10 Die Fourier-Transformation

Bisher betrachteten wir boolesche Konzepte als Teilmengen $c \subseteq \{0,1\}^n$ bzw. als deren charakteristische Funktionen $\chi_c : \{0,1\}^n \to \{0,1\}$. In diesem Abschnitt betrachten wir die charakteristische Funktion eines booleschen Konzepts als spezielle reellen Funktionen $f : \{0,1\}^n \to \mathbb{R}$. Dabei wird es sich als nützlich erweisen, den Wertebereich $\{0,1\}$ von χ_c durch die Menge $\{-1,+1\}$ zu ersetzen, d.h. wir identifizieren Teilmengen $c \subseteq \{0,1\}^n$ mit Funktionen $f : \{0,1\}^n \to \{-1,+1\}$ definiert durch

$$f(x) = \begin{cases} +1 & \text{falls } x \in c \\ -1 & \text{falls } x \notin c \end{cases}$$

Die Menge aller reellen Funktionen $f : \{0,1\}^n \to \mathbb{R}$ können wir als *reellen Vektorraum* der Dimension 2^n auffassen. Für diesen Vektorraum betrachten wir das folgende *innere Produkt* auf diesem Vektorraum, das je zwei Funktionen $f, g : \{0,1\}^n \to \mathbb{R}$ eine reelle Zahl zuordnet,

$$\langle f, g \rangle = \mathrm{E}[f(x) \cdot g(x)] = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x) \cdot g(x)$$

Wir schreiben im folgendem auch einfach $\mathrm{E}[f \cdot g]$, wenn der Erwartungswert $\mathrm{E}[f(x) \cdot g(x)]$ nur über ein zufälliges $x \in_{\mathcal{U}} \{0,1\}^n$ genommen wird.

Unmittelbar aus der Definition und der Linearität des Erwartungswertes folgt für alle Funktionen $f, g, h : \{0,1\}^n \to \mathbb{R}$ und $a, b \in \mathbb{R}$,

1. $\langle af + bg, h \rangle = a\langle f, g \rangle + b\langle g, h \rangle$
2. $\langle f, g \rangle = \langle g, f \rangle$
3. $\langle f, f \rangle \geq 0$, und $\langle f, f \rangle = 0$ genau dann wenn $f = 0$

Die Abbildung $\langle \cdot, \cdot \rangle$ erfüllt also tatsächlich die Bedingungen eines inneren Produkts, und die Menge aller reellen Funktionen $f : \{0,1\}^n \to \mathbb{R}$ zusammen mit diesem inneren Produkt bilden einen *inneren Produktraum.*

Mit Hilfe des inneren Produkts können wir nun die *Norm* einer Funktion $f : \{0,1\}^n \to \mathbb{R}$ wie üblich durch $\|f\| = \sqrt{\langle f, f \rangle}$ definieren.

10.1 Die Fourier-Basis

Die *Fourier-Basis* für die Menge aller Funktionen $f : \{0,1\}^n \to \mathbb{R}$ ist die Menge $\{\chi_\alpha \mid \alpha \in \{0,1\}^n\}$, wobei jede Funktion $\chi_\alpha : \{0,1\}^n \to \{-1,+1\}$ durch

$$\chi_\alpha(x) = (-1)^{\sum_{i=1}^n x_i \alpha_i}$$

für $\alpha = \alpha_1 \dots \alpha_n$ und $x = x_1 \dots x_n$ definiert ist. Alternativ können wir die Fourier-Basis auch durch die Menge $\{\chi_A \mid A \subseteq \{1,...,n\}\}$ beschreiben, wobei

$$\chi_A(x) = \prod_{i \in A} (-1)^{x_i} = \begin{cases} +1 & \text{falls } \sum_{i \in A} x_i \bmod 2 = 0 \\ -1 & \text{sonst.} \end{cases}$$

Zwei Funktionen f und g heißen *orthogonal*, wenn $\langle f, g \rangle = 0$ gilt. Entsprechenden heißt eine Menge B von Funktionen *orthogonal*, wenn je zwei verschieden Funktionen von B orthogonal sind. Darüberhinaus heißt eine orthogonale Menge B *orthonormal*, wenn alle Funktionen f in B die Norm eins besitzen, d.h. es gilt $\|f\| = \langle f, f \rangle = 1$. Die Menge B ist also genau dann orthonormal, wenn für alle Funktionen f und g in B gilt:

$$\langle f, g \rangle = \begin{cases} 0 & \text{falls } f \neq g \\ 1 & \text{falls } f = g \end{cases}$$

Lemma 10.1. *Die Fourier-Basis ist orthonormal.*

Beweis. Für alle $\alpha \in \{0,1\}^n$ und $x \in \{0,1\}^n$ gilt $\chi_\alpha^2(x) = 1$ und somit $\|\chi_\alpha\| = \sqrt{E(\chi_\alpha^2(x))} = 1$. Es bleibt zu zeigen, daß $\langle \chi_\alpha, \chi_\beta \rangle = 0$ für alle $\alpha \neq \beta$ gilt. (Übung) □

Ist eine Teilmenge eines inneren Produktraums orthonormal, dann ist diese Teilmenge linear unabhängig, und falls darüberhinaus die Kardinalität dieser Teilmenge mit der Dimension des inneren Produktraums übereinstimmt, dann bildet die Teilmenge eine Basis für den Produktraum.

Lemma 10.2. *Die Fourier-Basis ist linear unabhängig. Insbesondere kann jede Funktion $f : \{0,1\}^n \to \mathbb{R}$ als Linearkombination*

$$f(x) = \sum_\alpha \hat{f}(\alpha)\chi_\alpha(x)$$

mit Koeffizienten $\hat{f}(\alpha) = \langle f, \chi_\alpha \rangle$ *dargestellt werden.*

Beweis. Für die lineare Unabhängigkeit nehmen wir an,

$$\sum_\alpha a_\alpha \chi_\alpha = 0$$

für Koeffizienten a_α mit $\alpha \in \{0,1\}^n$ gilt. Da $\chi_\beta \neq 0$ für alle $\beta \in \{0,1\}^n$ gilt, folgt aus der Orthogonalität der Fourier-Basis

$$\langle \chi_\beta, \sum_\alpha a_\alpha \chi_\alpha \rangle = \langle \chi_\beta, 0 \rangle = 0.$$

Außerdem folgt aus den Eigenschaften des inneren Produkts und der Orthonormalität

$$\langle \chi_\beta, \sum_\alpha a_\alpha \chi_\alpha \rangle = \sum_\alpha a_\alpha \langle \chi_\beta, \chi_\alpha \rangle = a_\beta$$

Also gilt $a_\beta = 0$ für alle $\beta \in \{0,1\}^n$, d.h. die Fourier-Basis ist linear unabhängig.

Da wir 2^n Funktionen in der Fourier-Basis haben, bilden diese Funktionen eine Basis für die Menge aller Funktionen $f : \{0,1\}^n \to \mathbb{R}$. Also hat jede Funktion $f : \{0,1\}^n \to \mathbb{R}$ eine eindeutige Darstellung

$$f = \sum_\alpha a_\alpha \chi_\alpha,$$

und es gilt

$$\langle f, \chi_\alpha \rangle = \langle \sum_\beta a_\beta \chi_\beta, \chi_\alpha \rangle = \sum_\beta a_\beta \langle \chi_\beta, \chi_\alpha \rangle = a_\alpha.$$

□

Beispiel 10.3. Sei $f(x) = x_1 \wedge x_2$. Wir können f und die Parityfunktionen durch die folgende Tabelle beschreiben.

x	f	χ_{00}	χ_{01}	χ_{10}	χ_{11}
00	-1	1	1	1	1
01	-1	1	-1	1	-1
10	-1	1	1	-1	-1
11	1	1	-1	-1	1

Mit Hilfe der Tabelle können wir die Koeffizienten wie folgt bestimmen.

$$\begin{aligned}
\hat{f}(00) &= \mathrm{E}[f \cdot \chi_{00}] = \tfrac{1}{4}(-1-1-1+1) = -\tfrac{1}{2} \\
\hat{f}(01) &= \mathrm{E}[f \cdot \chi_{01}] = \tfrac{1}{4}(-1+1-1-1) = -\tfrac{1}{2} \\
\hat{f}(10) &= \mathrm{E}[f \cdot \chi_{10}] = \tfrac{1}{4}(-1-1+1-1) = -\tfrac{1}{2} \\
\hat{f}(11) &= \mathrm{E}[f \cdot \chi_{11}] = \tfrac{1}{4}(-1+1+1+1) = +\tfrac{1}{2}
\end{aligned}$$

Also ist $f(x) = -\frac{1}{2}\left(\chi_{00}(x) + \chi_{10}(x) + \chi_{01}(x)\right) + \frac{1}{2}\chi_{11}(x)$.

Lemma 10.4 (Parseval). *Für alle $f : \{0,1\}^n \to \mathbb{R}$ gilt*

$$\langle f, f \rangle = \sum_{\alpha \in \{0,1\}^n} \hat{f}^2(\alpha)$$

Beweis. Aus den Eigenschaften des inneren Produkts und der Orthonormalität der Fourier-Basis folgt

$$\begin{aligned}
\langle f, f \rangle &= \langle \sum_\alpha \hat{f}(\alpha)\chi_\alpha, \sum_\beta \hat{f}(\beta)\chi_\beta \rangle \\
&= \sum_\alpha \sum_\beta \hat{f}(\alpha)\hat{f}(\beta)\langle \chi_\alpha, \chi_\beta \rangle \\
&= \sum_\alpha \hat{f}^2(\alpha)
\end{aligned}$$

□

Für boolesche Funktionen $f : \{0,1\}^n \to \{-1,+1\}$ gilt $f^2(x) = 1$ für alle $x \in \{0,1\}^n$, und somit $\mathrm{E}[f^2] = 1$. Damit folgt aus der Parsevalgleichung, daß für solche Funktionen insbesondere

$$\sum_{\alpha \in \{0,1\}^n} \hat{f}^2(\alpha) = 1$$

gilt. Außerdem beachten wir, daß sich die Fourier-Transformation linear verhält.

Lemma 10.5. *Sei* $h = c \cdot f + g$ *für* $f, g : \{0,1\}^n \to \mathbb{R}$ *und* $c \in \mathbb{R}$. *Dann gilt* $\hat{h}(\alpha) = c \cdot \hat{f}(\alpha) + \hat{g}(\alpha)$ *für alle* $\alpha \in \{0,1\}^n$.

10.2 Approximation mit großen Fourier-Koeffizienten

Sei $f : \{0,1\}^n \to \{-1,+1\}$ eine boolesche Funktion mit $|\hat{f}(\beta)| \geq \theta$ für ein Parity $\beta \in \{0,1\}^n$ und eine reelle Schranke $\theta > 0$, und sei $h(x) = \hat{f}(\beta) \cdot \chi_\beta$. Wir beachten, daß h nicht unbedingt eine boolesche Funktion mit Wertebereich $\{-1,+1\}$ ist , sondern in der Regel eine reelwertige Funktion ist. Daher betrachten wir zur Bestimmung der Approximationsgüte das “erwartete Fehlerquadrat”,

$$\mathrm{E}[(f-h)^2].$$

Für $g(x) = f(x) - h(x)$ gilt wegen der Linearität der Fourier-Transformation für alle Koeffizienten mit Parity $\alpha \in \{0,1\}^n$,

$$\hat{g}(\alpha) = \begin{cases} \hat{f}(\alpha) & \text{falls } \alpha \neq \beta \\ 0 & \text{sonst} \end{cases}$$

Mit der Parsevalgleichung folgt

$$\mathrm{E}[(f-h)^2] = \mathrm{E}[g^2] = \sum_\alpha \hat{g}^2(\alpha) = \sum_{\alpha\neq\beta} \hat{f}^2(\alpha) = \underbrace{\sum_\alpha \hat{f}^2(\alpha)}_{=1} - \underbrace{\hat{f}^2(\beta)}_{\geq\theta^2} \leq 1-\theta^2$$

Für eine gute Approximation von f genügt also ein einzelner Fourier-Koeffizient $\hat{f}(\beta)$, sofern dessen Betrag nur ausreichend groß ist. Eine solche reelwertige Approximation von f kann einfach in eine boolesche Funktion überführt werden, so daß der Approximationsfehler im üblichen Sinne durch das erwartete Fehlerquadrat beschränkt ist. Hierzu betrachten wir das Vorzeichen der reelwertigen Approximation.

Definition. *Sei* $sign : \mathbb{R} \to \{-1,+1\}$ *definiert durch*

$$sign(x) = \begin{cases} +1 & \textit{falls } x \geq 0 \\ -1 & \textit{falls } x < 0 \end{cases}$$

Lemma 10.6. *Sei $f : \{0,1\}^n \to \{-1,+1\}$ und $h : \{0,1\}^n \to \mathbb{R}$. Dann gilt*

$$\Pr[f(x) \neq sign(h(x))] \leq \mathrm{E}[(f-h)^2]$$

Beweis. Sei $I(x)$ die Indikatorfunktion definiert durch

$$I(x) = \begin{cases} 1 & \text{falls } f(x) \neq sign(h(x)) \\ 0 & \text{sonst} \end{cases}$$

Dann gilt

$$\Pr[f(x) \neq sign(h(x))] = \mathrm{E}[I(x)] = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} I(x).$$

Falls $f(x) = sign(h(x))$, dann gilt $I(x) = 0 \leq (f(x) - h(x))^2$. Falls $f(x) \neq sign(h(x))$, dann gilt $|f(x) - h(x)| \geq 1$ und somit auch in diesem Fall $I(x) = 1 \leq (f(x) - h(x))^2$. $\square$

10.3 Lernen mit der Fourier-Transformation

Wie im vorigen Abschnitt sei $f : \{0,1\}^n \to \{-1,+1\}$ eine boolesche Funktion mit $|\hat{f}(\beta)| \geq \theta$ für $\beta \in \{0,1\}^n$ und $\theta > 0$. Nun aber sei $h(x) = a_\beta \chi_\beta(x)$ für eine Approximation a_β des Koeffizienten $\hat{f}(\beta)$ mit

$$|\hat{f}(\beta) - a_\beta| \leq \lambda.$$

Dann gilt für das erwartete Fehlerquadrat

$$\begin{aligned} E\left((f-h)^2\right) &= \sum_{\alpha \neq \beta} \hat{f}^2(\alpha) + \underbrace{\left(\hat{f}(\beta) - a_\beta\right)^2}_{\leq \lambda^2} \\ &\leq \sum_{\alpha} \hat{f}^2(\alpha) - \hat{f}^2(\beta) + \lambda^2 \\ &\leq 1 - \theta^2 + \lambda^2 \end{aligned}$$

Gegenüber einer Approximation von f mit dem tatsächlichen Wert des Koeffizienten $\hat{f}(\beta)$ erhöht sich der Fehler einer Approximation mit dem Wert a_β nur um λ^2.

Der Koeffizient

$$\hat{f}(\beta) = \langle f, \chi_\beta \rangle = \mathop{\mathrm{E}}_{x \in \{0,1\}^n} (f(x) \cdot \chi_\beta(x))$$

ist der Erwartungswert der $\{-1,+1\}$-wertigen Zufallsvariablen $Y = f(x)\chi_\beta(x)$, für $x \in_{\mathcal{U}} \{0,1\}^n$. Dieser Erwartungswert kann mit Hilfe der folgenden Chernoff-Schranke mit einer kleinen Stichprobe mit hoher Wahrscheinlichkeit gut approximiert werden.

Chernoff-Schranke. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte $\{-1,+1\}$-wertige Zufallsvariablen mit $E(X_i) = p$. Dann gilt für $\lambda > 0$:

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} x_i - p\right| \geq \lambda\right] \leq 2 \cdot e^{-\lambda^2 m/2}$$

Also können wir einen Koeffizient $\hat{f}(\beta)$ mit einer kleinen kleinen Anzahl von klassifizierten Beispielen $(x, f(x))$, $x \in_{\mathcal{U}} \{0,1\}^n$, mit hoher Wahrscheinlichkeit gut approximiert werden. Dies ist nicht nur für einen einzelnen Koeffizienten, sondern auch für eine ganze Menge von Koeffizienten möglich.

Algorithmus *Approx*(S, ε, δ)

1. Generiere eine zufällige Beispielfolge $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$ der Größe
$$m = \frac{2|S|}{\varepsilon}\sqrt{\ln\frac{2|S|}{\delta}}$$
2. Für alle $\alpha \in S$ berechne $a_\alpha = \frac{1}{m}\sum_{i=1}^{m} f(x_i) \cdot \chi_\alpha(x_i)$
3. Ausgabe $h(x) = \sum_{\alpha \in S} a_\alpha \chi_\alpha(x)$

Definition. *Die* Trägermenge (support) *einer Funktion* $f : \{0,1\}^n \to \mathbb{R}$ *ist* $S = \{\alpha \mid \hat{f}(\alpha) \neq 0\}$.

Lemma 10.7. *Für alle* $f : \{0,1\}^n \to \{-1,+1\}$, *für alle* $\varepsilon, \delta > 0$ *und* $S \subseteq \{0,1\}^n$ *gibt Approx*(S, ε, δ) *mit Wahrscheinlichkeit* $\geq 1 - \delta$ *eine Hypothese* h *aus, so daß gilt:*

1. die Trägermenge von h ist in S enthalten

2. $\mathrm{E}[(f-h)^2] \le 1 - \sum_{\alpha \in S} \hat{f}^2(\alpha) + \varepsilon$

Die Laufzeit von Approx(S, ε, δ) ist polynomiell in n, $|S|$, $\frac{1}{\varepsilon}$ und $\log(\frac{1}{\delta})$.

Beweis. Aus der Chernoff-Schranke mit $\lambda = \sqrt{\frac{\varepsilon}{|S|}}$ folgt

$$\begin{aligned}
\Pr\left[\exists \alpha \in S : |a_\alpha - \hat{f}(\alpha)| \ge \lambda\right] &\le \sum_{\alpha \in S} \Pr\left[|a_\alpha - \hat{f}(\alpha)| \ge \lambda\right] \\
&\le |S| \cdot 2e^{-\lambda^2 m/2} \\
&= |S| 2e^{-(\varepsilon/|S|)m/2} \le \delta
\end{aligned}$$

Also gilt mit Wahrscheinlichkeit $\ge 1-\delta$, daß $|a_\alpha - \hat{f}(\alpha)| < \lambda$ für alle $\alpha \in S$, und da $\hat{h}(\alpha) = 0$ für alle $\alpha \notin S$, folgt

$$\begin{aligned}
\mathrm{E}[(f-h)^2] &= \sum_{\alpha} (\hat{f}(\alpha) - \hat{h}(\alpha))^2 \\
&= \sum_{\alpha \in S} (\hat{f}(\alpha) - \hat{h}(\alpha))^2 + \sum_{\alpha \notin S} (\hat{f}(\alpha) - \hat{h}(\alpha))^2 \\
&= \sum_{\alpha \in S} (\hat{f}(\alpha) - a_\alpha)^2 + \sum_{\alpha \notin S} \hat{f}^2(\alpha) \\
&< \sum_{\alpha \in S} \lambda^2 + \sum_{\alpha \notin S} \hat{f}^2(\alpha) \\
&= \underbrace{|S|\lambda^2}_{=\varepsilon} + 1 - \sum_{\alpha \in S} \hat{f}^2(\alpha) \\
&= \varepsilon + 1 - \sum_{\alpha \in S} \hat{f}^2(\alpha).
\end{aligned}$$

□

Beispiel 10.8. Sei C_n die Menge aller Funktionen $f : \{0,1\}^n \to \{-1,+1\}$, so dass für alle $k \ge 0$ gilt:

$$\sum_{A \in S_k} \hat{f}(A) \ge 1 - \frac{1}{k}, \quad S_k = \{A \subseteq \{1, ..., n\} \mid |A| \le k\}.$$

Für eine gegebene Funktion $f \in \mathcal{C}_n$ und für $\varepsilon > 0$, sei k die kleinste Zahl mit $n^k \ge \frac{2}{\varepsilon}$, und sei h die Ausgabe von *Approx*$(S_k, \epsilon/2, \delta)$. Dann gilt mit

Wahrscheinlichkeit $\geq 1-\delta$:

$$\begin{aligned}\mathrm{E}[(f-h)^2] &\leq 1 - \sum_{A\in S_k} \hat{f}^2(A) + \frac{\varepsilon}{2} \\ &\leq 1 - (1 - \underbrace{\frac{1}{n^k}}_{\leq \varepsilon/2}) + \frac{\varepsilon}{2} \leq \varepsilon,\end{aligned}$$

und damit $\Pr[f(x) \neq sign(h(x))] \leq \varepsilon$. Die Laufzeit von $Approx(S_k, \varepsilon/2, \delta)$ ist wegen $|S_k| = n^k = O(\frac{1}{\varepsilon})$ polynomiell in n, $\frac{1}{\varepsilon}$, und $\log\frac{1}{\delta}$.

10.4 Bestimmung großer Fourier-Koeffizienten mittels Präfixsuche

In den beiden letzten Abschnitten haben wir gesehen, wie eine boolesche Funktionen $f : \{0,1\}^n \to \{-1,+1\}$ durch Approximation ihrer großen Koeffizienten $\hat{f}(\alpha)$ erlernt werden kann. Dabei haben wir vorausgesetzt, daß die entsprechenden Parities $\alpha \in \{0,1\}^n$ bereits bekannt sind. In diesem Abschnitt sehen wir, wie eine Menge von Parities effizient durch Präfixsuche bestimmt werden kann.

Bei dieser Präfixsuche beginnen wir mit $\alpha = \lambda$, und für jeden gegebenen Präfix α der Länge $k \leq n$ setzten wir die Suche rekursiv mit $\alpha 0$ und $\alpha 1$ fort, bis wir schließlich bei Parities α der Länge n angelangt sind. Im Idealfall setzten wir dabei die Suche nur dann fort, falls es für den aktuellen Präfix α eine Fortsetzung β gibt, für die der entsprechende Koeffizient $\hat{f}(\alpha\beta)$ genügend groß ist. Leider ist nicht klar, wie wir diese Bedingung effizient überprüfen können. Die Aufgabe besteht also darin, eine einfache Abbruchbedingung für die Präfixsuche zu finden, die einerseits garantiert, daß wir alle Parities mit großen Koeffizienten finden, und wir anderseits nicht zu vielen rekursive Aufrufe benötigen.

Zunächst aber halten wir fest, daß die Anzahl der Parities $\alpha \in \{0,1\}^n$, für die der Betrag der Koeffizienten $\hat{f}(\alpha)$ einen bestimmten Schwellenwert $\theta >$ überschreiten, höchstens polynomiell in $1/\theta$ sein kann. Somit kann die Präfixsuche deutlich effizienter als die naive Betrachtung aller Parities $\alpha \in \{0,1\}^n$ sein.

Lemma 10.9. *Sei $f : \{0,1\}^n \to \{-1,+1\}$ und $\theta > 0$. Dann gilt*

$$|\{\alpha \in \{0,1\}^n \mid |\hat{f}(\alpha)| \geq \theta\}| \leq \frac{1}{\theta^2}.$$

Beweis. Aus der Annahme $|\{\alpha \in \{0,1\}^n \mid |\hat{f}(\alpha)| \geq \theta\}| > \frac{1}{\theta^2}$ folgt

$$\sum_\alpha \hat{f}^2(\alpha) > \frac{1}{\theta^2} \cdot \theta^2 = 1.$$

Dies widerspricht aber der Parsevalgleichung $\sum_\alpha \hat{f}^2(\alpha) = 1$. □

Für einen gegeben Präfix α der Größe $k \leq n$ definieren wir nun eine Funktion f_α, deren Trägermenge aus allen Parities mit Präfix α besteht. Wir benutzen dabei die Konvention $\chi_\lambda(\lambda) = 1$.

Definition. *Für $f : \{0,1\}^n \to \mathbb{R}$ und $\alpha \in \{0,1\}^k$ mit $0 \leq k \leq n$ sei die Funktion $f_\alpha : \{0,1\}^{n-k} \to \mathbb{R}$ definiert durch*

$$f_\alpha(x) = \sum_{\beta \in \{0,1\}^{n-k}} \hat{f}(\alpha\beta)\chi_\beta(x).$$

Insbesondere ist $f_\alpha(x) = f(x)$ für $\alpha = \lambda$, und für alle $\alpha \in \{0,1\}^n$ besteht der Definitionsbereich von f_α nur aus $x = \lambda$ und es gilt $f_\alpha(x) = \hat{f}(\alpha)$.

Für die Abbruchbedingung in der Präfixsuche betrachten wir nun den Erwartungswert $\mathrm{E}[f_\alpha^2]$. Wir zeigen zunächst, daß wir im Fall $\mathrm{E}[f_\alpha^2] < \theta^2$ die Suche abbrechen können, da es dann keine Fortsetzungen β mit einem genügend großen Koeffizienten $\hat{f}(\alpha\beta)$ gibt.

Lemma 10.10. *Sei $f : \{0,1\}^n \to \{-1,+1\}$, $\theta > 0$ und $\alpha \in \{0,1\}^k$ mit $0 \leq k \leq n$. Falls es ein $\beta \in \{0,1\}^{n-k}$ mit $|\hat{f}(\alpha\beta)| \geq \theta$ gibt, dann ist $\mathrm{E}[f_\alpha^2] \geq \theta^2$.*

Beweis. Falls es ein $\beta \in \{0,1\}^{n-k}$ mit $|\hat{f}(\alpha\beta)| \geq \theta$ gibt, dann gilt $|f_\alpha(x)| = |\sum_\beta \hat{f}(\alpha\beta) \cdot \chi_\beta(x)| \geq \theta$, und somit $\mathrm{E}_x[f_\alpha^2(x)] \geq \theta^2$. □

Wir können die Anzahl der Präfixe α mit $\mathrm{E}[f_\alpha^2] \geq \theta^2$ polynomiell in $1/\theta$ beschränken. Daraus folgt, daß mit der Abbruchbedingung $\mathrm{E}[f_\alpha^2] < \theta^2$ höchsten polynomiell in n/θ rekursive Aufrufe erfolgen.

Lemma 10.11. *Sei $f : \{0,1\}^n \to \{-1,+1\}$, $\theta > 0$ und $0 \le k \le n$. Dann gilt*

$$|\{\alpha \in \{0,1\}^k \mid \mathrm{E}[f_\alpha^2] \ge \theta^2\}| \le \frac{1}{\theta^2}$$

Beweis. Aus der Annahme $|\{\alpha \in \{0,1\}^k \mid \mathrm{E}[f_\alpha^2] \ge \theta^2\}| > \frac{1}{\theta^2}$ folgt

$$\sum_\alpha \mathrm{E}[f_\alpha^2] > \frac{1}{\theta^2} \cdot \theta^2 = 1.$$

Analog zur Parsevalgleichung gilt aber

$$\begin{aligned}
\mathrm{E}[f_\alpha^2] &= \langle f_\alpha, f_\alpha \rangle \\
&= \langle \sum_\beta \hat{f}(\alpha\beta)\chi_\beta, \sum_\gamma \hat{f}(\alpha\gamma)\chi_\gamma \rangle \\
&= \sum_\beta \sum_\gamma \hat{f}(\alpha\beta)\hat{f}(\alpha\gamma)\langle \chi_\beta, \chi_\gamma \rangle \\
&= \sum_\beta \hat{f}(\alpha\beta)\hat{f}(\alpha\beta) \\
&= \sum_\beta \hat{f}^2(\alpha\beta),
\end{aligned}$$

und somit

$$\sum_\alpha E(f_\alpha^2) = \sum_\alpha \sum_\beta \hat{f}^2(\alpha\beta) = 1.$$

□

Nun können wir die Präfixsuche durch die folgende Funktion Coeff_{det} beschreiben.

Funktion $\mathit{Coeff}_{det}(\alpha, \theta)$ (deterministische Version)

1. Falls $\mathrm{E}[f_\alpha^2] < \theta^2$, dann Ausgabe $\emptyset$
2. Falls $\mathrm{E}[f_\alpha^2] \ge \theta^2$ und $|\alpha| = n$, dann Ausgabe $\{\alpha\}$
3. Sonst Ausgabe $\mathit{Coeff}_{det}(\alpha 0, \theta) \cup \mathit{Coeff}_{det}(\alpha 1, \theta)$

Aus Lemma 10.10 und Lemma 10.11 folgen unmittelbar die folgenden Eigenschaften der Funktion Coeff_{det}.

1. Für alle $\alpha \in \{0,1\}^n$ mit $|\hat{f}(\alpha)| \geq \theta$ ist $\alpha \in \mathit{Coeff}_{det}(\lambda, \theta)$.
2. Für alle $\alpha \in \mathit{Coeff}_{det}(\lambda, \theta)$ ist $|\hat{f}(\alpha)| \geq \theta$, und damit gilt $|\mathit{Coeff}_{det}(\lambda, \theta)| \leq 1/\theta^2$.
3. Es erfolgen höchstens n/θ^2 rekursive Aufrufe.

Es ist nicht klar, wie wir den Erwartungswert $\mathrm{E}[f_\alpha^2]$ effizient und mit höchsten polynomiell vielen klassifizierten Bespielen für f bestimmen können. Wir werden aber sehen, daß wir für die Präfixsuche tatsächlich nur eine Approximation des Erwartungswert $\mathrm{E}[f_\alpha^2]$ benötigen. Für diese Approximation benützen wir die folgende Darstellung der Funktion f_α als Erwartungswert bestimmter Werte von f.

Lemma 10.12. *Für alle $f : \{0,1\}^n \to \mathbb{R}$, $\alpha \in \{0,1\}^k$ mit $0 \leq k \leq n$, für alle $x \in \{0,1\}^{n-k}$ und für ein zufälliges $y \in_{\mathcal{U}} \{0,1\}^k$ gilt*

$$f_\alpha(x) = \mathop{\mathrm{E}}_y[f(yx)\chi_\alpha(y)]$$

Beweis. Aus $\chi_{\gamma\beta}(yx) = \chi_\gamma(y) \cdot \chi_\beta(x)$ für $\beta \in \{0,1\}^{n-k}$ und $\gamma \in \{0,1\}^k$ folgt

$$\begin{aligned} f(yx) &= \sum_\gamma \sum_\beta \hat{f}(\gamma\beta)\chi_{\gamma\beta}(yx) \\ &= \sum_\gamma \sum_\beta \hat{f}(\gamma\beta)\chi_\gamma(y)\chi_\beta(x) \end{aligned}$$

Die Orthonormalität der Fourier-Basis besagt

$$\langle \chi_\gamma, \chi_\alpha \rangle = \begin{cases} 1 & \text{falls } \alpha = \gamma \\ 0 & \text{falls } \alpha \neq \gamma. \end{cases}$$

Daraus folgt

$$\begin{aligned} \mathop{\mathrm{E}}_y[f(yx)\chi_\alpha(y)] &= \mathop{\mathrm{E}}_y\left[\sum_\gamma \sum_\beta \hat{f}(\gamma\beta)\chi_\gamma(y)\chi_\beta(x)\chi_\alpha(y)\right] \\ &= \sum_\gamma \sum_\beta \hat{f}(\gamma\beta)\chi_\gamma(y) \underbrace{\mathop{\mathrm{E}}_y[\chi_\beta(x)\chi_\alpha(y)]}_{=\langle \chi_\gamma, \chi_\alpha \rangle} \\ &= \sum_\beta \hat{f}(\alpha\beta)\chi_\beta(x) \\ &= f_\alpha(x). \end{aligned}$$

□

Aus Lemma 10.12 folgt

$$\mathrm{E}[f_\alpha^2] = \mathop{\mathrm{E}}_x \left[\left(\mathop{\mathrm{E}}_y [f(yx)\chi_\alpha(y)] \right)^2 \right].$$

Mit Hilfe der Chernoff-Schranken läßt sich ein solcher Erwartungswert mit wenigen Mem-Fragen an f gut approximieren. Hierzu wählen wir zunächst ein bestimmt Anzahl von Suffixen $x_1, \ldots, x_{m_1} \in_{\mathcal{U}} \{0,1\}^{n-k}$, und für jedes x_i berechnen wir dann Approximationen A_i der inneren Erwartungswerte $\mathrm{E}_y[f(yx_i)\chi_\alpha(y)]$ mit der Hilfe von zusätzlich gewählten Präfixe $y_{i1}, \ldots, y_{im_2} \in_{\mathcal{U}} \{0,1\}^k$ und Mem-Fragen $f(y_{ij}x_i)$. Die Approximation des Erwartungswerts $\mathrm{E}[f_\alpha^2]$ ergibt sich dann als die gewichtete Summe der Approximationen A_i^2 der inneren Erwartungswerte. Auf die genaue Anzahl der benötigten Mem-Fragen werden wir im folgenden nicht eingehen, sondern setzen bei der folgenden Beschreibung der Funktion *Coeff* eine geeignete Wahl der Parameter m_1 und m_2 voraus.

Funktion *Coeff*(α, θ, δ) (probabilistische Version)

1. Wähle $x_1, \ldots, x_{m_1} \in_{\mathcal{U}} \{0,1\}^{n-k}$
2. Für alle $i = 1, \ldots, m_1$ wähle $y_{i,1}, \ldots, y_{i,m_2} \in_{\mathcal{U}} \{0,1\}^k$ und berechne
$$A_i = \frac{1}{m_2} \sum_{j=1}^{m_2} f(y_{ij}x_i)\chi_\alpha(y_{ij})$$
3. Berechne
$$B_\alpha = \frac{1}{m_1} \sum_{i=1}^{m_1} A_i^2$$
4. Falls $B_\alpha < \theta^2/2$, dann Ausgabe $\emptyset$.
5. Falls $B_\alpha < \theta^2/2$ und $|\alpha| = n$, dann Ausgabe $\{\alpha\}$.
6. Sonst, Ausgabe *Coeff*$(\alpha 0, \theta) \cup$ *Coeff*$(\alpha 1, \theta)$.

Lemma 10.13. *Für alle $f : \{0,1\}^n \to \{-1,+1\}$, und für alle $\theta, \delta > 0$ gibt Coeff$(\lambda, \theta, \delta)$ eine Menge $S \subseteq \{0,1\}^n$ aus, so dass mit Wahrscheinlichkeit $\geq 1 - \delta$ gilt:*

1. *Für alle $\alpha \in \{0,1\}^n$ mit $|\hat{f}(\alpha)| \geq \theta$ ist $\alpha \in S$.*

2. *$|S|$ ist polynomiell in $\frac{1}{\theta}$ beschränkt.*

Die Laufzeit von Coeff$(\lambda, \theta, \delta)$ ist polynomiell in n, $\frac{1}{\theta}$ und $\log \frac{1}{\delta}$ beschränkt, und Coeff benutzt Mem-Fragen an f.

Beweis. Mit Hilfe der Chernoff-Schranken können wir m_1 und m_2 polynomiell in n, $\frac{1}{\theta}$ und $\log \frac{1}{\delta}$ wählen, so dass für jede Approximation B_α von $\mathrm{E}[f_\alpha^2]$ mit Wahrscheinlichkeit wenigstens

$$1 - \frac{\delta\theta^2}{4n}$$

die beiden folgenden Implikationen gelten.

1. $\mathrm{E}[f_\alpha^2] < \theta^2/4 \implies B_\alpha < \theta^2/2$

2. $B_\alpha < \theta^2/2 \implies \mathrm{E}[f_\alpha^2] < \theta^2$

Mit dieser Wahl von m_1 und m_2 benötigen wir für jeden Aufruf höchstens polynomiell in viele Schritte und ebenso viele Mem-Fragen.

Zur Bestimmung der Laufzeit und der Ausgabemenge der probabilistischen Prozedur *Coeff* vergleichen wir die rekursiven Aufrufe *Coeff*(α, θ, δ) mit den rekursiven Aufrufen *Coeff*$_{det}(\alpha, \theta/2)$ der deterministischen Version, wobei zu Beginn jeweils $\alpha = \lambda$ gewählt wird.

Aus der Annahme, daß für alle Präfixe α mit Aufrufen *Coeff*$_{det}(\alpha, \theta/2)$ und *Coeff*(α, θ, δ) die erste Implikation $\mathrm{E}[f_\alpha^2] < \theta^2/4 \implies B_\alpha < \theta^2/2$ gilt, folgt mit Induktion über die Länge der Präfixe α, daß es für jeden Aufruf der probabilistischen Version auch einen entsprechenden Aufruf der deterministischen Version gibt, und daß ein α mit $|\alpha| = n$ nur dann von der probabilistischen Version ausgegeben wird, falls α auch von der deterministischen Version ausgegeben wird.

Es erfolgen höchstens $4n/\theta^2$ rekursive Aufrufe der deterministischen Version mit Parameter $\theta/2$. Damit ist die Wahrscheinlichkeit, daß es einen Präfix α mit Aufrufen $\mathit{Coeff}_{det}(\alpha, \theta/2)$ und $\mathit{Coeff}(\alpha, \theta, \delta)$ gibt, so daß die Implikation $\mathrm{E}[f_\alpha^2] < \theta^2/4 \implies B_\alpha < \theta^2/2$ verletzt ist, höchstens

$$\frac{4n}{\theta^2} \cdot \frac{\delta\theta^2}{4n} = \delta.$$

Also gilt mit Wahrscheinlichkeit $\geq 1 - \delta$, daß auch für die probabilistische Version höchstens $4n/\theta^2$ Aufrufe erfolgen, und daß $|\hat{f}(a)| \geq \theta$ für jedes Parity $\alpha \in \{0,1\}^n$ der Ausgabemenge $\mathit{Coeff}_{det}(\lambda, \theta/2)$ gilt.

In analoger Weise kann durch einen Vergleich von $\mathit{Coeff}(\alpha, \theta, \delta)$ und $\mathit{Coeff}_{det}(\alpha, \theta)$, und mit Hilfe der zweiten Implikation gezeigt werden, dass mit Wahrscheinlichkeit $\geq 1 - \delta$ alle $\alpha \in \{0,1\}^n$ mit $|\hat{f}(\alpha)| \geq \theta$ in der Ausgabemenge $\mathit{Coeff}(\alpha, \theta, \delta)$ enthalten sind. □

Algorithmus $\mathit{KM}(\theta, \varepsilon, \delta)$

1. Bestimme durch Aufruf von $\mathit{Coeff}(\lambda, \theta, \frac{\delta}{2})$ eine Menge $S \subseteq \{0,1\}^n$.

2. Gib die von $\mathit{Approx}(S, \frac{\delta}{2})$ produzierte Hypothese h aus.

Theorem 10.14. *Für alle Funktionen* $f : \{0,1\}^n \to \{-1,+1\}$, *und für alle* $\varepsilon, \delta, \theta > 0$ *gibt* $\mathit{KM}(\theta, \varepsilon, \delta)$ *eine Hypothese* h *mit Trägermenge* S *aus, so dass mit Wahrscheinlichkeit* $\geq 1 - \delta$ *gilt:*

1. *Für alle* $\alpha \in \{0,1\}^n$ *mit* $\hat{f}(\alpha) \geq \theta$ *ist* $\alpha \in S$

2. $|S|$ *ist polynomiell in* $\frac{1}{\theta}$ *beschränkt*

3. $\mathrm{E}[(f-h)^2] \leq 1 - \sum_{\alpha \in S} \hat{f}^2(\alpha) + \varepsilon$

Die Laufzeit von $\mathit{KM}(\theta, \varepsilon, \delta)$ *ist polynomiell in* n, $\frac{1}{\theta}$, $\frac{1}{\varepsilon}$ *und* $\log \frac{1}{\delta}$ *beschränkt, und KM benutzt Mem-Fragen an* f.

Beweis. Folgt unmittelbar aus den Eigenschaften der Funktionen *Approx* und *Coeff* (Lemma 10.7 und Lemma 10.13). □

10.5 Entscheidungsbäume und kleine L_1-Norm

Ein Entscheidungsbaum über der Variablenmenge $\{x_1, \ldots, x_n\}$ ist ein Binärbaum, wobei die inneren Knoten mit einer Variablen und die Blätter mit 0 oder 1 markiert sind. Jede Belegung der Variablen $x_1, \ldots, x_n$ bestimmt eindeutig einen Pfad von der Wurzel zu einem Blatt, wobei bei jedem inneren Knoten mit Markierung x_i der linke (bzw. rechte) Nachfolger gewählt wird, falls $x_i = 0$ (bzw. $x_i = 1$). Der Wert des Entscheidungsbaums für eine Belegung von $x_1, \ldots, x_n$ ist dann die Markierung des Blattes.

Jede 1-Entscheidungsliste entspricht einem Entscheidungsbaum, und es ist nicht schwer zu sehen, daß jede Funktion mit einem Entscheidungsbaum mit m Knoten eine äquivalente DNF-Formel der Größe $O(m^2)$ hat.

Definition. *Die* L_1-Norm *einer Funktion* $f : \{0,1\}^n \to \mathbb{R}$ *ist gegeben durch*

$$L_1(f) = \sum_{\alpha \in \{0,1\}^n} |\hat{f}(\alpha)|.$$

Lemma 10.15. *Sei* $f, g : \{0,1\}^n \to \mathbb{R}$*. Dann gilt*

1. $L_1(f) + L_1(g) \geq L_1(f+g)$
2. $L_1(f) \cdot L_1(g) \geq L_1(f \cdot g)$

Beweis. ... □

Theorem 10.16. *Sei f eine Funktion mit einem Entscheidungsbaum mit m Blättern. Dann gilt $L_1(f) \leq m$.*

Beweis. Für $d \geq 1$, $b = b_1 \ldots b_d$, und $x = x_1 \ldots x_d$ sei

$$Eq_{(b_1 \ldots b_d)}(x) = \begin{cases} 1 & \text{falls } b_i = x_i \text{ für alle } i \in \{1, ..., d\} \\ 0 & \text{sonst} \end{cases}$$

Dann gilt

$$Eq_{(b_1 \ldots b_d)}(x) = \prod_{i=1}^{d} \frac{1 + (-1)^{b_i} \chi_{\{i\}}(x)}{2}$$

und somit

$$L_1(Eq_{(b_1 \ldots b_d)}(x)) \leq \prod_{i=1}^{d} \underbrace{L_1\left(\frac{1+(-1)^{b_i}\chi_{\{i\}}(x)}{2}\right)}_{\leq 1} \leq 1.$$

Für ein Blatt v im Entscheidungsbaum von f sei $d(v)$ die Anzahl der inneren Knoten zum Blatt v (mit der Wurzel), $x_{i_1}, \ldots, x_{i_{d(v)}}$ die Variablen in diesem Pfad, $b^v_{i_1}, \ldots, b^v_{i_{d(v)}}$ die eindeutige Belegung der Variablen $x_{i_1}, \ldots, x_{i_{d(v)}}$, die zum Blatt v führt, und σ_v die Markierung von v. Dann lässt sich f mit Hilfe der Funktion Eq durch

$$f(x) = \sum_{\text{Blätter } v} \sigma_v \cdot E_{b^v_{i_1}, \ldots, b^v_{i_{d(v)}}}(x_{i_1}, \ldots, x_{i_{d(v)}})$$

darstellen. Es folgt

$$L_1(f) \leq \sum_v \underbrace{L_1(\sigma_v)}_{\leq 1} \cdot \underbrace{L_1(Eq)}_{\leq 1} \leq m.$$

$\square$

Theorem 10.17. *Sei $\mathcal{C} = \bigcup \mathcal{C}_n$ eine Konzeptklasse, so dass es ein Polynom p mit $L_1(f) \leq p(n,s)$ für alle $f \in \mathcal{C}_n$ der Größe s und $n \geq 0$ gibt. Dann ist $\mathcal{C}$ bezüglich der uniformen Verteilung mit Mem-Fragen und mit einer p-evaluierbaren Hypothesenklasse effizient PAC-erlernbar.*

Beweis. Für alle $f \in \mathcal{C}_n$ liefert $KM(\theta, \varepsilon/2, \delta)$ mit $\theta = \varepsilon/(2p(n,s))$ mit Wahrscheinlichkeit $\geq 1-\delta$ eine Hypothese h mit Trägermenge S, so dass gilt:

1. $\alpha \notin S \implies |\hat{f}(\alpha)| \leq \dfrac{\varepsilon}{2p(\alpha,s)} \leq \dfrac{\varepsilon}{2L_1(f)}$
2. $\mathrm{E}[(f-h)^2] \leq 1 - \sum_{\alpha \in S} \hat{f}^2(\alpha) + \dfrac{\varepsilon}{2}$

Es folgt

$$
\begin{aligned}
1 - \sum_{\alpha \in S} \hat{f}^2(\alpha) &= \sum_{\alpha \notin S} \hat{f}^2(\alpha) \\
&\leq \max_{\alpha \notin S}(\hat{f}(\alpha)) \cdot \sum_{\alpha \notin S} |\hat{f}(\alpha)| \\
&\leq \frac{\varepsilon}{2L_1(f)} L_1(f) \\
&= \frac{\varepsilon}{2}
\end{aligned}
$$

und somit gilt für $x \in_{\mathcal{U}} \{0,1\}^n$,

$$\Pr[f(x) \neq sign(h(x))] \leq \varepsilon.$$

Die Laufzeit von $KM(\theta, \varepsilon/2, \delta)$ ist mit $\theta = \varepsilon/(2p(n,s))$ polynomiell in n, s, $1/\varepsilon$ und $\log \frac{1}{\delta}$. Also können wir $\mathcal{C}$ durch Berechnung von h und Ausgabe $sign(h)$ effizient PAC-erlernen. □

Unmittelbar aus Theorem 10.16 und Theorem 10.17 können wir auf die Erlernbarkeit von Entscheidungsbäumen schließen.

Korollar 10.18. *Entscheidungsbäume sind mit Mem-Fragen und einer p-evaluierbaren Hypothesenklasse bezüglich der uniformen Verteilung effizient PAC-erlernbar.*

10.6 Schwache Erlernbarkeit von DNF-Formeln

Definition. *Eine Konzeptklasse $\mathcal{C} = \bigcup \mathcal{C}_n$ ist schwach PAC-erlernbar, falls der Lernalgorithmus nur für $\varepsilon = \frac{1}{2} - \frac{1}{p(n,s)}$ das Lernkriterium erfüllt, wobei p ein Polynom und s die Größe des Zielkonzepts bezeichet.*

Lemma 10.19. *Sei f eine DNF-Formel mit einem Term T der Größe t. Dann gibt es ein $\alpha \in \{0,1\}^n$ mit $|\hat{f}(\alpha)| \geq 2^{-t}$.*

Beweis. O.B.d.A. nehmen wir an, daß T monoton ist, d.h. T besteht nur aus positiven Variablen $x_1, \ldots, x_{i_t}$. Sei $V = \{i_1, \ldots, i_t\}$ die entsprechende Menge der Indizes. Wir betrachten zunächst die $\{0,1\}$-wertige Version T' von T,

$$T'(x) = \frac{T(x)+1}{2}.$$

Mit $x = x_1 \dots x_n$ und wegen $x_i = (1 - \chi_{\{i\}}(x))/2$ gilt

$$T'(x) = \prod_{i \in V} x_i = \prod_{i \in V} \frac{1 - \chi_{\{i\}}(x)}{2} = 2^{-t} \prod_{i \in V} (1 - \chi_{\{i\}}(x)).$$

Außerdem ist $\chi_A(x)\chi_B(x) = \chi_{A \cup B}(x)$ für alle disjunkten Teilmengen $A, B \subseteq \{1, \dots, n\}$, und damit läßt sich das Produkt $\prod_{i \in V}(1 - \chi_{\{i\}}(x))$ durch Ausmultiplizieren in die Summe $\sum_{A \subseteq V} (-1)^{|A|} \chi_A(x)$ umformen. Es folgt

$$T'(x) = 2^{-t} \sum_{A \subseteq V} (-1)^{|A|} \chi_A(x),$$

und da die Anzahl aller Teilmengen $A \subseteq V$ gerade 2^t ist, erhalten wir für eine zufällige und unter Gleichverteilung gewählte Teilmenge $A \subseteq V$,

$$T'(x) = \mathop{\mathrm{E}}_{A \subseteq V}[(-1)^{|A|} \chi_A(x)].$$

Betrachten wir nun den Erwartungswert $\mathrm{E}_x[f(x)T'(x)]$. Durch Einsetzten des Erwartungswerts für T' erhalten wir

$$\begin{aligned}
\mathop{\mathrm{E}}_x[f(x)T'(x)] &= \mathop{\mathrm{E}}_x[f(x) \mathop{\mathrm{E}}_{A \subseteq V}[(-1)^{|A|} \chi_A(x)]] \\
&= \mathop{\mathrm{E}}_{A \subseteq V}[(-1)^{|A|} \mathop{\mathrm{E}}_x[f(x)\chi_A(x))]] \\
&\leq \mathop{\mathrm{E}}_{A \subseteq V}[| \underbrace{\mathop{\mathrm{E}}_x[f(x) \cdot \chi_A(x)]}_{=\hat{f}(A)} |] \\
&= \mathop{\mathrm{E}}_{A \subseteq V}[|\hat{f}(A)|].
\end{aligned}$$

Auf der anderen Seite ist T ein Term der DNF-Formel f. Aus $f(x) = -1$ folgt daher $T'(x) = 0$ und somit $f(x)T'(x) = T'(x)$. Natürlich gilt die Gleichung $f(x)T'(x) = T'(x)$ auch im Fall $f(x) = 1$, und somit gilt

$$\mathop{\mathrm{E}}_x[f(x)T'(x)] = \mathop{\mathrm{E}}_x[T'(x)].$$

Für den $\{0, 1\}$-wertigen Term T' ist $\mathrm{E}_x[T'(x)] = \Pr[T'(x) = 1]$. Da T ein Term mit t Variablen ist, ist die Wahrscheinlichkeit für $T'(x) = 1$ gerade 2^{-t}, und es folgt

$$\mathop{\mathrm{E}}_x[f(x)T'(x)] = 2^{-t}.$$

Also erhalten wir

$$\mathop{\mathrm{E}}_{A\subseteq V}[|\hat{f}(A)|] \geq 2^{-t}$$

für eine zufällige Teilmenge $A \subseteq V$, und daraus können wir auf die Existenz einer Teilmenge $A \subseteq V$ mit $|\hat{f}(A)| \geq 2^{-t}$ schließen. $\square$

Theorem 10.20. *DNF-Formeln sind effizient schwach PAC-erlernbar mit Mem-Fragen und mit einer p-evaluierbaren Hypothesenklasse bezüglich der uniformen Verteilung.*

Beweis. Sei f eine DNF-Formel $T_1 \vee \cdots \vee T_m$ der Größe s. Für einen Term T_i mit mehr als t Variablen ist die Wahrscheinlichkeit für $T_i(x) = 1$ mit $x \in_{\mathcal{U}} \{0,1\}^n$ kleiner als 2^{-t}. Falls alle Terme T_i von f mehr als $t = \log(3s)$ Variable haben, gilt also

$$\begin{aligned}\Pr[f(x) = 1] &= \Pr[T_1(x) = 1 \vee T_2(x) = 1 \vee \cdots \vee T_m(x) = 1] \\ &\leq \sum_{i=1}^{m} \Pr[T_i(x) = 1] < m \cdot 2^{-t} \leq \frac{1}{3},\end{aligned}$$

d.h. die konstante -1-Funktion approximiert f mit Fehler $\Pr[f(x) \neq -1] = \Pr[f(x) = 1] < 1/3$.

Mit den Chernoff-Schranken und mit Hilfe von zufälligen Beispielen $(x, f(x))$ mit $x \in_{\mathcal{U}} \{0,1\}^n$ können wir eine Approximation $\tilde{P}$ für $P = \Pr[f(x) = 1]$ berechnen, so daß mit hoher Wahrscheinlichkeit $|\tilde{P} - P| \leq 1/24$ gilt.

Falls $\tilde{P} \leq 9/24$ ist, dann gilt $P \leq \tilde{P} + 1/24 \leq 5/12$, und die Hypothese $h(x) = -1$ hat einen einen (für die schwache Erlernbarkeit) genügend kleinem Fehler.

Andernfalls, d.h. falls $\tilde{P} > 9/24$ ist, dann folgt aus $P \geq \tilde{P} - 1/24 > 1/3$, daß f einen Term mit höchstens $\log(3s)$ Variablen hat. Daher gibt es ein $\alpha \in \{0,1\}^n$ mit $|\hat{f}(\alpha)| \geq \frac{1}{3s}$, und ein solches α können wir mit Hilfe der Funktion *Coeff* mit Parameter $\theta = 1/(3s)$ effizient berechnen. Durch Approximation des Koeffizienten $\hat{f}(\alpha)$ mit einem Fehler kleiner als $1/(3s)$ können wir dann mit hoher Wahrscheinlichkeit das Vorzeichen $sign(\hat{f}(\alpha))$ bestimmen.

Betrachten wir nun den Fehler der Hypothese $h(x) = sign(\hat{f}(\alpha))\chi_\alpha(x)$. Für alle $\alpha \in \{0,1\}^n$ gilt

$$\begin{aligned}
\hat{f}(\alpha) &= \frac{1}{2^n}\sum_x f(x)\chi_\alpha(x) \\
&= \frac{1}{2^n}(|\{x \mid f(x) = \chi_\alpha(x)\}| - |\{x \mid f(x) \neq \chi_\alpha(x)\}|) \\
&= \Pr_x[f(x) = \chi_\alpha(x)] - \Pr_x[f(x) \neq \chi_\alpha(x)] \\
&= 1 - 2\Pr_x[f(x) \neq \chi_\alpha(x)],
\end{aligned}$$

und somit

$$\Pr[f(x) \neq \chi_\alpha(x)] = \frac{1-\hat{f}(\alpha)}{2}.$$

Falls $\hat{f}(\alpha)$ positiv ist, dann ist $h(x) = \chi_\alpha(x)$ und $\hat{f}(\alpha) \geq 1/(3s)$, und es folgt

$$\Pr[f(x) \neq h(x)] = \frac{1}{2} - \frac{\hat{f}(\alpha)}{2} \leq \frac{1}{2} - \frac{1}{6s}.$$

Falls $\hat{f}(\alpha)$ negativ ist, dann ist $h(x) = -\chi_\alpha(x)$ und $\hat{f}(\alpha) \leq -1/(3s)$, und es folgt entsprechend

$$\Pr[f(x) \neq h(x)] = 1 - \left(\frac{1-\hat{f}(\alpha)}{2}\right) \leq \frac{1}{2} - \frac{1}{6s}.$$

Also ist der Fehler der Hypothese h in beiden Fällen höchstens $1/2 - 1/(6s)$. Die benötigte Zeit zur Berechnung von h ist polynomiell in n, s und $\log\frac{1}{\delta}$. $\square$