right\rceil} \geq 1 - \frac{\delta}{2}
$$

ist.

Da die zweite Phase unabhängig von der ersten ist, ist die Wahrscheinlichkeit, daß die Transformation eine Hypothese mit einem Fehler kleiner als ϵ liefert, mindestens $1 - \delta$. Also ist mit Wahrscheinlichkeit von mindestens $1 - \delta$ die gelieferte Hypothese ein (ϵ, γ)-gutes Modell der Zielfunktion f. □

[10] im Sinne von $(a.s.)$

Ist bei einem on-line Algorithmus die Gesamtanzahl der Irrtümer m nicht durch eine Konstante, sondern nur durch eine Funktion, abhängig von der Anzahl der Runden, beschränkt, so können wir obigen Satz nicht direkt anwenden. Wir können jedoch aus obigem Satz obere Schranken für die Anzahl der Beispiele, die die Transformation benötigt, ableiten, falls die Gesamtanzahl der Irrtümer beim on-line Lernen durch $m(t) = O(\ln t)$ oder $m(t) = O(t^{1/q})$ für $q > 1$ beschränkt ist. Dafür zeigen wir zuerst folgendes Lemma.

Lemma 5.3 *1. Sei $k > 0$ und $s \geq 2k\ln(2k)$, dann gilt $s \geq k \ln s$.*

2. Sei $k > 0$, $s \geq 12$ und $s \geq 2k\ln(13k/6)$, dann gilt $s \geq k\ln(s+1)$.

3. Sei $k > 0$, $q > 1$ und $s \geq k^{q/(q-1)}$, dann gilt $s \geq ks^{1/q}$.

4. Sei $k > 0$, $q > 1$, $s \geq 12$ und $s \geq (13/12)^{1/(q-1)}k^{q/(q-1)}$, dann gilt $s \geq k(s+1)^{1/q}$.

Beweis:

1.

$$\begin{array}{rcl} s & \geq & 2k\ln(2k) \wedge k > 0 \\ \Rightarrow s & \geq & 2k\ln(2k) - 2k \\ \Leftrightarrow \dfrac{s}{2} & \geq & k\ln(2k) - k \\ \Leftrightarrow s & \geq & k\ln(2k) + \dfrac{s}{2} - k \end{array}$$

Aus $x - 1 \geq \ln x$ folgt mit $x = s/(2k)$

$$k\ln(2k) + \frac{s}{2} - k \geq k \ln s$$

und somit $s \geq k\ln s$.

2.

$$\begin{array}{rcl} s & \geq & 2k\ln(13k/6) \wedge k > 0 \\ \Rightarrow s & \geq & 2k\ln(13k/6) - 2k \\ \Leftrightarrow \dfrac{s}{2} & \geq & k\ln(13k/6) - k \\ \Leftrightarrow s & \geq & k\ln(13k/6) + \dfrac{s}{2} - k \\ \Leftrightarrow s & \geq & k\ln(2k) + \dfrac{s}{2} - k + k\ln(13/12) \end{array}$$

Aus $x-1 \geq \ln x$ folgt mit $x = s/(2k)$

$$k\ln(2k) + \frac{s}{2} - k \geq k \ln s$$

und somit $s \geq k\ln s + k\ln(13/12) = k\ln(13s/12)$. Wegen $s \geq 12 \Leftrightarrow s+1 \leq s + s/12 = 13s/12$ erhalten wir $s \geq k\ln(s+1)$

3. Für $k > 0$ und $q > 1$ gilt

$$s \geq k^{q/(q-1)} \Leftrightarrow s^{1-(1/q)} \geq k \Leftrightarrow s \geq ks^{1/q}.$$

4. Es gilt

$$s \geq \left(\frac{13}{12}\right)^{1/(q-1)} k^{q/(q-1)} \Leftrightarrow s^{q-1} \geq \frac{13}{12}k^q.$$

Wir erhalten nun

$$s\left(s^{q-1} - k^q\right) \geq s\left(\frac{13}{12}k^q - k^q\right) = \frac{s}{12}k^q \geq k^q.$$

Durch Umformung dieser Ungleichung erhalten wir schließlich die gewünschte Aussage:

$$s\left(s^{q-1} - k^q\right) \geq k^q \Leftrightarrow s^q \geq (s+1)k^q \Leftrightarrow s \geq k(s+1)^{1/q}.$$

□

Korollar 5.4 *Sei EX ein Orakel für eine Zielfunktion $f : X \to \mathbb{R}$ aus einer Funktionenklasse F. Die Anzahl der Irrtümer (bzgl. γ), die der konservative on-line Lernalgorithmus A auf Folgen mit Maximallänge t beim Lernen der Zielfunktion f macht, sei durch $m_\gamma(t)$ beschränkt. Falls nun $0 < \epsilon < 1$, $0 < \delta < 1$, $0 < \gamma$, $m_\gamma(t) \geq 1$ für alle $t \in \mathbb{N}$, $m_\gamma(t) = c_\gamma \ln t$ für alle $t \in \mathbb{N}$ mit $t \geq 12$ und für eine Konstante $c_\gamma > 0$,*

$$t_1 = \left\lceil \max\left\{\frac{16}{\epsilon}\ln\frac{2}{\delta}, \frac{8c_\gamma}{\epsilon}\ln\frac{26c_\gamma}{3\epsilon}\right\}\right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon}\ln\frac{2(c_\gamma \ln t_1 + 1)}{\delta}\right\rceil,$$

dann liefert $MBPAC(A, EX, \gamma, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1-\delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, die ein (ϵ,γ)-gutes

Modell für f ist.
Die Anzahl der Beispiele, die $MBPAC(A, EX, \gamma, t_1, t_2)$ benutzt ist für $c_\gamma \geq \frac{3\epsilon}{26}$ durch

$$t_1 + t_2 = O\left(\frac{1}{\epsilon}\left(c_\gamma \ln \frac{c_\gamma}{\epsilon} + \ln \frac{1}{\delta}\right)\right)$$

beschränkt.

Beweis: Da $t_1 \geq \lceil (16/\epsilon)\ln(2/\delta)\rceil \geq 12$ gilt, wird A in der ersten Phase mindestens 12 Runden simuliert, so daß $m_\gamma(t)$ für $t < 12$ hier unbedeutend ist[11].

Setzt man nun $t = t_1 + 1$, so kann der Beweis von Phase 1 aus dem Beweis von Satz 5.2 übernommen werden[12]. Für Phase 1 verbleibt zu zeigen, daß

$$t_1 \geq \left\lceil \frac{16}{\epsilon} \ln \frac{2}{\delta} \right\rceil$$

und

$$t_1 \geq \left\lceil \frac{4(m_\gamma(t_1+1) - 1)}{\epsilon} \right\rceil$$

gilt. Ersteres ist offensichtlich erfüllt und die andere Ungleichung läßt sich wie folgt zeigen.

Es gilt $t_1 \geq \lceil (16/\epsilon)\ln(2/\delta)\rceil \geq 12$. Wegen $t_1 \geq \frac{8c_\gamma}{\epsilon} \ln \frac{26c_\gamma}{3\epsilon}$ läßt sich dann Lemma 5.3 2. mit $k = 4c_\gamma/\epsilon$ und $s = t_1$ anwenden. Wir erhalten

$$t_1 \geq \frac{4c_\gamma}{\epsilon} \ln(t_1 + 1) = \frac{4m_\gamma(t_1+1)}{\epsilon} \geq \frac{4(m_\gamma(t_1+1) - 1)}{\epsilon}.$$

Da $t_1 \in \mathbb{N}$ gilt nun

$$t_1 \geq \left\lceil \frac{4(m_\gamma(t_1+1) - 1)}{\epsilon} \right\rceil.$$

[11]Es ist notwendig eine Schranke größer als 1 anzugeben, unterhalb der $m_\gamma(t) = c_\gamma \ln t$ nicht gelten muß, da ansonsten $m_\gamma(1) = 0$ gilt, weshalb dann die Starthypothese des on-line Algorithmus die „zu lernende" Zielfunktion sein muß. Außerdem wäre dann $m_\gamma(1) = 0 \not\geq 1$.

[12]Da die Anzahl der Irrtümer des on-line Lernalgorithmus mit der Anzahl der Runden wächst, muß die Hypothese nach t_1 Runden, bei Ausschöpfung der Anzahl der Irrtümer in den ersten t_1 Runden, nicht die Zielfunktion sein, so daß in der $(t_1 + 1)$-ten Runde ein Irrtum möglich ist. Deshalb wird $m_\gamma(t_1 + 1)$ benutzt, so daß, falls diese Schranke bereits nach t_1 Runden erreicht wurde, in der $(t_1 + 1)$-ten Runde kein Irrtum mehr möglich ist.

Auch für die zweite Phase kann der Beweis von Satz 5.2 benutzt werden. Die Anzahl der verschiedenen Hypothesen ist hier durch $m_\gamma(t_1) + 1$ beschränkt, so daß

$$t_2 = \left\lceil \frac{28}{\epsilon} \ln \frac{2(m_\gamma(t_1) + 1)}{\delta} \right\rceil = \left\lceil \frac{28}{\epsilon} \ln \frac{2(c_\gamma \ln t_1 + 1)}{\delta} \right\rceil$$

für $t_1 \geq 12$ ausreicht.

Somit verbleibt nur noch zu zeigen, daß

$$t_1 + t_2 = O\left(\frac{1}{\epsilon}\left(c_\gamma \ln \frac{c_\gamma}{\epsilon} + \ln \frac{1}{\delta}\right)\right).$$

Mit $c_\gamma \geq \frac{3\epsilon}{26}$ erfüllt t_1 offensichtlich die Schranke und für t_2 gilt

$$\begin{aligned}
t_2 &= O\left(\frac{1}{\epsilon}\left(\ln \frac{1}{\delta} + \ln(c_\gamma \ln t_1)\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln \frac{1}{\delta} + \ln c_\gamma + \ln\ln\left(\frac{1}{\epsilon}\left(c_\gamma \ln \frac{c_\gamma}{\epsilon} + \ln \frac{1}{\delta}\right)\right)\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\ln \frac{1}{\epsilon} + \ln c_\gamma + \ln \frac{1}{\delta}\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln \frac{c_\gamma}{\epsilon} + \ln \frac{1}{\delta}\right)\right).
\end{aligned}$$

□

Korollar 5.5 *Sei EX ein Orakel für eine Zielfunktion $f : X \to \mathbb{R}$ aus einer Funktionenklasse F. Die Anzahl der Irrtümer (bzgl. γ), die der konservative on-line Lernalgorithmus A auf Folgen mit Maximallänge t beim Lernen der Zielfunktion f macht, sei durch $m_\gamma(t)$ beschränkt. Falls nun $0 < \epsilon < 1$, $0 < \delta < 1$, $0 < \gamma$, $m_\gamma(t) \geq 1$, $m_\gamma(t) = c_\gamma t^{1/q}$ für alle $t \in \mathbb{N}$ und für Konstanten $c_\gamma > 0$ und $q > 1$,*

$$t_1 = \left\lceil \max\left\{\frac{16}{\epsilon} \ln \frac{2}{\delta}, \left(\frac{13}{12}\right)^{\frac{1}{q-1}} \left(\frac{4c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}}\right\} \right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon} \ln \frac{2\left(c_\gamma t_1^{1/q} + 1\right)}{\delta} \right\rceil,$$

dann liefert $MBPAC(A, EX, \gamma, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, die ein (ϵ, γ)-gutes

Modell für f ist.
Die Anzahl der Beispiele, die $MBPAC(A, EX, \gamma, t_1, t_2)$ benutzt ist durch

$$t_1 + t_2 = O\left(\left(\frac{c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

beschränkt. Betrachten wir Folgen von Funktionenklassen, so dürfen wir q in der O-Notation nicht als Konstante ansehen. Somit ergibt sich dann

$$t_1 + t_2 = O\left(\left(\frac{13}{3}\right)^{\frac{1}{q-1}}\left(\frac{c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{q}{q-1}\ln\frac{c_\gamma}{\epsilon}\right)\right).$$

Beweis: Analog zu Korollar 5.4 reicht es zu zeigen, daß

$$t_1 \geq \left\lceil\frac{4\left(c_\gamma(t_1+1)^{1/q} - 1\right)}{\epsilon}\right\rceil$$

und

$$t_2 = O\left(\left(\frac{13}{3}\right)^{\frac{1}{q-1}}\left(\frac{c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{q}{q-1}\ln\frac{c_\gamma}{\epsilon}\right)\right),$$

da der Rest offensichtlich ist.

Wegen $t_1 \geq \left(\frac{13}{12}\right)^{\frac{1}{q-1}}\left(\frac{4c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}}$ läßt sich Lemma 5.3 4. mit $k = 4c_\gamma/\epsilon$ und $s = t_1$ anwenden. Wir erhalten

$$t_1 \geq \frac{4c_\gamma}{\epsilon}(t_1+1)^{1/q} \geq \frac{4\left(c_\gamma(t_1+1)^{1/q} - 1\right)}{\epsilon}$$

und da $t_1 \in \mathbb{N}$

$$t_1 \geq \left\lceil\frac{4\left(c_\gamma(t_1+1)^{1/q} - 1\right)}{\epsilon}\right\rceil.$$

Es verbleibt also nur noch t_2 abzuschätzen:

$$\begin{aligned}
t_2 &= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln\left(c_\gamma t_1^{1/q}\right)\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\frac{1}{q}\ln t_1\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\ln t_1\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\ln\left(\frac{1}{\epsilon}\ln\frac{1}{\delta}+\left(\frac{13}{12}\right)^{\frac{1}{q-1}}\left(\frac{4c_\gamma}{\epsilon}\right)^{\frac{q}{q-1}}\right)\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\ln\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\left(\frac{13}{3\epsilon}\right)^{\frac{1}{q-1}}c_\gamma^{\frac{q}{q-1}}\right)\right)\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\ln\frac{1}{\epsilon}+\ln\left(\ln\frac{1}{\delta}+\left(\frac{13}{3\epsilon}\right)^{\frac{1}{q-1}}c_\gamma^{\frac{q}{q-1}}\right)\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\ln c_\gamma+\ln\frac{1}{\epsilon}+\ln\ln\frac{1}{\delta}+\ln\left(\frac{13}{3\epsilon}\right)^{\frac{1}{q-1}}+\ln c_\gamma^{\frac{q}{q-1}}\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\frac{q}{q-1}\ln\frac{1}{\epsilon}+\frac{q}{q-1}\ln c_\gamma\right)\right)\\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\frac{q}{q-1}\ln\frac{c_\gamma}{\epsilon}\right)\right).
\end{aligned}$$

□

Bei dem vorherigen Korollar reicht es aus, wenn $m_\gamma(t)=c_\gamma t^{1/q}$ für $t\geq 12$ gilt (s. Beweis von Korollar 5.4), wodurch evtl. eine kleinere Konstante c_γ möglich ist.

Mit den Lemmata 4.6 und 4.7 läßt sich der Satz und die Korollare für on-line Algorithmen mit beschränktem Abstand bzgl. L_p für ein $p>0$ umschreiben.

Korollar 5.6 *Sei EX ein Orakel für eine Zielfunktion $f: X\to\mathbb{R}$ aus einer Funktionenklasse F. Sei weiterhin $0<\epsilon<1$, $0<\delta<1$ und $0<\gamma$.*

1. *A lerne F im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p für ein $p>0$. Sei $l_p/\gamma^p\geq 1$, dann läßt sich mit*

$$\frac{1}{\epsilon}\left(44\ln\frac{2}{\delta}+4\left\lfloor\frac{l_p}{\gamma^p}\right\rfloor+28\ln\left(\left\lfloor\frac{l_p}{\gamma^p}\right\rfloor+1\right)-2\right)=O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta}+\frac{l_p}{\gamma^p}\right)\right)$$

Beispielen mit Wahrscheinlichkeit von mindestens $1-\delta$ eine Hypothese h berechnen, die ein (ϵ,γ)-gutes Modell für f ist.

2. *A lerne F im on-line Modell mit $l_p(t)$-beschränktem Abstand bzgl. L_p für ein $p>0$. Sei $l_p(t)/\gamma^p \geq 1$ für alle $t \in \mathbb{N}$, $l_p(t) = c_p \ln t$ für alle $t \in \mathbb{N}$ mit $t \geq 12$ und für $c_p/\gamma^p \geq \frac{3\epsilon}{26}$, dann läßt sich mit*

$$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{c_p}{\gamma^p}\ln\frac{c_p}{\epsilon\gamma^p}\right)\right)$$

Beispielen mit Wahrscheinlichkeit von mindestens $1-\delta$ eine Hypothese h berechnen, die ein (ϵ,γ)-gutes Modell für f ist.

3. *A lerne F im on-line Modell mit $l_p(t)$-beschränktem Abstand bzgl. L_p für ein $p>0$. Sei $l_p(t)/\gamma^p \geq 1$, $l_p(t) = c_p t^{1/q}$ für alle $t \in \mathbb{N}$ und für eine Konstante $q>1$, $c_p/\gamma^p > 0$, dann läßt sich mit*

$$O\left(\frac{1}{\epsilon}\ln\frac{1}{\delta} + \left(\frac{c_p}{\epsilon\gamma^p}\right)^{\frac{q}{q-1}}\right)$$

Beispielen mit Wahrscheinlichkeit von mindestens $1-\delta$ eine Hypothese h berechnen, die ein (ϵ,γ)-gutes Modell für f ist.

Das Korollar ist mit den obigen Bemerkungen offensichtlich, wenn wir für 2. und 3. $c_\gamma = c_p/\gamma^p$ setzen.

5.2 Transformationen vom on-line Modell in die Modelle mit ϵ-beschränkter Differenz

Wir geben nun eine Konversion von on-line Lernalgorithmen, die mit beschränktem Abstand bzgl. einer Abstandsfunktion lernen, in off-line Lernalgorithmen, die mit ϵ-beschränkter Differenz bzgl. der gleichen Abstandsfunktion lernen, an. Statt der Konservativität benutzten wir hier nun (optional[13]) eine Schranke für die maximale Anzahl verschiedener Hypothesen des benutzten on-line Lernalgorithmus auf einer Folge von Beispielen.

[13]Falls keine Schranke bekannt ist, setzt man sie einfach auf ∞.

Algorithmus $BLPAC(A, EX, L, t_1, t_2)$

Parameter:
A ist ein on-line Lernalgorithmus für Funktionen $X \to Y$.
EX ist ein Orakel für ein Zielkonzept $f : X \to Y$.
$L : Y^2 \to [0, 1]$ ist die Abstandsfunktion.
t_1 ist die „Länge“ der ersten Phase.
t_2 ist die „Länge“ der zweiten Phase.

Beschreibung:
Der Transformationsalgorithmus arbeitet in zwei Phasen. In der ersten Phase simuliert er t_1 Runden lang den on-line Algorithmus A, indem er t_1 mal das Orakel EX aufruft und die erhaltenen Beispiele als Instanzen und tatsächliche Werte nutzt. Die Menge der dabei auftretenden Hypothesen (ohne die letzte Hypothese; s. u.) $H = \{h_0, \ldots, h_{t_1-1}\}$ wird für die zweite Phase gespeichert.
In der zweiten Phase ruft der Transformationsalgorithmus t_2 mal das Orakel EX auf und erhält somit eine Folge S von t_2 Beispielen. Dann gibt er die Hypothese $MI(H, S, L)$ aus. •

Die letzte Hypothese, die $BLPAC$ in der 1. Phase berechnet, wird nicht benutzt. Es reicht also aus, A nur $t_1 - 1$ Runden lang zu simulieren, jedoch vereinfacht diese Betrachtung die Argumentation im Beweis zur 1. Phase im folgenden Satz und es liefert eine vereinfachte Rechnung bei den Beweisen zu den Korollaren 5.9 und 5.11 unten.

Satz 5.7 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0, 1]$ eine Funktion für den Abstand. A lerne F im on-line Modell mit l-beschränktem Abstand bzgl. L. Sei nun g (evtl. ∞) eine obere Schranke für die Anzahl verschiedener Hypothesen, die A auf einer beliebigen endlichen Folge von Beispielen produziert[14]. Falls nun $\epsilon > 0$, $0 < \delta < 1$, $l > 0$,*

$$t_1 = \left\lceil \max\left\{\frac{16}{\epsilon}\ln\frac{2}{\delta}, \frac{4l}{\epsilon}\right\}\right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon}\ln\frac{2\min\{g, t_1\}}{\delta}\right\rceil,$$

[14] g muß keine Konstante sein, sondern darf von Parametern abhängig sein. Sei z. B. A konservativ (bzgl. einer Konstante c) und $L = L_1$, so irrt sich A wegen Lemma 4.6 maximal $\lfloor l/c \rfloor$ mal (bzgl. c). Somit ist die Anzahl der verschiedenen Hypothesen durch $\lfloor l/c \rfloor + 1$ beschränkt.

dann liefert $BLPAC(A, EX, L, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L, t_1, t_2)$ benutzt ist durch

$$\begin{aligned} t_1 + t_2 &= \frac{1}{\epsilon}\left(44 \ln \frac{2}{\delta} + 4l + 28 \ln \min\left\{g, \frac{16}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + 1\right\}\right) + 2 \\ &= O\left(\frac{1}{\epsilon}\left(l + \ln \min\left\{\frac{g}{\delta}, \frac{1}{\epsilon\delta}\right\}\right)\right) \end{aligned}$$

beschränkt.

Auch hier benutzen wir, wie schon bei Satz 5.2, die Beweisidee von Littlestone [Lit89].

Beweis: Zuerst berechnen wir die Schranke. Es gilt

$$t_2 \leq \left\lceil \frac{28}{\epsilon} \ln \frac{2}{\delta} + \frac{28}{\epsilon} \ln \min\left\{g, \frac{16}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + 1\right\}\right\rceil.$$

Somit ist die Gesamtanzahl der benutzten Beispiele ist durch

$$\begin{aligned} t_1 + t_2 &\leq \frac{16}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + \frac{28}{\epsilon} \ln \frac{2}{\delta} + \frac{28}{\epsilon} \ln \min\left\{g, \frac{16}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + 1\right\} + 2 \\ &= \frac{1}{\epsilon}\left(44 \ln \frac{2}{\delta} + 4l + 28 \ln \min\left\{g, \frac{16}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + 1\right\}\right) + 2 \end{aligned}$$

beschränkt.

Da der Fall $\epsilon > 1$ offensichtlich trivial ist, reicht es, den Satz für $0 < \epsilon \leq 1$ zu zeigen.

Das Ziel der ersten Phase ist, eine Menge von Hypothesen zu erzeugen, so daß für zumindest eine von ihnen $e_L(h) < \epsilon/2$ gilt.

Wähle nun eine Zielfunktion $f \in F$ beliebig, aber fest. Die tatsächlichen Werte, die der Algorithmus A erhält, sind aus den Instanzen unter Berücksichtigung von f berechenbar. Somit sind die Vorhersagen des Algorithmus eine Funktion der Folge der Instanzen. Sei $\lambda_j(x_1, \ldots, x_j)$ die Vorhersage des Algorithmus A in Runde j, falls die Folge der Instanzen in den ersten j Runden $x_1, \ldots, x_j$ ist. Sei h_0 die Starthypothese von A und $h_j(x_1, \ldots, x_j)$ die Hypothese, die von A am Ende der j-ten Runde erzeugt wird. Somit ist der Wert von $h_j(x_1, \ldots, x_j)$ für ein bestimmtes $(x_1, \ldots, x_j)$ die Funktion

$h : X \to Y$ mit $h(x) = \lambda_{j+1}(x_1, \ldots, x_j, x)$. Wir haben $e_L(h_0) = E_{x \in X}(L(\lambda_1(x), f(x)))$ und für $j \geq 1$

$$\begin{aligned} e_L(h_j(x_1, \ldots, x_j)) &= E_{x\in X}(L(\lambda_{j+1}(x_1, \ldots, x_j, x), f(x))) \\ &= \int_X L(\lambda_{j+1}(x_1, \ldots, x_j, x_{j+1}), f(x_{j+1}))\, dP(x_{j+1}). \end{aligned}$$

Für $j \geq 1$ sei $B_j = \{(x_1, \ldots, x_j) \mid e_L(h_j(x_1, \ldots, x_j)) \geq \epsilon/2\}$ und $\tilde{r}_j : X^j \to [0, 1]$ gegeben durch

$$\tilde{r}_1(x_1) = \begin{cases} L(\lambda_1(x_1), f(x_1)) & \text{falls } e_L(h_0) \geq \frac{\epsilon}{2} \\ 1 & \text{sonst} \end{cases}$$

und für $j > 1$ durch

$$\tilde{r}_j(x_1, \ldots, x_j) = \begin{cases} L(\lambda_j(x_1, \ldots, x_j), f(x_j)) & \text{falls } (x_1, \ldots, x_{j-1}) \in B_{j-1} \\ 1 & \text{sonst} \end{cases}.$$

Grob gesprochen ist $\tilde{r}_j$ eine Funktion, die uns angibt, wie günstig (bzgl. dem Erreichen des Ziels dieser Phase) die Runde j verläuft. Je größer der Wert ist, den $\tilde{r}_j$ liefert, desto günstiger verläuft die Runde j. Dabei steht 1 für sehr günstig und 0 für sehr ungünstig. Z. B. ist $\tilde{r}_j = 1$, wenn $e_L(h) < \epsilon/2$ für die zur Vorhersage benutzte Hypothese h gilt, was das Ziel dieser Phase erfüllt, oder wenn die (schlechte) Hypothese, die zur Vorhersage benutzt wird, auf der gegebenen Instanz den größten möglichen Abstand (= 1) bzgl. L hat (dadurch verkleinert sich der „Abstandsvorrat“ des on-line Algorithmus stark). Dagegen ist $\tilde{r}_j$ klein, wenn eine schlechte ($e_L(h) \geq \epsilon/2$) Hypothese h auf der gegebenen Instanz nur einen kleinen ($= \tilde{r}_j$) Abstand bzgl. L hat.

Um die Wahrscheinlichkeit, daß die Summe der $\tilde{r}_j$ in den t_1 Runden klein ist, mit Korollar 4.5 nach oben abzuschätzen, müssen wir $E(\tilde{r}_1)$ und $E(\tilde{r}_j|\tilde{r}_1, \ldots, \tilde{r}_{j-1})$ für $j \geq 2$ nach unten abschätzen.

Falls $e_L(h_0) < \epsilon/2$, dann ist $E(\tilde{r}_1) = 1$. Ansonsten ist $E(\tilde{r}_1) = E_{x\in X}(L(\lambda_1(x), f(x))) = e_L(h_0) \geq \epsilon/2$. Für $j \geq 2$ bezeichne $\sigma(\tilde{r}_1, \ldots, \tilde{r}_{j-1})$ die durch $\tilde{r}_1, \ldots, \tilde{r}_{j-1}$ erzeugte σ-Algebra. Da $\tilde{r}_1, \ldots, \tilde{r}_{j-1}$ nur von $x_1, \ldots, x_{j-1}$ und nicht von x_j (für $(x_1, \ldots, x_j) \in X^j$) abhängen, kann jede Menge aus $\sigma(\tilde{r}_1, \ldots, \tilde{r}_{j-1})$ in der Form $U \times X$ für ein $U \subseteq X^{j-1}$ dargestellt werden. Für jede solche Menge U gilt:

$$\int_{U\times X} E(\tilde{r}_j|\tilde{r}_1, \ldots, \tilde{r}_{j-1})\, dP^j = \int_{U\times X} \tilde{r}_j\, dP^j$$

$$= \int_{(U\cap B_{j-1})\times X} L(\lambda_j(x_1, \ldots, x_j), f(x_j))\, dP^j + \int_{(U\setminus B_{j-1})\times X} 1\, dP^j$$

$$= \int_{U\cap B_{j-1}} e_L(h_{j-1})\, dP^{j-1} + \int_{U\setminus B_{j-1}} 1\, dP^{j-1}$$
$$\geq \int_{U\cap B_{j-1}} \frac{\epsilon}{2}\, dP^{j-1} + \int_{U\setminus B_{j-1}} \frac{\epsilon}{2}\, dP^{j-1}$$
$$= \int_U \frac{\epsilon}{2}\, dP^{j-1} = \int_{U\times X} \frac{\epsilon}{2}\, dP^j.$$

Da dieses für jede Menge $U \times X \in \sigma(\tilde{r}_1,\ldots,\tilde{r}_{j-1})$ gilt und da $E(\tilde{r}_j|\tilde{r}_1,\ldots,\tilde{r}_{j-1})$ lt. Definition $\sigma(\tilde{r}_1,\ldots,\tilde{r}_{j-1})$-meßbar ist, gilt $E(\tilde{r}_j|\tilde{r}_1,\ldots,\tilde{r}_{j-1}) \geq \epsilon/2$ $(a.s.)$.

Wenden wir nun Korollar 4.5 mit $c = \epsilon/2$ und $\alpha = 1/2$ an, so erhalten wir dann $P(\sum_{i=1}^{t_1} \tilde{r}_i \leq \epsilon t_1/4) \leq e^{-\epsilon t_1/16}$ für fast jeden Punkt[15] aus X^{t_1}. Wegen $t_1 \geq (16/\epsilon)\ln(2/\delta)$ ist $e^{-\epsilon t_1/16} \leq \delta/2$.

Wenn $\tilde{r}_j = 1$, dann gilt $e_L(h_{j-1}) < \epsilon/2$ oder der Abstand in Runde j ist 1. Gilt also $\sum_{j=1}^{t_1} \tilde{r}_j > l$, dann gilt $e_L(h) < \epsilon/2$ für zumindest eine der Hypothesen $h \in H = \{h_0,\ldots,h_{t_1-1}\}$, da der Gesamtabstand, den der on-line Algorithmus A auf beliebigen Folgen von Beispielen verursacht, durch l beschränkt ist. Mit der Wahl $t_1 \geq 4l/\epsilon$ erhalten wir mit $\sum_{j=1}^{t_1} \tilde{r}_j > \epsilon t_1/4$ dann $\sum_{j=1}^{t_1} \tilde{r}_j > l$. Somit gilt mit Wahrscheinlichkeit von mindestens $1-\delta/2$ zumindest für eine der Hypothesen $h \in H$ dann $e_L(h) < \epsilon/2$.

Da $g \geq |H|$, gilt $|H| \leq \min\{g, t_1\}$. Mit Lemma 5.1 erhalten wir nun, daß die Wahrscheinlichkeit dafür, daß in der zweiten Phase eine Hypothese mit $e_L(h) < \epsilon$ ausgegeben wird, mindestens

$$1 - |H| e^{-\frac{\epsilon}{28}\left\lceil \frac{28}{\epsilon} \ln \frac{2\min\{g,t_1\}}{\delta}\right\rceil} \geq 1 - \frac{\delta}{2}$$

ist.

Die Wahrscheinlichkeit dafür, daß die erste Phase eine gewünschte Menge von Hypothesen liefert ist mindestens $1-\delta/2$. Da die zweite Phase unabhängig von der ersten ist, ist die Wahrscheinlichkeit, daß die Transformation eine Hypothese mit $e_L(h) < \epsilon$ liefert, mindestens $1-\delta$. □

Zusammen mit Lemma 4.8 liefert dieser Satz für die „Situation“ von Satz 5.2 fast die gleiche obere Schranke für die Anzahl der benötigten Beispiele wie Satz 5.2. Sei m_γ eine obere Schranke für die Anzahl der Irrtümer, die der konservative on-line Lernalgorithmus A auf beliebigen endlichen Folgen von Beispielen macht. Für $L = L_{0,\gamma}$ sind dann

[15] im Sinne von $(a.s.)$

$l = m_\gamma$ und $g = m_\gamma + 1$. Somit erhalten wir mit Lemma 4.8 und Satz 5.7, daß

$$\begin{aligned} t &= \frac{1}{\epsilon}\left(44 \ln \frac{2}{\delta} + 4l + 28 \ln g + 2\right) \\ &= \frac{1}{\epsilon}\left(44 \ln \frac{2}{\delta} + 4m_\gamma + 28 \ln(m_\gamma + 1) + 2\right) \end{aligned}$$

Beispiele ausreichen, um mit Wahrscheinlichkeit von mindestens $1 - \delta$ eine (ϵ, γ)-gute Hypothese zu erhalten. Mit den nun folgenden Korollaren und Lemma 4.8 gilt entsprechendes im Vergleich zu den Korollaren aus Abschnitt 5.1.

Haben wir Abstandsfunktionen L, die bei einem gegebenen on-line Algorithmus A in einer Runde einen Abstand größer als 1 liefern können, so können wir, sofern der maximale Abstand einer Runde im Intervall $[0, D]$ für ein vorher bekanntes $D \in \mathbb{R}$ liegt, obige Transformation benutzen, indem wir einige Parameter, sowie die Abstandsfunktion anpassen. Dabei definieren wir für eine Abstandsfunktion L und obiges D die Funktion L/D durch $(L/D)(\hat{y}, y) = L(\hat{y}, y)/D$.

Korollar 5.8 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0, D]$ eine Funktion für den Abstand. A lerne F im on-line Modell mit l-beschränktem Abstand bzgl. L. Sei nun g (evtl. ∞) eine obere Schranke für die Anzahl verschiedener Hypothesen, die A auf einer beliebigen endlichen Folge von Beispielen produziert. Falls nun $\epsilon > 0$, $0 < \delta < 1$, $l > 0$,*

$$t_1 = \left\lceil \max\left\{\frac{16D}{\epsilon} \ln \frac{2}{\delta}, \frac{4l}{\epsilon}\right\}\right\rceil$$

und

$$t_2 = \left\lceil \frac{28D}{\epsilon} \ln \frac{2 \min\{g, t_1\}}{\delta} \right\rceil,$$

dann liefert $BLPAC(A, EX, L/D, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L/D, t_1, t_2)$ benutzt ist durch

$$\begin{aligned} t_1 + t_2 &= \frac{1}{\epsilon}\left(44D \ln \frac{2}{\delta} + 4l + 28D \ln \min\left\{g, \frac{16D}{\epsilon} \ln \frac{2}{\delta} + \frac{4l}{\epsilon} + 1\right\}\right) + 2 \\ &= O\left(\frac{1}{\epsilon}\left(l + D \ln \min\left\{\frac{g}{\delta}, \frac{D}{\epsilon\delta}\right\}\right)\right) \end{aligned}$$

beschränkt.

Beweis: Setze $L' = L/D$ und $\epsilon' = \epsilon/D$, dann liegt der Bildbereich von L' in $[0,1]$ und $e_{L'}(h) < \epsilon' \Leftrightarrow e_L(h) < \epsilon$. $l' = l/D$ ist dann eine Schranke für den Gesamtabstand bzgl. L'. Somit folgt nun das Korollar aus Satz 5.7. $\square$

Analog zum vorherigen Abschnitt läßt sich der obige Satz nicht direkt anwenden, falls bei einem on-line Algorithmus der Gesamtabstand bzgl. einer Abstandsfunktion L nicht durch eine Konstante, sondern nur durch eine Funktion, abhängig von der Anzahl der Runden, beschränkt ist. Wir können jedoch auch hier aus obigem Satz obere Schranken für die Anzahl der Beispiele, die die Transformation benötigt, ableiten, falls der Gesamtabstand bzgl. L beim on-line Lernen durch $l(t) = O(\ln t)$ oder $l(t) = O(t^{1/q})$ für $q > 1$ beschränkt ist. Da in diesen Fällen eine nicht triviale Schranke g selten bekannt ist und keinen (nennenswerten) Vorteil in der O-Notation bringt, berücksichtigen wir sie hier nicht. Soll g trotzdem berücksichtigt werden, so kann in der entsprechenden Formel für t_2 einfach „t_1" durch „$\min\{g, t_1\}$" ersetzt werden.

Korollar 5.9 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0,1]$ eine Funktion für den Abstand. A lerne F im on-line Modell mit $l(t)$-beschränktem Abstand bzgl. L. Falls nun $0 < \epsilon \leq 1$, $0 < \delta < 1$, $l(t) = c \ln t$ für alle $t \in \mathbb{N}$ mit $t \geq 12$ und für eine Konstante $c > 0$,*

$$t_1 = \left\lceil \max\left\{\frac{16}{\epsilon}\ln\frac{2}{\delta}, \frac{8c}{\epsilon}\ln\frac{8c}{\epsilon}\right\}\right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon}\ln\frac{2t_1}{\delta}\right\rceil,$$

dann liefert $BLPAC(A, EX, L, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L, t_1, t_2)$ benutzt ist für $c \geq \frac{\epsilon}{8}$ durch

$$t_1 + t_2 = O\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon\delta}\right)$$

beschränkt. Betrachten wir Folgen von Funktionenklassen, so dürfen wir c in der O-Notation nicht als Konstante ansehen. Somit ergibt sich dann

$$t_1 + t_2 = O\left(\frac{1}{\epsilon}\left(c\ln\frac{c}{\epsilon} + \ln\frac{1}{\delta}\right)\right).$$

Beweis: Dieses Korollar läßt sich analog zu Korollar 5.4, jedoch bezogen auf Satz 5.7 statt Satz 5.2, zeigen.

Für Phase 1 ist somit nur zu zeigen, daß

$$t_1 \geq \left\lceil \frac{4}{\epsilon} l(t_1) \right\rceil .$$

Wir wenden Lemma 5.3 1. mit $k = 4c/\epsilon$ und $s = t_1$ an und erhalten

$$t_1 \geq \frac{4}{\epsilon} c \ln t_1 = \frac{4}{\epsilon} l(t_1).$$

Da $t_1 \in \mathbb{N}$ ist, gilt nun

$$t_1 \geq \left\lceil \frac{4}{\epsilon} l(t_1) \right\rceil .$$

Auch für die zweite Phase kann der Beweis von Satz 5.7 benutzt werden. Der Wert von t_2 folgt direkt.

Somit verbleibt nur noch zu zeigen, daß

$$t_1 + t_2 = O\left(\frac{1}{\epsilon}\left(c \ln \frac{c}{\epsilon} + \ln \frac{1}{\delta}\right)\right)$$

für nicht konstantes c gilt. Daraus folgt dann sofort die O-Notation für konstantes c. Für $c \geq \frac{\epsilon}{8}$ erfüllt t_1 offensichtlich die Schranke und für t_2 gilt:

$$\begin{aligned}
t_2 &= O\left(\frac{1}{\epsilon}\left(\ln \frac{1}{\delta} + \ln t_1\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln \frac{1}{\delta} + \ln \frac{1}{\epsilon} + \ln\left(c \ln \frac{c}{\epsilon} + \ln \frac{1}{\delta}\right)\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln \frac{1}{\delta} + \ln \frac{c}{\epsilon}\right)\right).
\end{aligned}$$

□

Auch hier können wir mit der gleichen Konstruktion wie bei Korollar 5.8 Abstandsfunktionen L, die Abstände größer als 1 liefern, benutzen.

Korollar 5.10 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0, D]$ eine Funktion für den Abstand. A lerne F im* *on-line Modell mit $l(t)$-beschränktem Abstand bzgl. L. Falls nun $0 < \epsilon \leq D$, $0 < \delta < 1$, $l(t) = c \ln t$ für alle $t \in \mathbb{N}$ mit $t \geq 12$ und für eine Konstante $c > 0$,*

$$t_1 = \left\lceil \max\left\{ \frac{16D}{\epsilon} \ln \frac{2}{\delta}, \frac{8c}{\epsilon} \ln \frac{8c}{\epsilon} \right\} \right\rceil$$

und

$$t_2 = \left\lceil \frac{28D}{\epsilon} \ln \frac{2t_1}{\delta} \right\rceil,$$

dann liefert $BLPAC(A, EX, L/D, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L/D, t_1, t_2)$ benutzt ist für $c \geq \frac{D\epsilon}{8}$ durch

$$t_1 + t_2 = O\left(\frac{D}{\epsilon} \ln \frac{D}{\epsilon\delta} \right)$$

beschränkt. Betrachten wir Folgen von Funktionenklassen, so dürfen wir c in der O-Notation nicht als Konstante ansehen. Somit ergibt sich dann

$$t_1 + t_2 = O\left(\frac{1}{\epsilon} \left(c \ln \frac{c}{\epsilon} + D \ln \frac{cD}{\epsilon\delta} \right) \right).$$

Korollar 5.11 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0, 1]$ eine Funktion für den Abstand. A lerne F im on-line Modell mit $l(t)$-beschränktem Abstand bzgl. L. Falls nun $0 < \epsilon \leq 1$, $0 < \delta < 1$, $l(t) = ct^{1/q}$ für alle $t \in \mathbb{N}$ und für Konstanten $c > 0$ und $q > 1$,*

$$t_1 = \left\lceil \max\left\{ \frac{16}{\epsilon} \ln \frac{2}{\delta}, \left(\frac{4c}{\epsilon} \right)^{\frac{q}{q-1}} \right\} \right\rceil$$

und

$$t_2 = \left\lceil \frac{28}{\epsilon} \ln \frac{2t_1}{\delta} \right\rceil,$$

dann liefert $BLPAC(A, EX, L, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L, t_1, t_2)$ benutzt ist durch

$$t_1 + t_2 = O\left(\left(\frac{1}{\epsilon} \right)^{\frac{q}{q-1}} + \frac{1}{\epsilon} \ln \frac{1}{\delta} \right)$$

beschränkt. Betrachten wir Folgen von Funktionenklassen, so dürfen wir c und q in der O-Notation nicht als Konstanten ansehen. Somit ergibt sich dann

$$t_1 + t_2 = O\left(\left(\frac{4c}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{q}{q-1}\ln\frac{c}{\epsilon}\right)\right).$$

Beweis: Analog zu Korollar 5.9 reicht es zu zeigen, daß

$$t_1 \geq \left\lceil \frac{4c}{\epsilon} t_1^{\frac{1}{q}} \right\rceil$$

und

$$t_2 = O\left(\left(\frac{4c}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{q}{q-1}\ln\frac{c}{\epsilon}\right)\right),$$

da der Rest offensichtlich ist.

Wegen $t_1 \geq \left(\frac{4c}{\epsilon}\right)^{\frac{q}{q-1}}$ läßt sich Lemma 5.3 3. mit $k = 4c/\epsilon$ und $s = t_1$ anwenden. Wir erhalten

$$t_1 \geq \frac{4c}{\epsilon} t_1^{\frac{1}{q}}$$

und da $t_1 \in \mathbb{N}$ ist, gilt nun

$$t_1 \geq \left\lceil \frac{4c}{\epsilon} t_1^{\frac{1}{q}} \right\rceil.$$

Es verbleibt also nur noch t_2 abzuschätzen:

$$\begin{aligned}
t_2 &= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln t_1\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\left(\left(\frac{c}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\frac{1}{\epsilon} + \ln\left(\left(\frac{c^q}{\epsilon}\right)^{\frac{1}{q-1}} + \ln\frac{1}{\delta}\right)\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\frac{1}{\epsilon} + \ln\left(\frac{c^q}{\epsilon}\right)^{\frac{1}{q-1}} + \ln\ln\frac{1}{\delta}\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\left(\frac{c}{\epsilon}\right)^{\frac{q}{q-1}}\right)\right) \\
&= O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \frac{q}{q-1}\ln\frac{c}{\epsilon}\right)\right).
\end{aligned}$$

□

Korollar 5.12 *Sei EX ein Orakel für eine Zielfunktion $f : X \to Y$ aus einer Funktionenklasse F und $L : Y^2 \to [0, D]$ eine Funktion für den Abstand. A lerne F im on-line Modell mit $l(t)$-beschränktem Abstand bzgl. L. Falls nun $0 < \epsilon \leq D$, $0 < \delta < 1$, $l(t) = ct^{1/q}$ für alle $t \in \mathbb{N}$ und für Konstanten $c > 0$ und $q > 1$,*

$$t_1 = \left\lceil \max\left\{ \frac{16D}{\epsilon} \ln \frac{2}{\delta}, \left(\frac{4c}{\epsilon}\right)^{\frac{q}{q-1}} \right\} \right\rceil$$

und

$$t_2 = \left\lceil \frac{28D}{\epsilon} \ln \frac{2t_1}{\delta} \right\rceil,$$

dann liefert $BLPAC(A, EX, L/D, t_1, t_2)$ mit Wahrscheinlichkeit von mindestens $1 - \delta$ (bzgl. der Wahrscheinlichkeitsverteilung von EX) eine Hypothese h, mit $e_L(h) < \epsilon$ (bzgl. f).
Die Anzahl der Beispiele, die $BLPAC(A, EX, L/D, t_1, t_2)$ benutzt ist durch

$$t_1 + t_2 = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{D}{\epsilon}\left(\ln \frac{D}{\delta} + \ln \frac{1}{\epsilon}\right) \right)$$

beschränkt. Betrachten wir Folgen von Funktionenklassen, so dürfen wir c und q in der O-Notation nicht als Konstanten ansehen. Somit ergibt sich dann

$$t_1 + t_2 = O\left(\left(\frac{4c}{\epsilon}\right)^{\frac{q}{q-1}} + \frac{D}{\epsilon}\left(\ln \frac{D}{\delta} + \frac{q}{q-1} \ln \frac{c}{\epsilon}\right) \right).$$

Bei den beiden vorherigen Korollaren reicht es aus, wenn $l(t) = ct^{1/q}$ für $t \geq 12$ gilt (s. Beweis von Korollar 5.9), wodurch evtl. eine kleinere Konstante c möglich ist.

5.3 Transformationen zwischen dem (ϵ, γ)-guten Modell und den Modellen mit ϵ-beschränkter Differenz

Wir geben nun noch Transformationen zwischen den hier betrachteten off-line Lernmodellen an.

Das folgende Lemma ist eine Erweiterung eines Lemmas von Kearns und Schapire [KS94] auf die Abstandsfunktionen L_p für $p > 0$.

Lemma 5.13 *Seien sowohl h als auch f Funktionen $X \to \mathbb{R}$. Sei $p > 0$. Falls $e_{L_p}(h) < \epsilon$, dann ist h ein $(\epsilon/\gamma^p, \gamma)$-gutes Modell für entsprechendes f.*

Beweis: Definiere

$$err_h(x) = \begin{cases} 1 & \text{falls } |f(x) - h(x)| \geq \gamma \\ 0 & \text{sonst} \end{cases} .$$

Für $p > 0$ gilt

$$\begin{aligned} \epsilon &> e_{L_p}(h) \\ &= \int_X |f(x) - h(x)|^p \, dP \\ &\geq \gamma^p \int_X err_h(x) \, dP \\ &= \gamma^p P(\{x \mid |f(x) - h(x)| \geq \gamma\}). \end{aligned}$$

Daraus folgt sofort die Behauptung des Lemmas. □

Korollar 5.14 *Seien sowohl h als auch f Funktionen $X \to \mathbb{R}$.*

1. *Falls $e_1(h) < \epsilon$, dann ist h ein $(\epsilon/\gamma, \gamma)$-gutes Modell für entsprechendes f.*
2. *Falls $e_2(h) < \epsilon$, dann ist h ein $(\epsilon/\gamma^2, \gamma)$-gutes Modell für entsprechendes f.*

Das folgende Lemma ist eine Erweiterung eines Lemmas von Kearns und Schapire [KS94] auf die Abstandsfunktionen L_p für $p > 0$ und auf Klassen von Funktionen $f : X \to [C, C + D]$ für beliebiges C, D, statt $f : X \to [0, 1]$.

Lemma 5.15 *Seien sowohl h als auch f Funktionen $X \to [C, C + D]$ für beliebige Konstanten $C \in \mathbb{R}$ und $D \in \mathbb{R}^{>0}$. Sei $p > 0$. Falls nun h ein (ϵ, γ)-gutes Modell für f ist, so gilt $e_{L_p}(h) < D^p\epsilon + \gamma^p$. Insbesondere gilt dann für Funktionen $f, g : X \to [0, 1]$: $e_1(h) < \epsilon + \gamma$ und $e_2(h) < \epsilon + \gamma^2$.*

Beweis:

$$\begin{aligned} e_{L_p}(h) &= \int_X |f(x) - h(x)|^p \, dP \\ &\leq D^p P(\{x \mid |f(x) - h(x)|^p \geq \gamma^p\}) + \gamma^p P(\{x \mid |f(x) - h(x)|^p < \gamma^p\}) \\ &\leq D^p Error(h) + \gamma^p \\ &< D^p\epsilon + \gamma^p \end{aligned}$$

□

6 On-line Resultate

6.1 Funktionen mit beschränkter Steigung

Kimber und Long [KL92] haben eine Hierarchie von Funktionenklassen mit beschränkter Steigung eingeführt, on-line Lernalgorithmen für einen Teil dieser Klassen vorgestellt und obere Schranken für deren Gesamtabstand angegeben.

Sei $\mathcal{F}^*$ die Menge aller stetigen und stückweise 2 mal differenzierbaren Funktionen f mit Definitionsbereich $[0,1]$ und $f(0) = 0$. Nun sei $\mathcal{F}_p = \{f \mid f \in \mathcal{F}^* \wedge p\text{-Norm von } f' \leq 1\}$ für $p \in \{1, 2, \infty\}$.

Es ist leicht einzusehen, daß $\mathcal{F}_\infty \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_1$.

Algorithmus $LININT(S, x)$

Parameter:
S ist die Folge der Instanz-Funktionswert-Paare der vorherigen Runden.
x ist die Instanz der aktuellen Runde.

Beschreibung:
$LININT$ approximiert die Zielfunktion, indem es die Werte für Instanzen, deren Funktionswerte noch nicht bekannt sind, mit Hilfe der bereits bekannten Instanz-Funktionswert-Paare linear interpoliert.
Formal: Sei $S' = \{(u_i, v_i) \mid 1 \leq i \leq m\}$ eine endliche Teilmenge von $[0,1] \times [-1,1]$, so daß $u_1 < u_2 < \cdots < u_m$. Definiere $f_{S'} : [0,1] \to [-1,1]$ wie folgt: für alle x, $f_\emptyset(x) = 0$ und

$$f_{S'}(x) = \begin{cases} v_1 & \text{falls } x \leq u_1 \\ v_i + \frac{(x-u_i)(v_{i+1}-v_i)}{u_{i+1}-u_i} & \text{falls } x \in (u_i, u_{i+1}] \\ v_m & \text{falls } x > u_m \end{cases}$$

für $|S'| \geq 1$. Sei S' die Menge der Elemente der Folge S. Dann liefert $LININT(S, x)$ die Vorhersage $f_{S'}(x)$. •

Algorithmus $NN(S, x)$

Parameter:
$S = ((x_i, y_i))_{|S|}$ ist die Folge der Instanz-Funktionswert-Paare der vorherigen Runden.
x ist die Instanz der aktuellen Runde.

Beschreibung:
Für $S \notin \emptyset$ liefert $NN(S, x)$ den Funktionswert des „nächsten Nachbarn" aus S und

$NN(\emptyset, x)$ liefert 0.
Formal: Für $S \notin \emptyset$ bestimmt $NN(S,x)$ zuerst das x_i mit $|x_i - x|$ minimal. (Falls es zwei x_i mit minimalem Abstand zu x gibt, nimm das kleinere.) Dann liefert $NN(S,x)$ das zu x_i gehörige y_i. •

Folgende Resultate haben Kimber und Long [KL92] gezeigt:

Lemma 6.1 *LININT lernt $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) im on-line Modell mit 1-beschränktem quadratischem Abstand.*

Korollar 6.2 *LININT lernt $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) für $p \geq 2$ im on-line Modell mit 1-beschränktem Abstand bzgl. L_p.*

Beweis: Da der quadratische Abstand in jeder einzelnen Runde durch 1 beschränkt ist, können höhere Potenzen den Wert für den Abstand höchstens verkleinern. Mit Lemma 6.1 folgt dann das Korollar. □

Lemma 6.3 *LININT lernt $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) für $\eta > 0$ und $p = 1 + \eta$ im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p für*

$$l_p = 2 + \frac{1}{2\eta \ln 2}.$$

Lemma 6.4 *NN lernt $\mathcal{F}_\infty$ für $\eta > 0$ und $p = 1+\eta$ im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p für*

$$l_p = 1 + \frac{1}{2\eta \ln 2}.$$

Lemma 6.5 *LININT lernt $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) für alle $t \geq 3$ im on-line Modell mit $l(t)$-beschränktem linearem Abstand für*

$$l(t) = 2e + \frac{e}{2 \ln 2} \ln t.$$

Lemma 6.6 *NN lernt $\mathcal{F}_\infty$ für alle $t \geq 3$ im on-line Modell mit $l(t)$-beschränktem linearem Abstand für*

$$l(t) = e + \frac{e}{2 \ln 2} \ln t.$$

Aus dem maximalen linearen Abstand D_1, den ein on-line Algorithmus A in einer einzelnen Runde verursachen kann, erhält man für $p > 0$ den Wert $D = D_1^p$, so daß der Bildbereich von L_p in $[0, D]$ liegt. Wir schätzen nun D_1 für $LININT$ und NN bei Zielfunktionen aus $\mathcal{F}_2$ (und somit auch bei Zielfunktionen aus $\mathcal{F}_\infty$) ab.

Lemma 6.7 *Für jedes $f \in \mathcal{F}_1$ und $0 \leq x_1 < x_2 \leq 1$ gilt $|f(x_2) - f(x_1)| \leq 1$.*

Beweis: Annahme: $|f(x_2) - f(x_1)| > 1$. Wir können o.B.d.A. annehmen, daß $f(x_1) < f(x_2)$. Es gilt $1 < f(x_2) - f(x_1) = \int_{x_1}^{x_2} f'(x)\, dx$. Somit gilt $f \notin \mathcal{F}_1$, was im Widerspruch zur Voraussetzung $f \in \mathcal{F}_1$ steht. □

Korollar 6.8 *$LININT$ und NN liefern in einer einzelnen Runde für jede Zielfunktion $f \in \mathcal{F}_2$ (und somit auch für jede Zielfunktion $f \in \mathcal{F}_\infty$) einen linearen Abstand von höchstens 1.*

Beweis: NN liefert als Vorhersage $f(0) = 0$ oder einen Wert, der Funktionswert von einer vorherigen Instanz ist. Somit gibt es zu jeder Vorhersage $\hat{y}$ von NN ein $x \in [0, 1]$ mit $f(x) = \hat{y}$. Wegen Lemma 6.7 und $\mathcal{F}_2 \subseteq \mathcal{F}_1$ kann also die Vorhersage von NN höchstens um 1 vom tatsächlichen Funktionswert abweichen.

$LININT$ liefert als Vorhersage $f(0) = 0$, einen Wert, der Funktionswert von einer vorherigen Instanz ist oder eine lineare Interpolation zwischen zwei Funktionswerten vorheriger Instanzen. Da die Zielfunktion f stetig im Intervall $[0, 1]$ ist, gibt es zu jeder Vorhersage $\hat{y}$ von $LININT$ ein $x \in [0, 1]$ mit $f(x) = \hat{y}$. Wie bei NN kann somit die Vorhersage von $LININT$ höchstens um 1 vom tatsächlichen Funktionswert abweichen. □

Da eine Schranke l für den quadratischen Gesamtabstand auch eine Schranke für den quadratischen Abstand in jeder einzelnen Runde ist, ist $\sqrt{l}$ eine Schranke für den linearen Abstand in jeder einzelnen Runde. Somit läßt sich obiges Korollar für $LININT$ auch aus Lemma 6.1 gewinnen.

6.2 Lineare Funktionen

$F_n(M, W)$ sei die Klasse aller linearen Funktionen $f : [0, M]^n \to [0, MW]$ mit $f(\vec{x}) = \vec{\mu}\vec{x}$, wobei $\vec{\mu} \in [0, W]^n$ und $\|\vec{\mu}\|_1 = W$.

Littlestone, Long und Warmuth [LLW91] haben einen on-line Lernalgorithmus für die Klasse $F_n(1,1)$ vorgestellt.[16] Dieser wurde von Kivinen und Warmuth [KW94] verbessert.

Der Algorithmus EG (angepaßt auf unsere Lernsituation) von Kivinen und Warmuth [KW94] eignet sich direkt zum Lernen von Funktionenklassen $F_n(M,1)$.

Algorithmus EG

Beschreibung:
Die Hypothese h_{i-1}, die in der i-ten Runde zur Vorhersage benutzt wird, wird durch einen Gewichtsvektor $\vec{w}_i$ dargestellt. Es gilt $\hat{y}_i = h_{i-1}(\vec{x}_i) = \vec{w}_i\vec{x}_i$. Für die erste Runde wird $\vec{w}_1 = (1/n, \ldots, 1/n)$ gesetzt und dann $\hat{y}_1 = h_0(x_1)$ vorhergesagt. Für die Runde $i+1$ mit $i \geq 1$ wird $\vec{w}_{i+1} = (w_{i+1,1}, \ldots, w_{i+1,n})$ wie folgt berechnet:

$$w_{i+1,j} = \frac{w_{i,j}r_{i,j}}{\sum_{k=1}^{n} w_{i,k}r_{i,k}},$$

wobei

$$r_{i,j} = \exp\left(\frac{4(y_i - \hat{y}_i)x_{i,j}}{R_i^2}\right).$$

und $R_i = \max_j x_{i,j} - \min_j x_{i,j}$. •

Mit einer einfachen Standardtransformation (s. z. B. [Lit88, LLW91]) erhalten wir für vorher bekanntes W den Lernalgorithmus EG_W für $F_n(M,W)$ aus EG.

Algorithmus EG_W

Parameter:
W ist der Wert für $\|\vec{\mu}\|_1$.

Beschreibung:
EG_W übergibt die Instanz $\vec{x}$ an EG. Statt der Vorhersage $\hat{y}$ von EG liefert EG_W dann die Vorhersage $\hat{y}W$. Nun erhält EG_W den tatsächlichen Wert y und übergibt y/W an EG als tatsächlichen Wert. •

Für einige Korollare benutzen wir die Voraussetzung, daß EG nur Instanzen $\vec{x} \in [0,1]^n$ erhält. Hier läßt sich, ebenfalls mit einer einfachen Standardtransformation, für vorher bekanntes M und W der Lernalgorithmus $EG_{M,W}$ für $F_n(M,W)$ aus EG gewinnen.

[16] Mit Standardtransformationen läßt sich damit ein Lernalgorithmus für die Klassen $F_n(M,W)$ gewinnen (s. z. B. Algorithmus $EG_{M,W}$ unten).

Algorithmus $EG_{M,W}$

Parameter:
M beschränkt die Instanzen wie folgt: $\vec{x} \in [0, M]^n$.
W ist der Wert für $\|\vec{\mu}\|_1$.

Beschreibung:
$EG_{M,W}$ gibt für eine Instanz $\vec{x}$ die Instanz $\vec{x}/M$ an EG. Statt der Vorhersage $\hat{y}$ von EG liefert $EG_{M,W}$ dann die Vorhersage $\hat{y}MW$. Nun erhält $EG_{M,W}$ den tatsächlichen Wert y und übergibt $y/(MW)$ an EG als tatsächlichen Wert. •

Lemma 6.9 ([KW94]) *EG lernt $F_n(M, 1)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit l_M-beschränktem quadratischem Abstand für*

$$l_M = \frac{1}{2} M^2 \ln n.$$

Korollar 6.10 *EG_W lernt $F_n(M, W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit $l_{M,W}$-beschränktem quadratischem Abstand für*

$$l_{M,W} = \frac{1}{2} (MW)^2 \ln n.$$

Beweis: Der quadratische Gesamtabstand von EG ist wegen Lemma 6.9 durch $(1/2)M^2$ $\ln n$ beschränkt. Durch den Faktor W erhält der quadratische Abstand den Faktor W^2. □

Korollar 6.11 *EG lernt $F_n(1, 1)$ für $n \in \mathbb{N}, n \geq 2$ und $p \geq 2$ im on-line Modell mit $(\ln n/2)$-beschränktem Abstand bzgl. L_p.*

Beweis: Da $\vec{x} \in [0, 1]$ und in jeder Runde $\|\vec{w}\|_1 = 1$, sowie $\vec{w} \in [0, 1]^n$, gilt $\hat{y} \in [0, 1]$. Da offensichtlich auch $y \in [0, 1]$, ist der quadratische Abstand in jeder einzelnen Runde durch 1 beschränkt. Somit können höhere Potenzen den Wert für den Abstand höchstens verkleinern. Mit Lemma 6.9 folgt dann das Korollar. □

Korollar 6.12 *$EG_{M,W}$ lernt $F_n(M, W)$ für $n \in \mathbb{N}, n \geq 2$ und $p \geq 2$ im on-line Modell mit $l_{M,W,p}$-beschränktem Abstand bzgl. L_p für*

$$l_{M,W,p} = \frac{1}{2} (MW)^p \ln n.$$

Beweis: Der Gesamtabstand (bzgl. L_p) von EG ist wegen Korollar 6.11 durch $\ln n/2$ beschränkt. Durch den Faktor MW erhält der Abstand bzgl. L_p den Faktor $(MW)^p$. □

Korollar 6.13 *EG_W lernt $F_n(M, W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit $m_{M,W,\gamma}$-beschränktem Abstand bzgl. $L_{0,\gamma}$ für*

$$m_{M,W,\gamma} = \left\lfloor \frac{1}{2} \left(\frac{MW}{\gamma} \right)^2 \ln n \right\rfloor .$$

Beweis: Das Korollar folgt sofort aus Korollar 6.10 und Lemma 4.6. □

Bernstein [Ber92] unterscheidet zwischen Lernalgorithmen mit und ohne Speicher. Ein Lernalgorithmus ohne Speicher merkt sich zwischen zwei Runden nur die Hypothese. Ein Lernalgorithmus mit Speicher merkt sich dagegen noch weitere Informationen, z. B. die bis dahin gezeigte Beispielfolge oder daraus berechnete Werte. Bernstein hat auch gezeigt, daß der Algorithmus von Littlestone, Long und Warmuth [LLW91] $F_n(1, 1)$ im on-line Modell nicht mit endlich beschränktem linearem Abstand lernt.

Bernstein [Ber92] hat nun folgenden Algorithmus mit Speicher vorgestellt.

Algorithmus $B(GUESS)$

Parameter:
$GUESS$ eine Funktion mit $GUESS(\vec{0}) = 0$ (z. B. ein anderer Lernalgorithmus oder die konstante Null-Funktion).

Beschreibung:
Der Start-Gewichtsvektor des Algorithmus ist $\vec{w}_1 = \vec{0}$. Der Algorithmus erhält in Runde t die Instanz $\vec{x}_t$ und spaltet diese in

$$\vec{v}_t = \sum_{j=1}^{t-1} \frac{\vec{x}_t \vec{u}_j}{\|\vec{u}_j\|_2^2} \vec{u}_j$$

und

$$\vec{u}_t = \vec{x}_t - \vec{v}_t$$

auf. Er liefert die Vorhersage $\hat{y}_t = \vec{w}_t \vec{v}_t + GUESS(\vec{u}_t)$. Nun erhält der Algorithmus den tatsächlichen Wert y_t für $\vec{x}_t$. Nun wird der neue Gewichtsvektor

$$\vec{w}_{t+1} = \vec{w}_t + \frac{y_t - \hat{y}_t}{\|\vec{u}_t\|_2^2} \vec{u}_t$$

berechnet. •

$B(GUESS)$ spaltet jede Instanz $\vec{x}_t$ in zwei Vektoren $\vec{u}_t$ und $\vec{v}_t$ auf. Dabei liegt $\vec{v}_t$ in dem von $\vec{x}_1$ bis $\vec{x}_{t-1}$ aufgespannten Unterraum $\mathcal{U}$ und $\vec{u}_t$ in dem zu $\mathcal{U}$ senkrechten Unterraum $\mathcal{U}^\perp$. Die Menge der Vektoren $\{\vec{u}_1, \ldots, \vec{u}_{t-1}\} \setminus \{\vec{0}\}$ bilden eine Orthogonalbasis von $\mathcal{U}$. Da die tatsächlichen Werte $y_1, \ldots, y_{t-1}$ bekannt sind, sind auch die tatsächlichen Werte für $\{\vec{u}_1, \ldots, \vec{u}_{t-1}\}$ und damit für jeden Vektor $\vec{v} \in \mathcal{U}$ bekannt. In Runde t ist $f(\vec{v}_t) = \vec{w}_t \vec{v}_t$ für $\vec{v}_t \in \mathcal{U}$. Wegen $y_t = f(\vec{x}_t) = f(\vec{u}_t) + f(\vec{v}_t)$ muß somit nur $f(\vec{u}_t)$ geschätzt werden, was mit $GUESS(\vec{u}_t)$ geschieht.

Lemma 6.14 *Der Algorithmus $B(GUESS)$ erhalte beim on-line Lernen einer linearen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ eine Folge $S = ((\vec{x}_1, y_1), \ldots, (\vec{x}_t, y_t)) \in (\mathbb{R}^n \times \mathbb{R})^t$ von t Beispielen, dann macht $B(GUESS)$ maximal n falsche Vorhersagen.*

Beweis: $B(GUESS)$ liefert auf Instanzen $\vec{x} \in \mathcal{U}$ wegen $GUESS(\vec{0}) = 0$ und obiger Beobachtung $f(\vec{x})$ als Vorhersage. Somit liefert $B(GUESS)$ nur dann eine falsche Vorhersage, wenn $\vec{x} \notin \mathcal{U}$. Dann vergrößert sich aber der bekannte Unterraum $\mathcal{U}$ um eine Dimension[17]. Aus der Tatsache, daß der gesamte Instanzenraum $X = \mathbb{R}^n$ n-dimensional ist, folgt sofort das Lemma. □

Korollar 6.15 *Für $n \in \mathbb{N}, n \geq 2$ liefert $B(EG_{M,W})$ beim Lernen von $F_n(M,W)$ maximal n falsche Vorhersagen. $B(EG_{M,W})$ lernt somit $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit n-beschränktem Abstand bzgl. L_0.*

Lemma 6.16 *Sei $0 < p < 2$ und $n \in \mathbb{N}$. $\sum_{i=1}^n d_i^p$ wird unter der Randbedingung $\sum_{i=1}^n d_i^2 \leq D$ maximal, wenn $d_i = \sqrt{D/n}$ für $i = 1, \ldots, n$.*

Beweis: Offensichtlich können wir die Bedingung $\sum_{i=1}^n d_i^2 \leq D$ durch $\sum_{i=1}^n d_i^2 = D$ ersetzen. Wir benutzen nun die Multiplikatorenregel von Lagrange. Wir definieren $\phi(x) = \sum_{i=1}^n d_i^p + \lambda \sum_{i=1}^n d_i^2 - D$ und setzen die partiellen Ableitungen

$$\frac{\partial \phi}{\partial d_i} = p d_i^{p-1} + 2\lambda d_i$$

[17] Genauer: Der Vektor $\vec{x}$ vergrößert den Unterraum $\mathcal{U}$ und der zu $\vec{x}$ gehörige Vektor $\vec{u}$ liefert mit den alten Basisvektoren des alten $\mathcal{U}$ wieder eine Orthogonalbasis für $\mathcal{U}$. Somit ist nun $\mathcal{U}$ um eine Dimension größer. Der tatsächliche Wert $y_{\vec{u}}$ für $\vec{u}$ ist $y - \hat{y}$. Somit ist der tatsächliche Wert für jeden Vektor aus $\mathcal{U}$ bekannt.

für $i = 1, \ldots, n$ gleich 0. Somit erhalten wir zusammen mit der Randbedingung ein Gleichungssystem mit $n + 1$ Gleichungen. Wir lösen dieses auf und erhalten $d_i = (-2\lambda/p)^{1/(p-2)}$ für $i = 1, \ldots, n$ aus den ersten n Gleichungen. Mit der Randbedingung erhalten wir $n(-2\lambda/p)^{2/(p-2)} = D$, also $\lambda = -(p/2)(D/n)^{(p-2)/2}$. Wir erhalten somit $d_i = \sqrt{D/n}$ für $i = 1, \ldots, n$ als Extremalstelle. Es ist leicht einzusehen, daß $d_i = \sqrt{D/n}$ für $i = 1, \ldots, n$ dann für $0 < p < 2$ eine Maximalstelle ist. □

Lemma 6.17 *$B(EG)$ lernt $F_n(1,1)$ für $n \in \mathbb{N}, n \geq 2$ und $p > 0$ im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p für*

$$l_p = \begin{cases} n^{1-(p/2)} \left(\frac{\ln n}{2}\right)^{p/2} & \textit{falls } 0 < p < 2 \\ \frac{\ln n}{2} & \textit{falls } p \geq 2 \end{cases} .$$

Beweis: Wegen Korollar 6.15 wird höchstens n mal durch EG geschätzt[18]. Nur in den Runden in denen geschätzt wird tritt ein Abstand > 0 (genau der durch die Schätzung verursachte Abstand) auf. Wegen Korollar 6.10 (mit $M = W = 1$) ist die Summe der Quadrate dieser n Abstände durch $\ln n/2$ beschränkt. Somit gilt das Lemma für $p = 2$.

Für $p > 2$ gilt die gleiche Schranke wie für $p = 2$, da die Einzelabstände alle durch 1 beschränkt sind.

Für $0 < p < 2$ liefert Lemma 6.16 (mit $D = \ln n/2$), daß der Abstand bzgl. L_p maximal wird, wenn der lineare Abstand in jeder der n Runden $\sqrt{\ln n/(2n)}$ beträgt. Somit ergibt sich für den Gesamtabstand bzgl. L_p dann $n\sqrt{\ln n/(2n)}^p = n^{1-(p/2)}(\ln n/2)^{p/2}$. □

Korollar 6.18 *$B(EG_{M,W})$ lernt $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ und $p > 0$ im on-line Modell mit $l_{M,W,p}$-beschränktem Abstand bzgl. L_p für*

$$l_{M,W,p} = \begin{cases} (MW)^p n^{1-(p/2)} \left(\frac{\ln n}{2}\right)^{p/2} & \textit{falls } 0 < p < 2 \\ (MW)^p \frac{\ln n}{2} & \textit{falls } p \geq 2 \end{cases} .$$

Beweis: Der Gesamtabstand (bzgl. L_p) von EG aus Lemma 6.17 erhält durch den Faktor MW den Faktor $(MW)^p$. □

[18] Die Schätzungen für den Vektor $\vec{0}$ sind exakt und werden deshalb nicht mitgezählt.

Korollar 6.19 *$B(EG_{M,W})$ lernt $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit $l_{M,W}$-beschränktem linearem Abstand für*

$$l_{M,W} = MW\sqrt{\frac{n \ln n}{2}}.$$

Korollar 6.20 *$B(EG_{M,W})$ lernt $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit $l_{M,W}$-beschränktem quadratischem Abstand für*

$$l_{M,W} = \frac{1}{2}(MW)^2 \ln n.$$

Für $p \geq 2$ sind die Schranken für den Abstand bzgl. L_p beim on-line Lernen von $F_n(M,W)$ für $B(EG_{M,W})$ und $EG_{M,W}$ gleich, jedoch wissen wir, wegen Korollar 6.15, daß $B(EG_{M,W})$ maximal n falsche Vorhersagen macht, so daß maximal $n+1$ verschiedene Hypothesen produziert werden, was ein Vorteil bei der Anwendung der Transformationen ist. Außerdem erhalten wir so eine bessere Schranke (als durch Korollar 6.13 gegeben) für den Abstand bzgl. $L_{0,\gamma}$ beim on-line Lernen von $F_n(M,W)$.

Korollar 6.21 *$B(EG_{M,W})$ lernt $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ im on-line Modell mit $m_{M,W,\gamma}$-beschränktem Abstand bzgl. $L_{0,\gamma}$ für*

$$m_{M,W,\gamma} = \min\left\{ \left\lfloor \frac{1}{2}\left(\frac{MW}{\gamma}\right)^2 \ln n \right\rfloor, n \right\}.$$

Beweis: Aus Korollar 6.20 und Lemma 4.6 folgt die Schranke $\left\lfloor \frac{1}{2}\left(\frac{MW}{\gamma}\right)^2 \ln n \right\rfloor$ und aus Korollar 6.15 die Schranke n. □

Wir schätzen nun den maximalen linearen Abstand D_1 einer Runde für EG und $B(EG)$ bei Zielfunktionen aus $F_n(M,1)$, sowie für EG_W, $EG_{M,W}$ und $B(EG_{M,W})$ bei Zielfunktionen aus $F_n(M,W)$ ab.

Lemma 6.22 *EG liefert in einer einzelnen Runde für jede Zielfunktion $f \in F_n(M,1)$ einen linearen Abstand von höchstens M. EG_W und $EG_{M,W}$ liefern in einer einzelnen Runde für jede Zielfunktion $f \in F_n(M,W)$ einen linearen Abstand von höchstens MW.*

Beweis: Die Algorithmen benutzen nur Hypothesen aus $F_n(M,1)$ bzw. $F_n(M,W)$. Das Bild jeder Funktion aus diesen Klassen liegt in $[0,M]$ bzw. $[0,MW]$, woraus sofort das Lemma folgt. □

Lemma 6.23 *$B(EG)$ liefert in einer einzelnen Runde für jede Zielfunktion $f \in F_n(M, 1)$ einen linearen Abstand von höchstens M. $B(EG_{M,W})$ liefert in einer einzelnen Runde für jede Zielfunktion $f \in F_n(M, W)$ einen linearen Abstand von höchstens MW.*

Beweis: Die Instanz $\vec{x} \in [0, M]$ spaltet sich in die Vektoren $\vec{v} \in \mathcal{U}$ und $\vec{u} \in \mathcal{U}^{\perp}$ auf. Es gilt $\vec{v}, \vec{u} \in [0, M]$, $f(\vec{v}), f(\vec{u}) \geq 0$ und $f(\vec{x}) = f(\vec{v}) + f(\vec{u})$. $f(\vec{v})$ kann berechnet werden und $f(\vec{u})$ wird von EG bzw. $EG_{M,W}$ abgeschätzt. $f(\vec{u}) \in [0, M]$ bzw. $\in [0, MW]$. Die Schätzung liefert einen Wert aus dem Intervall $[0, M]$ bzw. $[0, MW]$. Somit ist der lineare Abstand durch M bzw. MW beschränkt. □

Wir definieren nun eine weitere Klasse linearer Funktionen.

$G_n(M, W)$ sei die Klasse aller linearen Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ mit $f(\vec{x}) = \vec{\mu}\vec{x}$, wobei $\|\vec{\mu}\|_2 \leq W$ und $\|\vec{x}\|_2 \leq M$ (d. h., f ist auf $\vec{x}$ mit $\|\vec{x}\|_r > M$ nicht definiert).

Bildlich gesprochen sind die Gewichtsvektoren aus der n-dimensionalen Kugel mit Radius W und die Instanzen aus der n-dimensionalen Kugel mit Radius M.

Algorithmus B

Beschreibung:
B ist $B(GUESS)$ mit $GUESS(\vec{x}) = \vec{0}$. •

Korollar 6.24 *Für $n \in \mathbb{N}$ liefert B beim Lernen von $G_n(M, W)$ maximal n falsche Vorhersagen. Somit wird $G_n(M, W)$ von B für $n \in \mathbb{N}$ im on-line Modell mit n-beschränktem Abstand bzgl. L_0 gelernt.*

Lemma 6.25 ([Ber92]) *B lernt $G_n(1, 1)$ für $n \in \mathbb{N}$ und $p > 0$ im on-line Modell mit l_p-beschränktem Abstand bzgl. L_p für*

$$l_p = n^{\frac{\max\{p,2\}-p}{2}}.$$

Korollar 6.26 *B lernt $G_n(M, W)$ für $n \in \mathbb{N}$ und $p > 0$ im on-line Modell mit $l_{M,W,p}$-beschränktem Abstand bzgl. L_p für*

$$l_{M,W,p} = (MW)^p n^{\frac{\max\{p,2\}-p}{2}}.$$

Korollar 6.27 *B lernt $G_n(M, W)$ für $n \in \mathbb{N}$ im on-line Modell mit $(MW)^2$-beschränktem quadratischem Abstand.*

Korollar 6.28 *B lernt $G_n(M,W)$ für $n \in \mathbb{N}$ im on-line Modell mit $(MW\sqrt{n})$-beschränktem linearem Abstand.*

Korollar 6.29 *B lernt $G_n(M,W)$ für $n \in \mathbb{N}$ und $\gamma > 0$ im on-line Modell mit $m_{M,W,\gamma}$-beschränktem Abstand bzgl. $L_{0,\gamma}$ für*

$$m_{M,W,\gamma} = \min\left\{ \left\lfloor \left(\frac{MW}{\gamma}\right)^2 \right\rfloor, n \right\}.$$

Beweis: Aus Korollar 6.27 und Lemma 4.6 folgt, daß sich B beim Lernen von $G_n(M,W)$ für $n \in \mathbb{N}$ im on-line Modell höchstens

$$\left\lfloor \left(\frac{MW}{\gamma}\right)^2 \right\rfloor$$

mal irren kann. Die Schranke n folgt aus Korollar 6.24. □

Wir schätzen nun den maximalen linearen Abstand D_1 einer Runde für B bei Zielfunktionen aus $G_n(M,W)$ ab.

Lemma 6.30 *B liefert in einer einzelnen Runde für jedes $f \in G_n(M,W)$ einen linearen Abstand von höchstens MW.*

Beweis: B spaltet jede Instanz in einen bekannten und einen unbekannten Anteil auf. Sei $\vec{v}$ der bekannte und $\vec{u}$ der unbekannte Anteil. Für $\vec{v}$ kann B den korrekten Funktionswert berechnen. Für $\vec{u}$ schätzt B als Funktionswert 0. Der lineare Abstand zwischen der Vorhersage 0 und dem Funktionswert für den unbekannten Anteil ist $|f(\vec{u})| = |\vec{\mu}\vec{u}|$. Dieser läßt sich mit der Cauchy-Schwartzschen Ungleichung abschätzen:

$$|\vec{\mu}\vec{u}| \leq \|\vec{\mu}\|_2 \|\vec{u}\|_2 \leq WM.$$

Wegen $f(\vec{x}) = f(\vec{v}) + f(\vec{u})$ ist somit der lineare Abstand zwischen der Vorhersage von B und dem tatsächlichen Funktionswert höchstens MW. □

7 Anwendung der Transformationen

Wir wenden nun die Transformationen aus Kapitel 5 auf die on-line Algorithmen aus Kapitel 6 an. So erhalten wir konstruktiv obere Schranken für die Längen der benötigten Proben. Diese Längen geben wir in O-Notation an, um sie später mit unteren Schranken zu vergleichen.

7.1 Funktionen mit beschränkter Steigung

Bei Anwendung der Transformationen auf den teilweise leicht besseren Algorithmus NN beim Lernen von $\mathcal{F}_\infty$ erhalten wir gegenüber Anwendung der Transformationen auf den Algorithmus $LININT$ keine Verbesserung in der O-Notation. Deshalb geben wir die Schranken für die Anwendung der Transformationen auf NN nicht extra an.

Bei Anwendung der Transformationen auf $LININT$ erhalten wir folgende Resultate:

Korollar 7.1 *Es gibt einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) mit einem (ϵ, γ)-guten Modell lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(\frac{1}{\gamma}\ln\frac{1}{\epsilon\gamma}+\ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Lemma 6.5 liefert für $t \geq 3$ die Schranke $2e + (e/(2\ln 2))\ln t < 7\ln t = l(t)$ für den linearen Gesamtabstand. Mit Korollar 5.6 2. folgt dann das Korollar. □

Korollar 7.2 *Sei $p > 1$ eine Konstante, dann gibt es einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) mit einem (ϵ, γ)-guten Modell lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(\frac{1}{\gamma^p}+\ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Lemma 6.3 liefert für $p = 1 + \eta$ die Schranke $l_p = 2 + 1/(2\eta\ln 2)$ für die Gesamtabweichung bzgl. L_p. Mit Korollar 5.6 1. folgt dann das Korollar. □

Korollar 7.3 *Es gibt einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) mit ϵ-beschränkter absoluter Differenz lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon\delta}\right)$$

auskommt.

Beweis: Das Korollar folgt aus Lemma 6.5 und den Korollaren 6.8 und 5.9. □

Korollar 7.4 *Sei $p > 1$ eine Konstante, dann gibt es einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon\delta}\right)$$

auskommt.

Beweis: Mit Lemma 6.3, Korollar 6.8 und Satz 5.7 erhalten wir das Korollar. □

Korollar 7.5 *Es gibt einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) für $p \geq 2$ mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon\delta}\right)$$

auskommt.

Beweis: Mit den Korollaren 6.2 und 6.8, sowie Satz 5.7 erhalten wir das Korollar. □

Korollar 7.6 *Es gibt einen Lernalgorithmus, der $\mathcal{F}_2$ (und somit auch $\mathcal{F}_\infty$) mit ϵ-beschränkter quadratischer Differenz lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon\delta}\right)$$

auskommt.

7.2 Lineare Funktionen

Da $B(EG_{M,W})$ beim on-line Lernen von $F_n(M,W)$ bei den hier betrachteten Abstandsfunktionen niemals schlechtere (oft jedoch bessere) Resultate[19] als $EG_{M,W}$ liefert, geben wir die Schranken für die Anwendung der Transformationen auf $EG_{M,W}$ nicht extra an.

Bei Anwendung der Transformationen auf $B(EG_{M,W})$ erhalten wir folgende Resultate:

[19] Zusätzlich ist die Anzahl der Runden mit falschen Vorhersagen (im Gegensatz zu $EG_{M,W}$) beschränkt, was ein Vorteil bei der Transformation $BLPAC$ aus Abschnitt 5.2 ist.

Korollar 7.7 *Für $n \in \mathbb{N}, n \geq 2$ gibt es einen Lernalgorithmus, der $F_n(M,W)$ mit einem (ϵ,γ)-guten Modell lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(\min\left\{\left(\frac{MW}{\gamma}\right)^2 \ln n, n\right\} + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Korollar 6.21 liefert die Schranke $m_{M,W,\gamma} = \min\left\{\left\lfloor \frac{1}{2}\left(\frac{MW}{\gamma}\right)^2 \ln n\right\rfloor, n\right\}$ für die maximale Anzahl der Irrtümer. Das Korollar folgt dann mit Satz 5.2. □

Korollar 7.8 *Für $n \in \mathbb{N}, n \geq 2$ und $p > 0$ gibt es einen Lernalgorithmus, der $F_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{(MW)^p}{\epsilon}\left(n^{1-\min\{p/2,1\}}\left(\frac{\ln n}{2}\right)^{\min\{p/2,1\}} + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Korollar 6.18 liefert die Schranke $(MW)^p n^{1-\min\{p/2,1\}}\left(\frac{\ln n}{2}\right)^{\min\{p/2,1\}}$ für den Gesamtabstand bzgl. L_p. Lemma 6.23 liefert $D_1 = MW$, woraus $D = (MW)^p$ folgt. Wegen Korollar 6.15 ist die Anzahl der verschiedenen Hypothesen, die $B(EG_{M,W})$ liefert durch $n+1$ beschränkt. Somit folgt nun aus Korollar 5.8 das Korollar. □

Korollar 7.9 *Für $n \in \mathbb{N}, n \geq 2$ gibt es einen Lernalgorithmus, der $F_n(M,W)$ mit ϵ-beschränkter quadratischer Differenz lernt und mit einer Probe der Länge*

$$O\left(\frac{(MW)^2}{\epsilon}\ln\frac{n}{\delta}\right)$$

auskommt.

Korollar 7.10 *Für $n \in \mathbb{N}, n \geq 2$ gibt es einen Lernalgorithmus, der $F_n(M,W)$ mit ϵ-beschränkter absoluter Differenz lernt und mit einer Probe der Länge*

$$O\left(\frac{MW}{\epsilon}\left(\sqrt{n\ln n} + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Korollar 7.11 *Für $n \in \mathbb{N}, n \geq 2$ gibt es einen Lernalgorithmus, der $F_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_0 lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(n + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Wegen Korollar 6.15 ist die Anzahl der verschiedenen Hypothesen von $B(EG_{M,W})$ durch $n+1$ und der Gesamtabstand (bzgl. L_0) durch n beschränkt. Somit folgt das Korollar aus Satz 5.7. □

Für $G_n(M,W)$ wenden wir die Transformationen auf den on-line Algorithmus B an und erhalten folgende Resultate:

Korollar 7.12 *Für $n \in \mathbb{N}$ gibt es einen Lernalgorithmus, der $G_n(M,W)$ mit einem (ϵ,γ)-guten Modell lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \min\left\{\left(\frac{MW}{\gamma}\right)^2, n\right\}\right)\right)$$

auskommt.

Beweis: Korollar 6.29 liefert die Schranke $m_{M,W,\gamma} = \min\{\lfloor (MW/\gamma)^2 \rfloor, n\}$ für die maximale Anzahl der Irrtümer. Das Korollar folgt dann mit Satz 5.2. □

Korollar 7.13 *Für $n \in \mathbb{N}$ und $p > 0$ gibt es einen Lernalgorithmus, der $G_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{(MW)^p}{\epsilon}\left(n^{\frac{\max\{p,2\}-p}{2}} + \ln\min\left\{\frac{n}{\delta}, \frac{(MW)^p}{\epsilon\delta}\right\}\right)\right)$$

auskommt.

Beweis: Korollar 6.26 liefert die Schranke $(MW)^p n^{(\max\{p,2\}-p)/2}$ für den Gesamtabstand bzgl. L_p. Lemma 6.30 liefert $D_1 = MW$, woraus $D = (MW)^p$ folgt. Wegen Korollar 6.24 ist die Anzahl der verschiedenen Hypothesen, die B liefert durch $n+1$ beschränkt. Somit folgt nun aus Korollar 5.8 das Korollar. □

Korollar 7.14 *Für $n \in \mathbb{N}$ und $0 < p < 2$ gibt es einen Lernalgorithmus, der $G_n(M, W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{(MW)^p}{\epsilon}\left(n^{1-\frac{p}{2}} + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Korollar 7.15 *Für $n \in \mathbb{N}$ und $p \geq 2$ gibt es einen Lernalgorithmus, der $G_n(M, W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt und mit einer Probe der Länge*

$$O\left(\frac{(MW)^p}{\epsilon}\ln\min\left\{\frac{n}{\delta}, \frac{(MW)^p}{\epsilon\delta}\right\}\right)$$

auskommt.

Korollar 7.16 *Für $n \in \mathbb{N}$ gibt es einen Lernalgorithmus, der $G_n(M, W)$ mit ϵ-beschränkter Differenz bzgl. L_0 lernt und mit einer Probe der Länge*

$$O\left(\frac{1}{\epsilon}\left(n + \ln\frac{1}{\delta}\right)\right)$$

auskommt.

Beweis: Wegen Korollar 6.24 ist die Anzahl der verschiedenen Hypothesen von B durch $n + 1$ und der Gesamtabstand (bzgl. L_0) durch n beschränkt. Somit folgt das Korollar aus Satz 5.7. □

8 Verschiedene Dimensionen

Zur Berechnung von oberen und unteren Schranken beim Lernen, sowie zur Entscheidung der Lernbarkeit in verschiedenen Lernmodellen, wurde eine Reihe von Dimensionen benutzt, die Erweiterungen der Vapnik-Chervonenkis Dimension sind. Hier sollen nun die Pseudo-Dimension [Hau92], die „fat-shattering function“ [KS94, AB94] und eine der Natarajan Dimension ähnliche Dimension [Sim93] vorgestellt werden. Weitere Dimensionen sind z. B. die Band-Dimension [ABIST94], die Graph-Dimension [Nat89] und die oben bereits angesprochene Natarajan Dimension [Nat89].

Die *Vapnik-Chervonenkis Dimension* von C (kurz $VCdim(C)$), ist für eine Klasse C von Funktionen $f : X \to \{0,1\}$ wie folgt definiert:

Eine Folge

$$(x_1, \ldots, x_d) \in X^d$$

ist durch C aufspaltbar, wenn

$$\forall b \in \{0,1\}^d : \exists f = f_b \in C : \forall i \in \{1, \ldots, d\} : f(x_i) = b_i.$$

$VCdim(C)$ bezeichne dann das Supremum aller d, so daß eine durch C aufspaltbare Folge der Länge d existiert.

Die *Pseudo-Dimension* (oder auch *Pollard Dimension* oder *kombinatorische Dimension* genannt) von F (kurz $Pdim(F)$) ist wie folgt definiert:

Eine Folge

$$S = ((x_1, r_1), \ldots, (x_d, r_d)) \in (X \times \mathbb{R})^d$$

ist durch F aufspaltbar, wenn

$$\forall b \in \{0,1\}^d : \exists f = f_b \in F : \forall i \in \{1, \ldots, d\} : f(x_i) \begin{cases} \leq r_i & \text{falls } b_i = 0 \\ > r_i & \text{falls } b_i = 1 \end{cases}$$

$Pdim(F)$ bezeichne dann das Supremum aller d, so daß eine durch F aufspaltbare Folge der Länge d existiert.

Falls F eine Klasse von Funktionen $f : X \to \{0,1\}$ ist, so ist offensichtlich $Pdim(F) = VCdim(F)$.

Die „*fat-shattering function*“ von F bzgl. γ (kurz $fat_F(\gamma)$) wurde von Kearns und Schapire [KS94] eingeführt und ist wie folgt definiert:

Eine Folge

$$S = ((x_1, r_1), \ldots, (x_d, r_d)) \in (X \times \mathbb{R})^d$$

ist γ-fat-aufspaltbar durch F, wenn

$$\forall b \in \{0,1\}^d : \exists f = f_b \in F : \forall i \in \{1, \ldots, d\} : f(x_i) \begin{cases} \leq r_i - \gamma & \text{falls } b_i = 0 \\ \geq r_i + \gamma & \text{falls } b_i = 1 \end{cases}$$

$fat_F(\gamma)$ bezeichne dann das Supremum aller d, so daß eine Folge der Länge d existiert, die γ-fat-aufspaltbar durch F ist.

Falls $\gamma \leq 1/2$ und F eine Klasse von Funktionen $f : X \to \{0,1\}$ ist, so ist offensichtlich $fat_F(\gamma) = VCdim(F)$.

Simon [Sim93] hat eine, der Natarajan Dimension ähnliche, Dimension eingeführt, die eine „Verschärfung“ der „fat-shattering function“ ist.

Eine Folge

$$S = ((x_1, r_1, s_1), \ldots, (x_d, r_d, s_d)) \in (X \times \mathbb{R} \times \mathbb{R})^d$$

ist durch F streng-γ-fat-aufspaltbar, wenn

1. $\forall b \in \{0,1\}^d : \exists f = f_b \in F : \forall i \in \{1, \ldots, d\} : f(x_i) = \begin{cases} r_i & \text{falls } b_i = 0 \\ s_i & \text{falls } b_i = 1 \end{cases}$
2. $\forall i \in \{1, \ldots, d\} : s_i \geq 2\gamma + r_i$.

$d_F(\gamma)$ bezeichne dann das Supremum aller d, so daß eine durch F streng-γ-fat-aufspaltbare Folge der Länge d existiert.

Falls $\gamma \leq 1/2$ und F eine Klasse von Funktionen $f : X \to \{0,1\}$ ist, so ist offensichtlich $d_F(\gamma) = VCdim(F)$.

Wir zeigen nun einfache Beziehungen zwischen den vorgestellten Dimensionen, sowie einem bzgl. der Abstandsfunktion $L_{0,\gamma}$ optimalen on-line Lernalgorithmus.

Lemma 8.1 *Für $\gamma > 0$ gilt:*

1. $d_F(\gamma) \leq fat_F(\gamma)$,
2. $fat_F(\gamma) \leq Pdim(F)$ *und*
3. $d_F(\gamma) \leq opt_{L_{0,\gamma}}(F)$.

Beweis:

1. Es reicht zu zeigen, daß zu jeder durch F streng-γ-fat-aufspaltbaren Folge S auch eine durch F γ-fat-aufspaltbare Folge S' gleicher Länge existiert. Sei nun $S = ((x_1, r_1, s_1), \ldots, (x_d, r_d, s_d))$ eine durch F streng-γ-fat-aufspaltbaren Folge, dann setze $S' = ((x_1, (r_1 + s_1)/2), \ldots, (x_d, (r_d + s_d)/2))$. Für alle $1 \leq i \leq d$ gilt wegen $s_i \geq 2\gamma + r_i$ offensichtlich $r_i \leq (r_i + s_i)/2) - \gamma$ und $s_i \geq (r_i + s_i)/2) + \gamma$.

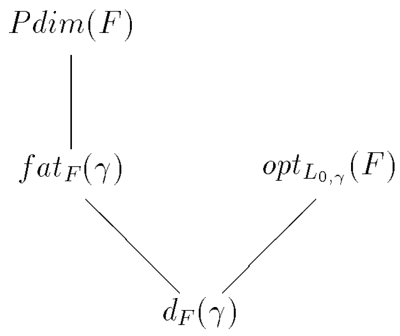

Abbildung 1: Die Hierarchie einiger Dimensionen und $opt_{L_{0,\gamma}}$.

2. Offensichtlich ist jede Folge S, die γ-fat-aufspaltbar durch F ist, auch durch F aufspaltbar.

3. Es reicht zu zeigen, daß zu jeder durch F streng-γ-fat-aufspaltbaren Folge S, der Länge d, eine Gegenspielerstrategie existiert, die bei jedem Algorithmus einen Gesamtabstand (bzgl. $L_{0,\gamma}$) von mindestens d verursacht. Sei nun $S = ((x_1, r_1, s_1), \ldots, (x_d, r_d, s_d))$. Für $1 \leq i \leq d$ ist die Instanz in der i-ten Runde x_i. Die Vorhersage von A in der i-ten Runde sei $\hat{y}_i$. Falls $\hat{y}_i < (r_i + s_i)/2$, so antwortet der Gegenspieler mit $y_i = s_i$, ansonsten mit $y_i = r_i$. Somit ist $|\hat{y}_i - y_i| \geq \gamma$ und damit $L_{0,\gamma}(\hat{y}_i, y_i) = 1$ für jedes $1 \leq i \leq d$. Also ist der Gesamtabstand (bzgl. $L_{0,\gamma}$) nach d Runden bereits d.

□

In Abbildung 1 ist die aus Lemma 8.1 resultierende Hierarchie dargestellt.

Wir können also mit Lemma 8.1 $d_F(\gamma)$ durch die Pseudo-Dimension und durch online Resultate bzgl. der Abstandsfunktion $L_{0,\gamma}$ nach oben abschätzen. Falls F ein n-dimensionaler Vektorraum von reellwertigen Funktionen ist, läßt sich mit dem folgenden Satz von Dudley [Dud78] (s. auch [Hau92]) die Pseudo-Dimension bestimmen.

Satz 8.2 ([Dud78]) *Sei F ein d-dimensionaler Vektorraum von Funktionen $f : X \to \mathbb{R}$, dann ist $Pdim(F) = d$.*

Wir schätzen nun $d_F(\gamma)$ für die hier benutzten Funktionenklassen ab. Untere Schranken für $d_F(\gamma)$ erhalten wir durch Folgen S, die durch F streng-γ-fat-aufgespalten werden.

Obere Schranken erhalten wir durch eine Konstruktion von Simon [Sim93], durch Satz 8.2 und durch die on-line Resultate aus Kapitel 6.

Lemma 8.3 $d_{\mathcal{F}_1}(\gamma) \geq d_{\mathcal{F}_2}(\gamma) \geq d_{\mathcal{F}_\infty}(\gamma) \geq \lfloor(1+1/\gamma)/2\rfloor$.

Beweis: Die Folge

$$S = ((\gamma, -\gamma, \gamma), (3\gamma, -\gamma, \gamma), (5\gamma, -\gamma, \gamma), \ldots, ((2\lfloor(1+1/\gamma)/2\rfloor - 1)\gamma, -\gamma, \gamma))$$

hat offensichtlich die Länge $\lfloor(1+1/\gamma)/2\rfloor$ und es gilt $(2\lfloor(1+1/\gamma)/2\rfloor - 1)\gamma \leq 1$.

Wir zeigen nun, daß S durch $\mathcal{F}_\infty$ streng-γ-fat-aufspaltbar ist. Die 2. Bedingung ist offensichtlich erfüllt und die 1. Bedingung läßt sich wie folgt zeigen.

Wir zeigen, daß für jedes b (vgl. Definition von $d_F(\gamma)$) der Betrag der Steigung von f_b zwischen 0 und γ, sowie zwischen $(2i-1)\gamma$ und $(2i+1)\gamma$ für $1 \leq i < \lfloor(1+1/\gamma)/2\rfloor$, durch 1 beschränkt ist. Falls $b_1 = 0$, dann setzen wir zwischen 0 und γ die Steigung auf -1, so daß wir von $f(0) = 0$ nach $f(\gamma) = -\gamma$ kommen. Entsprechend setzen wir für $b_1 = 1$ die Steigung zwischen 0 und γ auf 1. Für $1 \leq i < \lfloor(1+1/\gamma)/2\rfloor$ kommen wir von r_i nach r_{i+1} ($b_i = b_{i+1} = 0$), sowie von s_i nach s_{i+1} ($b_i = b_{i+1} = 1$) mit Steigung 0, von r_i nach s_{i+1} ($b_i = 0$ und $b_{i+1} = 1$) mit Steigung 1 und von s_i nach r_{i+1} ($b_i = 1$ und $b_{i+1} = 0$) mit Steigung -1.

Da $\mathcal{F}_\infty \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_1$, ist $d_{\mathcal{F}_\infty}(\gamma) \leq d_{\mathcal{F}_2}(\gamma) \leq d_{\mathcal{F}_1}(\gamma)$. □

Lemma 8.4 $d_{\mathcal{F}_1}(\gamma) \leq \lfloor(1+1/\gamma)/2\rfloor$.

Beweis: s. Anhang A.2 □

Korollar 8.5 $d_{\mathcal{F}_1}(\gamma) = d_{\mathcal{F}_2}(\gamma) = d_{\mathcal{F}_\infty}(\gamma) = \lfloor(1+1/\gamma)/2\rfloor$.

Lemma 8.6 *Für* $n \in \mathbb{N}, n \geq 2$ *ist* $d_{F_n(M,W)}(\gamma) \geq \min\{\lfloor MW/(2\gamma)\rfloor, n-1\}$.

Beweis: Für $\gamma > MW/2$ ist die Aussage des Lemmas trivial. Somit nehmen wir an, daß $0 < \gamma \leq MW/2$ gilt. Die Funktionen $f \in F_n(M,W)$ werden durch ihren Koeffizientenvektor $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ dargestellt.

Sei $k = \min\{\lfloor MW/(2\gamma)\rfloor, n-1\}$, dann setze

$$S = ((M\vec{e}_1, 0, 2\gamma), \ldots, (M\vec{e}_k, 0, 2\gamma)).$$

Wir zeigen nun, daß S durch $F_n(M,W)$ streng-γ-fat-aufspaltbar ist. Die 2. Bedingung ist offensichtlich erfüllt und die 1. Bedingung läßt sich wie folgt zeigen.

Zu jedem $b \in \{0,1\}^k$ gibt es ein $\vec{\mu}_b = (\mu_{b,1}, \ldots, \mu_{b,n})$, dabei ist $\mu_{b,i} = 2\gamma b_i/M$ für $1 \leq i \leq k$, $\mu_{b,i} = 0$ für $k < i < n$ und $\mu_{b,n} = W - \sum_{i=1}^{k} \mu_{b,i}$. Es gilt $\sum_{i=1}^{n-1} \mu_{b,i} = \sum_{i=1}^{k} \mu_{b,i} \leq 2\gamma k/M \leq 2\gamma \lfloor MW/(2\gamma) \rfloor / M \leq W$. Somit ist $\mu_{b,n} \geq 0$. □

$F'_n(W)$ sei die Klasse aller linearen Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ mit $f(\vec{\mu}) = \vec{\mu}\vec{x}$, wobei $\vec{\mu} \in [0,W]^n$ und $\|\vec{\mu}\|_1 = W$. Also ist $F_n(M,W) \subset F'_n(W)$ für $n \in \mathbb{N}, n \geq 2$, $W > 0$ und $M > 0$ und damit offensichtlich $Pdim(F_n(M,W)) \leq Pdim(F'_n(W))$.

Lemma 8.7 *Für $n \in \mathbb{N}, n \geq 2$ ist $F'_n(W)$ ein $(n-1)$-dimensionaler Vektorraum von Funktionen.*

Beweis: s. Anhang A.3. □

Korollar 8.8 *Für $n \in \mathbb{N}, n \geq 2$ ist $Pdim(F_n(M,W)) \leq n-1$.*

Korollar 8.9 *Für $n \in \mathbb{N}, n \geq 2$ ist*

$$\min\left\{\left\lfloor \frac{MW}{2\gamma} \right\rfloor, n-1\right\} \leq d_{F_n(M,W)}(\gamma) \leq \min\left\{\left\lfloor \frac{(MW)^2 \ln n}{2\gamma^2} \right\rfloor, n-1\right\}.$$

D. h. für $n-1 \leq \lfloor MW/(2\gamma) \rfloor$ ist

$$d_{F_n(M,W)}(\gamma) = n-1$$

und für $n-1 > \lfloor MW/(2\gamma) \rfloor$ ist

$$\left\lfloor \frac{MW}{2\gamma} \right\rfloor \leq d_{F_n(M,W)}(\gamma) \leq \min\left\{\left\lfloor \frac{(MW)^2 \ln n}{2\gamma^2} \right\rfloor, n-1\right\}.$$

Beweis: Die Abschätzung für $d_{F_n(M,W)}(\gamma)$ nach unten stammt aus Lemma 8.6. Die Abschätzung nach oben folgt aus den Korollaren 6.13 und 8.8, sowie dem Lemma 8.1. □

Lemma 8.10 *Für $n \in \mathbb{N}$ ist $d_{G_n(M,W)}(\gamma) \geq \min\{\lfloor (MW/\gamma)^2 \rfloor, n\}$.*

Beweis: Für $\gamma > MW$ ist die Aussage des Lemmas trivial. Somit nehmen wir an, daß $0 < \gamma \leq MW$ gilt. Die Funktionen $f \in G_n(M,W)$ werden durch ihren Koeffizientenvektor $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ dargestellt.

Sei $k = \min\{\lfloor (MW/\gamma)^2 \rfloor, n\}$, dann setze

$$S = ((M\vec{e}_1, -\gamma, \gamma), \ldots, (M\vec{e}_k, -\gamma, \gamma)).$$

Wir zeigen nun, daß S durch $G_n(M,W)$ streng-γ-fat-aufspaltbar ist. Die 2. Bedingung ist offensichtlich erfüllt und die 1. Bedingung läßt sich wie folgt zeigen.

Zu jedem $b \in \{0,1\}^k$ gibt es ein $\vec{\mu}_b = (\mu_{b,1}, \ldots, \mu_{b,n})$, dabei ist $\mu_{b,i} = (2b_i - 1)\gamma/M$ für $1 \leq i \leq k$ und $\mu_{b,i} = 0$ für $k < i \leq n$. Es gilt $\|\vec{\mu}_b\|_2 = \sqrt{\sum_{i=1}^{k} (\pm\gamma/M)^2} = \sqrt{k}\gamma/M \leq \sqrt{\lfloor (MW/\gamma)^2 \rfloor}(\gamma/M) \leq W$. □

Korollar 8.11 *Für $n \in \mathbb{N}$ ist $d_{G_n(M,W)}(\gamma) = \min\{\lfloor (MW/\gamma)^2 \rfloor, n\}$.*

Beweis: Wegen Korollar 6.29 ist $opt_{L_{0,\gamma}}(G_n(M,W)) \leq \min\{\lfloor (MW/\gamma)^2 \rfloor, n\}$ für jedes $n \in \mathbb{N}$. Somit folgt das Korollar nun aus den Lemmata 8.1 und 8.10. □

9 Untere Schranken in den verteilungsfreien off-line Modellen

Zur Bestimmung unterer Schranken benutzen wir die Dimension $d_F(\gamma)$ (s. Kapitel 8) und das Maß $\Delta(F)$ aus [Sim93], welches wir nun definieren.

Zwei Funktionen f, g auf dem Grundbereich X heißen *disjunkt*, wenn $f(x) \neq g(x)$ für alle $x \in X$. Eine Funktionenklasse F heißt *trivial*, falls ihre Funktionen paarweise disjunkt sind. Da für triviale Funktionenklassen ein Beispiel zum Lernen ausreicht, betrachten wir nur nicht triviale Funktionenklassen. Wir führen nun folgendes Maß ein:

$$\Delta(F) = \sup \left\{ |g(x) - f(x)| \;\middle|\; \begin{array}{l} x \in X \text{ und } f,g \text{ sind nicht disjunkte} \\ \text{Funktionen aus } F \end{array} \right\}.$$

Es liefert den maximalen Abstand, den zwei nicht disjunkte Funktionen aus einer Klasse F an einem Punkt haben können.

Lemma 9.1 $\Delta(\mathcal{F}_1) = \Delta(\mathcal{F}_2) = \Delta(\mathcal{F}_\infty) = 2.$

Beweis: Da offensichtlich $-1 \leq f(x) \leq 1$ für alle $f \in \mathcal{F}_1$ und $x \in [0,1]$ gilt, ist $\Delta(\mathcal{F}_1) \leq 2$. Wegen $\mathcal{F}_\infty \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_1$ ist dann $\Delta(\mathcal{F}_\infty) \leq \Delta(\mathcal{F}_2) \leq \Delta(\mathcal{F}_1) \leq 2$. Da $\Delta(\mathcal{F}_\infty) \geq 2$ von den Funktionen $f_1(x) = x$ und $f_2(x) = -x$ an der Stelle $x = 1$ bezeugt wird, folgt nun das Lemma. $\square$

Lemma 9.2 *Für $n \in \mathbb{N}, n \geq 2$ ist $\Delta(F_n(M,W)) = MW$.*

Beweis: Es gilt $0 \leq \vec{\mu}\vec{x} = \sum_{i=1}^n w_i x_i \leq M \sum_{i=1}^n w_i = MW$, also $\Delta(F_n(M,W)) \leq MW$. $\Delta(F_n(M,W)) \geq MW$ wird von den Funktionen $f_1(\vec{x}) = (W\vec{e}_1)\vec{x}$ und $f_2(\vec{x}) = (W\vec{e}_2)\vec{x}$ an der Stelle $\vec{x} = M\vec{e}_1$ bezeugt. $\square$

Lemma 9.3 *Für $n \in \mathbb{N}$ ist $\Delta(G_n(M,W)) = 2MW$.*

Beweis: Mit der Cauchy-Schwartzschen Ungleichung folgt $|\vec{\mu}\vec{x}| \leq \|\vec{\mu}\|_2 \|\vec{x}\|_2 \leq MW$. Somit gilt $\Delta(G_n(M,W)) \leq 2MW$. $\Delta(G_n(M,W)) \geq 2MW$ wird von den Funktionen $f_1(\vec{x}) = (W\vec{e}_1)\vec{x}$ und $f_2(\vec{x}) = (-W\vec{e}_1)\vec{x}$ an der Stelle $\vec{x} = M\vec{e}_1$ bezeugt. $\square$

Simon [Sim93] hat abhängig von $d_F(\gamma)$ und $\Delta(F)$ folgende allgemeine untere Schranken für die Anzahl der benötigten Beispiele beim Lernen mit einem (ϵ,γ)-guten Modell gezeigt.

Satz 9.4 ([Sim93]) *Sei A ein Algorithmus, der die Funktionenklasse F mit einem (ϵ,γ)-guten Modell lernt.*

1. *Falls F nicht trivial, $\epsilon \leq 1/2$ und $\gamma \leq \Delta(F)/2$, dann benötigt A*

$$\Omega\left(\frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

   Beispiele.

2. *Falls $\epsilon \leq 1/8$ und $\delta \leq 1/100$, dann benötigt A*

$$\Omega\left(\frac{d_F(\gamma) - 1}{\epsilon}\right)$$

   Beispiele.

Simon [Sim93] benutzt für 1. die schärfere Voraussetzung „$\epsilon < 1/2$ und $\gamma < \Delta(F)/2$“ statt „$\epsilon \leq 1/2$ und $\gamma \leq \Delta(F)/2$“. Der Satz gilt jedoch auch mit der hier angegebenen schwächeren Voraussetzung[20].

Wir wenden obigen Satz nun auf die hier betrachteten Funktionenklassen an und erhalten die folgenden unteren Schranken beim Lernen mit einem (ϵ, γ)-guten Modell.

Korollar 9.5 *Für $\epsilon \leq 1/8$, $\gamma \leq 1$ und $\delta \leq 1/100$ benötigt jeder Algorithmus, der $\mathcal{F}_2$ oder $\mathcal{F}_\infty$ mit einem (ϵ, γ)-guten Modell lernt,*

$$\Omega\left(\frac{1}{\epsilon}\left(\frac{1}{\gamma} + \ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Korollar 9.6 *Für $n \in \mathbb{N}, n \geq 2$, $\epsilon \leq 1/8$, $\gamma \leq MW/2$ und $\delta \leq 1/100$ benötigt jeder Algorithmus, der $F_n(M, W)$ mit einem (ϵ, γ)-guten Modell lernt,*

$$\Omega\left(\frac{1}{\epsilon}\left(\min\left\{\frac{MW}{\gamma}, n\right\} + \ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Korollar 9.7 *Für $n \in \mathbb{N}$, $\epsilon \leq 1/8$, $\gamma \leq MW$ und $\delta \leq 1/100$ benötigt jeder Algorithmus, der $G_n(M, W)$ mit einem (ϵ, γ)-guten Modell lernt,*

$$\Omega\left(\frac{1}{\epsilon}\left(\min\left\{\left(\frac{MW}{\gamma}\right)^2, n\right\} + \ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Mit Hilfe von Lemma 5.13 erhalten wir aus den allgemeinen unteren Schranken für die Anzahl der benötigten Beispiele beim Lernen mit einem (ϵ, γ)-guten Modell aus Satz 9.4 auch allgemeine untere Schranken für die Anzahl der benötigten Beispiele beim Lernen mit ϵ-beschränkter Differenz.

[20] $\gamma \leq \Delta(F)/2$ ist bei Betrachtung des Beweises sofort offensichtlich. $\epsilon \leq 1/2$ folgt aus der Tatsache, daß hier (und in [Sim93]) ein „schärferes“ Modell als in der im Beweis angeführten Literatur ([BEHW89]; dort wird ein Fehler von genau ϵ noch akzeptiert) benutzt wird.

Korollar 9.8 *1. Sei F eine nicht triviale Funktionenklasse, $p > 0$ und $\epsilon \leq \Delta^p(F)/2^{p+1}$. Jeder Algorithmus, der F mit ϵ-beschränkter Differenz bzgl. L_p lernt benötigt*

$$\Omega\left(\frac{\Delta^p(F)}{2^p\epsilon}\ln\frac{1}{\delta}\right)$$

Beispiele.

2. Sei $p > 0$, $\delta \leq 1/100$, $\epsilon \leq \gamma^p/8$. Jeder Algorithmus, der F mit ϵ-beschränkter Differenz bzgl. L_p lernt benötigt

$$\Omega\left(\frac{\gamma^p(d_F(\gamma)-1)}{\epsilon}\right)$$

Beispiele.

Beweis:

1. Sei A ein Algorithmus, der eine nicht triviale Funktionenklasse F für $p > 0$ und $\epsilon \leq \Delta^p(F)/2^{p+1}$ mit ϵ-beschränkter Differenz bzgl. L_p lernt. Setze $\gamma = \Delta(F)/2$. Somit gilt offensichtlich $\epsilon/\gamma^p = 2^p\epsilon/\Delta^p(F) \leq 1/2$. Wegen Lemma 5.13 liefert der Algorithmus dann mit einer Wahrscheinlichkeit von mindestens $1-\delta$ eine Hypothese, die ein $(\epsilon/\gamma^p, \gamma)$-gutes Modell für die Zielfunktion ist, so daß sich die untere Schranke von Satz 9.4 1. auf A überträgt.

2. Sei A ein Algorithmus, der eine Funktionenklasse F für ein $p > 0$, $\delta \leq 1/100$ und $\epsilon \leq \gamma^p/8$ mit ϵ-beschränkter Differenz bzgl. L_p lernt. Wegen Lemma 5.13 liefert der Algorithmus dann mit einer Wahrscheinlichkeit von mindestens $1-\delta$ eine Hypothese, die ein $(\epsilon/\gamma^p, \gamma)$-gutes Modell für die Zielfunktion ist, so daß sich die untere Schranke von Satz 9.4 2. auf A überträgt.

□

Wir wenden nun die allgemeinen Schranken auf die hier betrachteten Funktionenklassen an und erhalten dabei die folgenden unteren Schranken beim Lernen mit ϵ-beschränkter Differenz.

Korollar 9.9 *Für $p > 0$ und $\epsilon \leq 1/2$ benötigt jeder Algorithmus, der $\mathcal{F}_2$ oder $\mathcal{F}_\infty$ mit ϵ-beschränkter Differenz bzgl. L_p lernt,*

$$\Omega\left(\frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

Beispiele.

Korollar 9.10 *Für $n \in \mathbb{N}, n \geq 2$, $p > 0$ und $\epsilon \leq (MW)^p/2^{p+1}$ benötigt jeder Algorithmus, der $F_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt,*

$$\Omega\left(\frac{(MW)^p}{2^p\epsilon}\ln\frac{1}{\delta}\right)$$

Beispiele.

Korollar 9.11 *Für $n \in \mathbb{N}, n \geq 2$ (falls n eine Konstante ist, dann $n \geq 3$), $0 < p < 1$, $\epsilon \leq (MW/(2n))^p/8$ und $\delta \leq 1/100$ benötigt jeder Algorithmus, der $F_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt,*

$$\Omega\left(\frac{(MW)^p}{2^p\epsilon}\left(n^{1-p}+\ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Beweis: Für $\gamma = MW/(2n)$ liefert Korollar 8.9 $d_{F_n(M,W)}(\gamma) = n-1$. Korollar 9.8 2. liefert dann für $n \geq 2$ (bzw. $n \geq 3$, falls n eine Konstante ist) die Schranke $\Omega((MW/(2n))^p n/\epsilon)$. Da offensichtlich $\epsilon \leq (MW)^p/2^{p+1}$ gilt, ist wegen Korollar 9.10 auch $\Omega(((MW)^p/(2^p\epsilon))\ln(1/\delta))$ eine untere Schranke. □

Korollar 9.12 *Sei $n \in \mathbb{N}, n \geq 2$, $\epsilon \leq 1/8$ und $\delta \leq 1/100$. Jeder Algorithmus, der $F_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_0 lernt, benötigt*

$$\Omega\left(\frac{1}{\epsilon}\left(n+\ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Beweis: Für jedes $\gamma > 0$ und $y, \hat{y} \in Y$ gilt $L_0(\hat{y},y) \geq L_{0,\gamma}(\hat{y},y)$. Mit $\gamma = MW/n$, Lemma 4.8 und Korollar 9.6 folgt dann das Korollar. □

Korollar 9.13 *Für $n \in \mathbb{N}$, $p > 0$ und $\epsilon \leq (MW)^p/2$ benötigt jeder Algorithmus, der $G_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt,*

$$\Omega\left(\frac{(MW)^p}{\epsilon}\ln\frac{1}{\delta}\right)$$

Beispiele.

Korollar 9.14 *Sei $n \in \mathbb{N}$ (falls n eine Konstante ist, dann $n \geq 2$), $0 < p < 2$, $\epsilon \leq (MW)^p/(8n^{p/2})$ und $\delta \leq 1/100$. Jeder Algorithmus, der $G_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_p lernt, benötigt*

$$\Omega\left(\frac{(MW)^p}{\epsilon}\left(\ln\frac{1}{\delta} + n^{1-\frac{p}{2}}\right)\right)$$

Beispiele.

Beweis: Für $\gamma = MW/\sqrt{n}$ liefert Korollar 8.11 $d_{G_n(M,W)}(\gamma) = n$. Korollar 9.8 2. liefert dann (falls n eine Konstante ist: für $n \geq 2$) die Schranke $\Omega((MW/\sqrt{n})^p n/\epsilon) = \Omega((MW)^p n^{1-p/2}/\epsilon)$. Da offensichtlich $\epsilon \leq (MW)^p/2$ gilt, ist wegen Korollar 9.13 auch $\Omega(((MW)^p/\epsilon)\ln(1/\delta))$ eine untere Schranke. □

Korollar 9.15 *Sei $n \in \mathbb{N}$, $\epsilon \leq 1/8$ und $\delta \leq 1/100$. Jeder Algorithmus, der $G_n(M,W)$ mit ϵ-beschränkter Differenz bzgl. L_0 lernt, benötigt*

$$\Omega\left(\frac{1}{\epsilon}\left(n + \ln\frac{1}{\delta}\right)\right)$$

Beispiele.

Beweis: Für jedes $\gamma > 0$ und $y, \hat{y} \in Y$ gilt $L_0(\hat{y}, y) \geq L_{0,\gamma}(\hat{y}, y)$. Mit $\gamma = MW/\sqrt{n}$, Lemma 4.8 und Korollar 9.7 folgt dann das Korollar. □

10 Vergleich zwischen unteren Schranken und Ergebnissen aus den Transformationen

Vergleichen wir die Probenlängen, die wir bei Anwendung der Transformationen auf Funktionen mit beschränkter Steigung, sowie lineare Funktionen erhalten, mit den entsprechenden unteren Schranken, so sehen wir, daß unsere Probenlängen meistens bis auf einen logarithmischen oder gar konstanten Faktor optimal sind.

Bemerkung 10.1 *Mit der Transformation MBPAC lassen sich $\mathcal{F}_2$ und $\mathcal{F}_\infty$ mit einem (ϵ,γ)-guten Modell bis auf einen logarithmischen Faktor optimal lernen (für $\epsilon \leq 1/8$, $\gamma \leq 1$ und $\delta \leq 1/100$).*

	$\mathcal{F}_2$, $\mathcal{F}_\infty$	$F_n(M,W)$	$G_n(M,W)$
(ϵ,γ)-gutes Modell	log Faktor	$n =$ konst. $\Rightarrow$ konst. Faktor	konst. Faktor
ϵ-beschränkte Differenz bzgl. L_p	$p \geq 1 \wedge p =$ konst. $\Rightarrow$ log Faktor $p \geq 2 \Rightarrow$ log Faktor	$p \geq 2 \wedge p =$ konst. $\Rightarrow$ log Faktor	$0 < p < 2 \Rightarrow$ konst. Faktor $p \geq 2 \Rightarrow$ log Faktor
ϵ-beschränkte Differenz bzgl. L_0	—	konst. Faktor	konst. Faktor

Tabelle 3: Zusammenfassung des Vergleichs zwischen oberen und unteren Schranken.

Bemerkung 10.2 *Mit der Transformation MBPAC läßt sich $F_n(M,W)$ für* konstantes $n \in \mathbb{N}, n \geq 2$ *mit einem (ϵ,γ)-guten Modell bis auf einen konstanten Faktor optimal lernen (für $\epsilon \leq 1/8$, $\gamma \leq MW/2$ und $\delta \leq 1/100$).*

Bemerkung 10.3 *Mit der Transformation MBPAC läßt sich $G_n(M,W)$ für $n \in \mathbb{N}$ mit einem (ϵ,γ)-guten Modell bis auf einen konstanten Faktor optimal lernen (für $\epsilon \leq 1/8$, $\gamma \leq MW$ und $\delta \leq 1/100$).*

Bemerkung 10.4 *Mit der Transformation BLPAC lassen sich $\mathcal{F}_2$ und $\mathcal{F}_\infty$ für* konstantes $p \geq 1$ *mit ϵ-beschränkter Differenz bzgl. L_p (also insbesondere absolute und quadratische Differenz) bis auf einen logarithmischen Faktor optimal lernen (für $\epsilon \leq 1/2$).*

Bemerkung 10.5 *Mit der Transformation BLPAC lassen sich $\mathcal{F}_2$ und $\mathcal{F}_\infty$ für $p \geq 2$ mit ϵ-beschränkter Differenz bzgl. L_p bis auf einen logarithmischen Faktor optimal lernen (für $\epsilon \leq 1/2$).*

Bemerkung 10.6 *Mit der Transformation BLPAC läßt sich $F_n(M,W)$ für $n \in \mathbb{N}, n \geq 2$ und* konstantes $p \geq 2$ *mit ϵ-beschränkter Differenz bzgl. L_p (also insbesondere quadratische Differenz) bis auf einen logarithmischen Faktor optimal lernen (für $\epsilon \leq (MW)^p/2^{p+1}$).*

Bemerkung 10.7 *Mit der Transformation BLPAC läßt sich $G_n(M,W)$ für $n \in \mathbb{N}$ und $0 < p < 2$ mit ϵ-beschränkter Differenz bzgl. L_p (also insbesondere absolute Differenz) bis auf einen konstanten Faktor optimal lernen (für $\epsilon \leq (MW)^p/(8n^{p/2})$ und $\delta \leq 1/100$).*

Bemerkung 10.8 *Mit der Transformation BLPAC läßt sich* $G_n(M, W)$ *für* $n \in \mathbb{N}$ *und* $p \geq 2$ *mit* ϵ*-beschränkter Differenz bzgl.* L_p *(also insbesondere quadratische Differenz) bis auf einen logarithmischen Faktor optimal lernen (für* $\epsilon \leq (MW)^p/2$*).*

Bemerkung 10.9 *Mit der Transformation BLPAC läßt sich* $F_n(M, W)$ *für* $n \in \mathbb{N}$, $n \geq 2$ *mit* ϵ*-beschränkter Differenz bzgl.* L_0 *bis auf einen konstanten Faktor optimal lernen (für* $\epsilon \leq 1/8$ *und* $\delta \leq 1/100$*).*

Bemerkung 10.10 *Mit der Transformation BLPAC läßt sich* $G_n(M, W)$ *für* $n \in \mathbb{N}$ *mit* ϵ*-beschränkter Differenz bzgl.* L_0 *bis auf einen konstanten Faktor optimal lernen (für* $\epsilon \leq 1/8$ *und* $\delta \leq 1/100$*).*

In Tabelle 3 sind die Resultate des Vergleichs zwischen oberen und unteren Schranken (ohne Angabe der Einschränkungen bzgl. ϵ, γ, δ und n) zusammengefaßt.

A Einige Beweise

A.1 Beweis von Lemma 4.2

Lemma 4.2 (Hoeffding Ungleichungen (multiplikative Form)) *Seien* $X_1, \ldots, X_t$ *unabhängige Wahrscheinlichkeitsvariablen mit* $0 \leq X_i \leq 1$, $T = \sum_{i=1}^{t} X_i$, $p = E(T/t)$. *Nun gilt für* $0 < \alpha < 1$

$$LE(p,t,(1-\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{2}}$$

und für $0 < \alpha < (1-p)/p$ *und* $\alpha \leq 1$

$$GE(p,t,(1+\alpha)tp) \leq e^{-\frac{\alpha^2 tp}{3}}.$$

Für den Beweis dieses Lemmas benötigen wir Lemma 4.1, sowie einige Hilfsresultate.

Lemma 4.1 (Hoeffding Ungleichungen (additive Form) [Hoe63]) *Seien* $X_1, \ldots, X_t$ *unabhängige Wahrscheinlichkeitsvariablen mit* $0 \leq X_i \leq 1$, $T = \sum_{i=1}^{t} X_i$, $p = E(T/t)$, *dann seien* $LE(p,t,r) = P(T \leq r)$ *und* $GE(p,t,r) = P(T \geq r)$. *Nun gilt für* $0 < c < p$

$$\begin{aligned} LE(p,t,(p-c)t) &\leq \left(\left(\frac{1-p}{1-p+c} \right)^{1-p+c} \left(\frac{p}{p-c} \right)^{p-c} \right)^t \\ &\leq e^{-2tc^2} \end{aligned}$$

und für $0 < c < 1-p$

$$\begin{aligned} GE(p,t,(p+c)t) &\leq \left(\left(\frac{p}{p+c} \right)^{p+c} \left(\frac{1-p}{1-p-c} \right)^{1-p-c} \right)^t \\ &\leq e^{-2tc^2}. \end{aligned}$$

Das folgende Lemma ist eine wohlbekannte Abschätzung mittels Taylor Reihen (s. z. B. [Lit89]).

Lemma A.1 *Für alle $x > 0$ gilt*

1.
$$x \ln x \geq x - 1$$

2.
$$x \ln x \geq (x-1) + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3$$

Beweis: Wir benutzen die Taylorsche Formel (mit Lagrangeschem Restglied) an dem Entwicklungspunkt $x_0 = 1$. Für $f(x) = x \ln x$ ist

$$f'(x) = 1 + \ln x,$$
$$f''(x) = 1/x,$$
$$f'''(x) = -1/x^2$$

und

$$f''''(x) = 2/x^3.$$

Für $x = 1$ ist das Lemma offensichtlich erfüllt. Für andere $x > 0$ gibt es $0 < \vartheta_1, \vartheta_2 < 1$, so daß

$$x \ln x = (x-1) + \frac{(x-1)^2}{2(1+\vartheta_1(x-1))}$$

und

$$x \ln x = (x-1) + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3 + \frac{(x-1)^4}{12(1+\vartheta_2(x-1))^3}.$$

Da die Terme, die ϑ_1 bzw. ϑ_2 enthalten, größer als 0 sind, erhalten wir die gewünschten Ungleichungen. □

Das folgende Lemma entspricht dem 2. Teil des Satzes 4.4, der in [Lit89] bewiesen ist.

Lemma A.2 *Für $0 < \alpha < 1$ gilt*

$$\left(\left(\frac{1-p}{1-p+\alpha p}\right)^{1-p+\alpha p}\left(\frac{p}{p-\alpha p}\right)^{p-\alpha p}\right)^t \leq e^{-\frac{\alpha^2 tp}{2}}$$

und für $0 < \alpha < (1-p)/p$ und $\alpha \leq 1$ gilt

$$\left(\left(\frac{p}{p+\alpha p}\right)^{p+\alpha p}\left(\frac{1-p}{1-p-\alpha p}\right)^{1-p-\alpha p}\right)^t \leq e^{-\frac{\alpha^2 tp}{3}}.$$

Beweis: Es gilt

$$
\begin{aligned}
&\left(\left(\frac{1-p}{1-p+\alpha p}\right)^{1-p+\alpha p}\left(\frac{p}{p-\alpha p}\right)^{p-\alpha p}\right)^{t} \\
= \;& \exp\left(t\ln\left(\left(\frac{1-p}{1-p+\alpha p}\right)^{1-p+\alpha p}\left(\frac{p}{p-\alpha p}\right)^{p-\alpha p}\right)\right) \\
= \;& \exp\left(t\left((1-p+\alpha p)\ln\frac{1-p}{1-p+\alpha p}+(p-\alpha p)\ln\frac{p}{p-\alpha p}\right)\right) \\
= \;& \exp\left(-t\left((1-p+\alpha p)\ln\frac{1-p+\alpha p}{1-p}+(p-\alpha p)\ln\frac{p-\alpha p}{p}\right)\right) \\
= \;& \exp\left(-t\left((1-p)\left(1+\frac{\alpha p}{1-p}\right)\ln\left(1+\frac{\alpha p}{1-p}\right)+p(1-\alpha)\ln(1-\alpha)\right)\right).
\end{aligned}
$$

Mit Lemma A.1 1. erhalten wir

$$(1-p)\left(1+\frac{\alpha p}{1-p}\right)\ln\left(1+\frac{\alpha p}{1-p}\right)\geq(1-p)\frac{\alpha p}{1-p}=\alpha p.$$

Mit Lemma A.1 2. erhalten wir

$$p(1-\alpha)\ln(1-\alpha)\geq p\left(-\alpha+\frac{\alpha^2}{2}+\frac{\alpha^3}{6}\right)\geq-\alpha p+\frac{\alpha^2 p}{2}.$$

Somit erhalten wir

$$(1-p)\left(1+\frac{\alpha p}{1-p}\right)\ln\left(1+\frac{\alpha p}{1-p}\right)+p(1-\alpha)\ln(1-\alpha)\geq\frac{\alpha^2 p}{2}$$

und damit die erste Ungleichung.

Es gilt

$$
\begin{aligned}
&\left(\left(\frac{p}{p+\alpha p}\right)^{p+\alpha p}\left(\frac{1-p}{1-p-\alpha p}\right)^{1-p-\alpha p}\right)^{t} \\
&= \exp\left(t\ln\left(\left(\frac{p}{p+\alpha p}\right)^{p+\alpha p}\left(\frac{1-p}{1-p-\alpha p}\right)^{1-p-\alpha p}\right)\right) \\
&= \exp\left(t\left((p+\alpha p)\ln\left(\frac{p}{p+\alpha p}\right)+(1-p-\alpha p)\ln\left(\frac{1-p}{1-p-\alpha p}\right)\right)\right) \\
&= \exp\left(-t\left((p+\alpha p)\ln\left(\frac{p+\alpha p}{p}\right)+(1-p-\alpha p)\ln\left(\frac{1-p-\alpha p}{1-p}\right)\right)\right) \\
&= \exp\left(-t\left(p(1+\alpha)\ln(1+\alpha)+(1-p)\left(1-\frac{\alpha p}{1-p}\right)\ln\left(1-\frac{\alpha p}{1-p}\right)\right)\right).
\end{aligned}
$$

Mit Lemma A.1 2. erhalten wir

$$p(1+\alpha)\ln(1+\alpha) \geq p\left(\alpha+\frac{\alpha^2}{2}-\frac{\alpha^3}{6}\right) = \alpha p\left(1+\frac{\alpha}{2}\left(1-\frac{\alpha}{3}\right)\right).$$

Für $\alpha \leq 1$ (also $1-\alpha/3 \geq 2/3$) erhalten wir

$$p(1+\alpha)\ln(1+\alpha) \geq \alpha p\left(1+\frac{\alpha}{3}\right) = \alpha p+\frac{\alpha^2 p}{3}.$$

Mit Lemma A.1 1. erhalten wir

$$(1-p)\left(1-\frac{\alpha p}{1-p}\right)\ln\left(1-\frac{\alpha p}{1-p}\right) \geq (1-p)\frac{-\alpha p}{1-p} = -\alpha p.$$

Somit erhalten wir

$$p(1+\alpha)\ln(1+\alpha)+(1-p)\left(1-\frac{\alpha p}{1-p}\right)\ln\left(1-\frac{\alpha p}{1-p}\right) \geq \frac{\alpha^2 p}{3}$$

und damit die zweite Ungleichung. □

Nun folgt Lemma 4.2 aus Lemma 4.1 mit $c = \alpha p$ und Lemma A.2. □

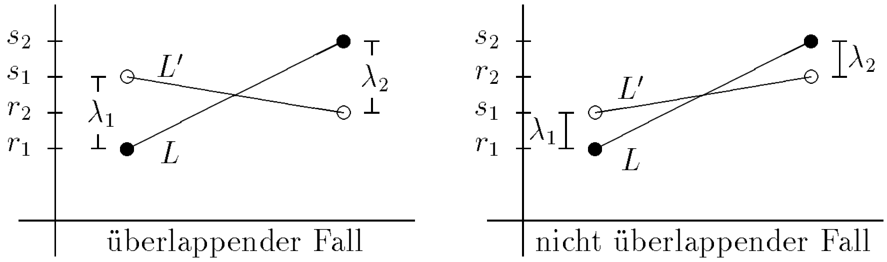

Abbildung 2: Zwei duale Liniensegmente L, L'. Die Summe ihrer Höhendifferenzen $|s_2 - r_1| + |s_1 - r_2|$ ist mindestens $\lambda_1 + \lambda_2$.

A.2 Beweis von Lemma 8.4

Lemma 8.4 $d_{\mathcal{F}_1}(\gamma) \leq \lfloor (1 + 1/\gamma)/2 \rfloor$.

Für ähnlich definiertes $\mathcal{F}_1$ hat Simon [Sim93] $d_{\mathcal{F}_1}(\gamma) \leq 1 + \lfloor 1/(2\gamma) \rfloor$ gezeigt. Wir benutzen die gleiche Beweisidee.

Das folgende Lemma von Simon [Sim93] gilt auch für unser (etwas anders definiertes) $\mathcal{F}_1$.

Lemma A.3 *Sei $f \in \mathcal{F}_1$, dann gilt*

$$\sum_{i=0}^{r-1} |f(x_{i+1}) - f(x_i)| \leq 1$$

für alle $r \geq 0$ und alle Folgen $x_0 < \cdots < x_r$ von Zahlen aus $[0, 1]$.

Das folgende Lemma ist eine Anpassung des Lemmas 4.2 aus [Sim93] für unsere Funktionenklassen.

Lemma A.4 *Sei $d \geq 2$. Sei $0 \leq r_i < s_i \leq 1$ und $\lambda_i = s_i - r_i$ für $i = 1, \ldots, d$, dann ist es möglich, $t_i \in \{r_i, s_i\}$ so zu wählen, daß*

$$|t_1| + \sum_{i=1}^{d-1} |t_{i+1} - t_i| \geq \frac{\lambda_d}{2} + \sum_{i=1}^{d-1} \lambda_i$$

gilt.

Beweis: Wähle die Folgen $T = (t_i), T' = (t'_i)$ durch

$$T = r_1, s_2, r_3, s_4, \dots \text{ und } T' = s_1, r_2, s_3, r_4, \dots.$$

Abbildung 2 zeigt, daß

$$\sum_{i=1}^{d-1}(|t_{i+1} - t_i| + |t'_{i+1} - t'_i|) \geq \sum_{i=1}^{d-1}(\lambda_i + \lambda_{i+1})$$

gilt. Außerdem gilt offensichtlich $|t_1| + |t'_1| \geq \lambda_1$ und somit

$$|t_1| + |t'_1| + \sum_{i=1}^{d-1}(|t_{i+1} - t_i| + |t'_{i+1} - t'_i|) \geq \lambda_d + 2\sum_{i=1}^{d-1}\lambda_i.$$

Also erfüllt T oder T' die Behauptung dieses Lemmas. □

Sei nun $S = ((x_i, r_i, s_i))_{1 \leq i \leq d}$ eine Folge, die durch $\mathcal{F}_1$ streng-γ-fat-aufgespalten wird. O.B.d.A. sei $x_1 < \cdots < x_d$. Wähle $t_i \in \{r_i, s_i\}$ entsprechend Lemma A.4. Wähle $b \in \{0,1\}^d$, so daß $f = f_b$ an dem Punkt x_i den Wert t_i annimmt. Setze $x_0 = 0$. Wegen $s_i - r_i \geq 2\gamma$ gilt mit den Lemmata A.3 und A.4

$$\gamma + (d-1)2\gamma \leq \sum_{i=0}^{d-1}|f(x_{i+1}) - f(x_i)| \leq 1.$$

Somit gilt $d \leq (1 + 1/\gamma)/2$ und wegen $d \in \mathbb{N}$ dann $d \leq \lfloor(1 + 1/\gamma)/2\rfloor$. □

A.3 Beweis von Lemma 8.7

Lemma 8.7 *Für $n \in \mathbb{N}, n \geq 2$ ist $F'_n(W)$ ein $(n-1)$-dimensionaler Vektorraum von Funktionen.*

Da sich alle Funktionen aus $F'_n(W)$ durch ihren Gewichtsvektor $\vec{w}$ darstellen lassen, reicht es z. z., daß

$$C_n = \left\{ \vec{w} = (w_1, \dots, w_n) \;\middle|\; \sum_{i=1}^{n} w_i = W \right\}$$

ein $(n-1)$-dimensionaler $\mathbb{R}$-Vektorraum ist. C_n ist also der affine Unterraum von $\mathbb{R}^n$, der Lösungsraum des Gleichungssystems $\sum_{i=1}^{n} w_i = W$ ist. (Hier besteht das Gleichungssystem nur aus einer Gleichung.) Wir zeigen nun, daß wenn ein gegebenes lineares Gleichungssystem $A\vec{x} = \vec{b}$ lösbar ist und $n - m$ die Anzahl der linear unabhängigen Spalten der Matrix A ist, zu dem Lösungsraum dieses Gleichungssystems eine Addition und eine (äußere) Multiplikation existiert, so daß der Lösungsraum mit diesen Verknüpfungen ein m-dimensionaler $\mathbb{R}$-Vektorraum ist.

Sei $A\vec{x} = \vec{b}$ ein lösbares Gleichungssystem mit $A \in \mathbb{R}^{m \times n}$ und $\vec{b} \in \mathbb{R}^m$, dann sei $L(A, \vec{b})$ der Lösungsraum dieses Gleichungssystems. Sei desweiteren $Sp(A)$ die Anzahl der linear unabhängigen Spalten in A, dann ist wohlbekannt (s. z. B. Satz 6.8 und 6.14 in [Weg91]), daß $L(A, \vec{b})$ affiner Unterraum von $\mathbb{R}^n$ mit Dimension $n - Sp(A)$ ist.

Lemma A.5 *Sei L_a ein affiner Unterraum von $\mathbb{R}^n$ mit Dimension $m \leq n$. Dann läßt sich eine Addition $\oplus : L_a \times L_a \to L_a$ und eine (äußere) Multiplikation $\odot : \mathbb{R} \times L_a \to L_a$ bestimmen, so daß $(L_a, \oplus, \odot)$ ein m-dimensionaler reeller Vektorraum ist.*

Beweis: Sei L_a ein affiner Unterraum von $\mathbb{R}^n$ mit Dimension $m \leq n$, dann lassen sich m Variablen bestimmen, so daß für jede Wahl dieser Variablen genau ein $\vec{x} \in L_a$ existiert. O.B.d.A. seien dies die Variablen $x_1, \ldots, x_m$. Dann lassen sich $x_{m+1}, \ldots, x_n$ mit linearen Gleichungen bestimmen. Sei $\vec{x}' = (x_1, \ldots, x_m)$, $\vec{x}'' = (x_{m+1}, \ldots, x_n)$, $k = m - n$. Seien $B \in \mathbb{R}^{k \times m}$ und $\vec{a} \in \mathbb{R}^k$ so gewählt, daß $\vec{x}'' = B\vec{x}' + \vec{a}$ gilt. (Wird der affine Unterraum durch Lösung eines linearen Gleichungssystems bestimmt, so erhält man dabei automatisch B und $\vec{a}$.) Seien $\vec{x}'$ und $\vec{x}''$ wie oben, sowie $\vec{y}'$ und $\vec{y}''$ analog für $\vec{y}$ definiert, dann können wir $(\vec{x}', \vec{x}'')$ für $\vec{x}$ schreiben ($\vec{y}$ analog). Sei $r \in \mathbb{R}$. Die Addition ist dann durch

$$(\vec{x}', \vec{x}'') \oplus (\vec{y}', \vec{y}'') = (\vec{x}' + \vec{y}', B(\vec{x}' + \vec{y}') + \vec{a})$$

und die (äußere) Multiplikation durch

$$r \odot (\vec{x}', \vec{x}'') = (r\vec{x}', B(r\vec{x}') + \vec{a})$$

definiert. Die Verknüpfungen erfüllen offensichtlich die Eigenschaften, die für einen Vektorraum notwendig sind. Da dieser Vektorraum isomorph zu $\mathbb{R}^m$ ist, hat er die Dimension m. $\square$

Wir setzen $A = (1, \ldots, 1)$ und $\vec{b} = W$. Da A nur eine linear unabhängige Spalte enthält, erhalten wir mit Lemma A.5 und der Bemerkung davor, daß C_n für $n \geq 2$ ein $(n-1)$-dimensionaler $\mathbb{R}$-Vektorraum ist. (Wir erhalten $B = (-1, \ldots, -1) \in \mathbb{R}^{1 \times (n-1)}$ und $\vec{a} = W \in \mathbb{R}$.) $\square$

Literatur

[AB94] Martin Anthony und Peter Bartlett. Function learning from interpolation. NeuroCOLT Technical Report Series NC-TR-94-013, Royal Holloway, University of London, Oktober 1994.

[ABIST94] Martin Anthony, Peter Bartlett, Yuval Ishai und John Shawe-Taylor. Valid generalisation from approximate interpolation. Technical Report, 1994. Erscheint in *Combinatorics, Probability and Computing*.

[Ang88] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, April 1988.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler und Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, Oktober 1989.

[Ber92] Ethan J. Bernstein. Absolute error bounds for learning linear functions online. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Seiten 160–163. ACM Press, New York, NY, 1992.

[Dud78] R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, September 1992.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, März 1963.

[KL92] Don Kimber und Philip M. Long. The learning complexity of smooth functions of a single variable. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Seiten 153–159. ACM Press, New York, NY, 1992.

[KS94] Michael Kearns und Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48:464–497, 1994.

[KW94] Jyrki Kivinen und Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. Technical Report UCSC-CRL-94-16, University of California, Santa Cruz, Juni 1994.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[Lit89] Nick Littlestone. From on-line to batch learning. In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory*, Seiten 269–284. Morgan Kaufmann, San Mateo, CA, 1989.

[LLW91] Nick Littlestone, Philip M. Long und Manfred K. Warmuth. On-line learning of linear functions. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*, Seiten 465–475, 1991.

[Nat89] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[Shi84] A. N. Shiryayev. *Probability*. Graduate Texts in Mathematics. Springer, New York, 1984.

[Sim93] Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. Forschungsbericht Nr. 501, FB Informatik, LS II, Universität Dortmund, 1993. Ein erweitertes Abstrakt ist erschienen in *Computational Learning Theory: EUROCOLT'93*.

[Val84] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Weg91] Gerd Wegner. Lineare Algebra für Informatiker. Skriptum zu den Grundvorlesungen Lineare Algebra für Informatiker I (WS 1990/91) und II (SS 1991), FB Mathematik, Universität Dortmund, 1991.