

# 1 Funkční závislosti stanovené z dat

Pro danou relaci  $\mathcal{D}$  chceme najít, co možná nejmenší teorii  $T$  tak, že  $\mathcal{D} \models A \Rightarrow B$  právě, když  $T \models A \Rightarrow B$ .

**Definice 1.** Teorie  $T$  se nazývá báze  $\mathcal{D}$ , pokud pro každou  $A \Rightarrow B$  platí  $\mathcal{D} \models A \Rightarrow B$  p.k.  $T \models A \Rightarrow B$ .

**Poznámka.** Bázi  $\mathcal{D}$  je obecně hodně. Např. pokud  $T$  je báze  $\mathcal{D}$  a navíc  $T \models A \Rightarrow B$  pro nějakou  $A \Rightarrow B \notin T$ , pak  $T \cup \{A \Rightarrow B\}$  je opět báze.

Z definice báze je zřejmé, že budeme-li mít dvě báze, budou mít stejné sémantické důsledky, je proto žádoucí si takový jev pojmenovat.

**Definice 2.** Teorie  $T_1$  a  $T_2$  jsou sémanticky ekvivalentní, značeno  $T_1 \equiv T_2$ , jestliže pro libovolnou  $A \Rightarrow B$  platí  $T_1 \models A \Rightarrow B$  právě, když  $T_2 \models A \Rightarrow B$ .

Sémanticky ekvivalentní teorie, pak mají úzký vztah k pojmu model teorie.

**Věta 1** (o charakterizaci sémantické ekvivalence). *Následující tvrzení jsou ekvivalentní:*

1.  $T_1 \equiv T_2$ ,
2.  $\text{Mod}(T_1) = \text{Mod}(T_2)$ ,
3.  $\text{Mod}_C(T_1) = \text{Mod}_C(T_2)$ ,
4. Pro libovolnou  $A \subseteq R$  máme  $[A]_{T_1} = [A]_{T_2}$ .

*Důkaz.*  $1 \Rightarrow 2$ : Pro libovolnou  $A \Rightarrow B$  máme  $\text{Mod}(T_1) \models A \Rightarrow B$  p.k.  $T_1 \models A \Rightarrow B$  p.k.  $T_2 \models A \Rightarrow B$  p.k.  $\text{Mod}(T_1) \models A \Rightarrow B$ .

$2 \Rightarrow 3$ : Speciální případ.

$3 \Rightarrow 4$ : Stejné uzávěrové systémy mají stejné uzávěrové operátory.

$4 \Rightarrow 1$ : Pro libovolnou  $A \Rightarrow B$  máme  $T_1 \models A \Rightarrow B$  p.k.  $B \subseteq [A]_{T_1} = [A]_{T_2}$  p.k.  $T_2 \models A \Rightarrow B$ .  $\square$

**Důsledek.** Pokud jsou  $T_1$  a  $T_2$  báze  $\mathcal{D}$ , pak  $T_1 \equiv T_2$ .

Pro snadnější charakterizaci pravdivosti v relaci si zavedeme operátor, který bude fungovat podobně jako sémantický uzávěr u teorie. Nejdříve však definujeme relaci na n-ticích.

**Definice 3.** Pro  $\mathcal{D} \subseteq \prod_{y \in R} D_y$  a  $M \subseteq R$  definujeme  $E_{\mathcal{D}} : 2^R \rightarrow 2^{\mathcal{D} \times \mathcal{D}}$  předpisem

$$E_{\mathcal{D}}(M) = \{\langle t, t' \rangle \in \mathcal{D} \times \mathcal{D} \mid t(M) = t'(M)\}.$$

**Poznámka.** • Z definice  $E_{\mathcal{D}}$  je hned zřejmé, že relace  $E_{\mathcal{D}}(M)$  je ekvivalencí na  $\mathcal{D}$  což také znamená, že můžeme udělat rozklad.

- Význam vztahu  $E_{\mathcal{D}}(M) \subseteq E_{\mathcal{D}}(M')$  je, že všechny dvojice n-tic, které se rovnají na  $M$  se také rovnají na  $M'$ .
- $E_{\mathcal{D}}$  je zřejmě antinonní, protože pokud  $M_1 \subseteq M_2$ , všechny n-tice, které se rovnají na  $M_2$  se tím spíš musí rovnat na  $M_1$ , tedy  $E_{\mathcal{D}}(M_2) \subseteq E_{\mathcal{D}}(M_1)$ .

**Definice 4.** Pro  $\mathcal{D} \subseteq \prod_{y \in R} D_y$  a  $M \subseteq R$  definujeme  $C_{\mathcal{D}} : 2^R \rightarrow 2^R$  předpisem

$$C_{\mathcal{D}}(M) = \{y \in R \mid E_{\mathcal{D}}(M) \subseteq E_{\mathcal{D}}(\{y\})\}.$$

**Poznámka.**  $C_{\mathcal{D}}(M)$  je vlastně množina atributů, na kterých jsou si rovny všechny dvojice n-tic z  $\mathcal{D}$ , které jsou si rovny na  $M$ . Důsledkem pak je, že  $E_{\mathcal{D}}(M) \subseteq E_{\mathcal{D}}(C_{\mathcal{D}}(M))$ . Důkaz je ponechán čtenáři.

**Věta 2.**  $C_{\mathcal{D}}$  je uzávěrový operátor na  $R$ .

*Důkaz.* • (*extenzivita*): Pokud  $y \in M$ , pak  $E_{\mathcal{D}}(M) \subseteq E_{\mathcal{D}}(\{y\})$ , protože pokud jsou si  $t, t'$  rovny na všech attributech z  $M$ , tím spíš jsou si rovny na  $y \in M$ . Odtud dle definice  $C_{\mathcal{D}}$  dostáváme  $y \in C_{\mathcal{D}}(M)$ .

- (*monotonie*): Předpokládejme  $M_1 \subseteq M_2$  a vezmeme  $y \in C_{\mathcal{D}}(M_1)$ . Poslední znamená, že  $E_{\mathcal{D}}(M_1) \subseteq E_{\mathcal{D}}(\{y\})$ . Z antitonie  $E_{\mathcal{D}}$  dostáváme  $E_{\mathcal{D}}(M_2) \subseteq E_{\mathcal{D}}(M_1) \subseteq E_{\mathcal{D}}(\{y\})$ . Z definice  $C_{\mathcal{D}}$  je  $y \in C_{\mathcal{D}}(M_2)$ .
- (*idempotence*):  $C_{\mathcal{D}}(M) \subseteq C_{\mathcal{D}}(C_{\mathcal{D}}(M))$  platí z extenzivity. Pro obrácenou inkluzi máme následující posloupnost argumentů:

$$\begin{aligned} E_{\mathcal{D}}(M) &\subseteq E_{\mathcal{D}}(C_{\mathcal{D}}(M)) \\ \{y \in R \mid E_{\mathcal{D}}(C_{\mathcal{D}}(M)) \subseteq E_{\mathcal{D}}(\{y\})\} &\subseteq \{y \in R \mid E_{\mathcal{D}}(M) \subseteq E_{\mathcal{D}}(\{y\})\} \\ C_{\mathcal{D}}(C_{\mathcal{D}}(M)) &\subseteq C_{\mathcal{D}}(M) \end{aligned}$$

□

**Věta 3** (o charakterizaci pravdivosti). *Následující jsou ekvivalentní:*

1.  $\mathcal{D} \models A \Rightarrow B$
2.  $E_{\mathcal{D}}(A) \subseteq E_{\mathcal{D}}(B)$
3.  $B \subseteq C_{\mathcal{D}}(A)$

*Důkaz.* 1  $\Rightarrow$  2: Z definice  $\mathcal{D} \models A \Rightarrow B$ , pokud  $t(A) = t'(A)$ , pak  $t(B) = t'(B)$ , tzn. pokud  $\langle t, t' \rangle \in E_{\mathcal{D}}(A)$ , pak  $\langle t, t' \rangle \in E_{\mathcal{D}}(B)$ , tj.  $E_{\mathcal{D}}(A) \subseteq E_{\mathcal{D}}(B)$ .

2  $\Rightarrow$  3: Předpokládejme  $E_{\mathcal{D}}(A) \subseteq E_{\mathcal{D}}(B)$ . Pro libovolný  $y \in B$  pak z antitonie platí  $E_{\mathcal{D}}(A) \subseteq E_{\mathcal{D}}(B) \subseteq E_{\mathcal{D}}(\{y\})$ . To podle definice  $C_{\mathcal{D}}$  znamená, že  $y \in C_{\mathcal{D}}(A)$ , tj.  $B \subseteq C_{\mathcal{D}}(A)$ .

3  $\Rightarrow$  1: Předpokládejme  $B \subseteq C_{\mathcal{D}}(A)$ . Dále mějme  $t, t' \in \mathcal{D}$  takové, že  $t(A) = t'(A)$  a vezmeme libovolné  $y \in B$ . Pak nutně  $\langle t, t' \rangle \in E_{\mathcal{D}}(A)$  a navíc  $E_{\mathcal{D}}(A) \subseteq E_{\mathcal{D}}(\{y\})$ . Důsledkem je, že  $t(y) = t'(y)$ , tedy  $t(B) = t'(B)$ . □

**Věta 4** (o charakterizaci báze).  *$T$  je báze  $\mathcal{D}$  právě, když pro libovolné  $A \subseteq R$  máme  $C_{\mathcal{D}}(A) = [A]_T$ .*

**Poznámka.** Ekvivalentně  $C_{\mathcal{D}}(M) = [M]_T = M_T^{\infty} = M_T^+$ .

*Důkaz.* " $\Rightarrow$ ": Nechť  $T$  je báze  $\mathcal{D}$ . Pak  $[M]_T \subseteq [M]_T$  p.k.  $T \models M \Rightarrow [M]_T$  p.k.  $\mathcal{D} \models M \Rightarrow [M]_T$  p.k.  $[M]_T \subseteq C_{\mathcal{D}}(M)$ . Obráceně máme  $C_{\mathcal{D}}(M) \subseteq [M]_T$  p.k.  $\mathcal{D} \models M \Rightarrow C_{\mathcal{D}}(M)$  p.k.  $T \models M \Rightarrow C_{\mathcal{D}}(M)$  p.k.  $C_{\mathcal{D}}(M) \subseteq [M]_T$ . Dohromady tedy  $C_{\mathcal{D}}(M) = [M]_T$ .

" $\Leftarrow$ ": Pokud  $C_{\mathcal{D}}$  má stejné pevné body jako  $[\dots]_T$ , pak  $\mathcal{D} \models A \Rightarrow B$  p.k.  $B \subseteq C_{\mathcal{D}}(A) = [A]_T$  p.k.  $T \models A \Rightarrow B$ .  $\square$

Následující věta ukazuje, že pro libovolnou relaci existuje minimálně jedna báze.

**Věta 5** (o existenci báze).  *$T = \{A \Rightarrow C_{\mathcal{D}}(A) \mid A \subseteq R\}$  je báze  $\mathcal{D}$ .*

*Důkaz.* Dle předchozí věty stačí ověřit, že  $C_{\mathcal{D}}(M) = [M]_T$  pro libovolnou  $M$ , tzn. ověřit, že  $M = C_{\mathcal{D}}(M)$  právě když  $M \in \mathcal{M}_T$ .

" $\Rightarrow$ ": Předpokládejme  $M = C_{\mathcal{D}}(M)$  a vezmeme libovolnou  $A \Rightarrow C_{\mathcal{D}}(A) \in T$  tak, že  $A \subseteq M$ . Z monotonie operátoru  $C_{\mathcal{D}}$  dostaneme  $C_{\mathcal{D}}(A) \subseteq C_{\mathcal{D}}(M) = M$ . Dohromady tedy  $\mathcal{D}_M \models A \Rightarrow C_{\mathcal{D}}(A)$  p.k.  $\mathcal{D}_M \in \text{Mod}_C(T)$  p.k.  $M \in \mathcal{M}_T$ .

" $\Leftarrow$ ": Předpokládejme, že  $M \in \mathcal{M}_T$ . To jest  $\mathcal{D}_M \in \text{Mod}_C(T)$ . Speciálně pro  $M \Rightarrow C_{\mathcal{D}}(M) \in T$  máme  $\mathcal{D}_M \models M \Rightarrow C_{\mathcal{D}}(M)$ . Odtud  $C_{\mathcal{D}}(M) \subseteq M$  a přidáme-li extenzivitu  $C_{\mathcal{D}}$  dostaneme  $C_{\mathcal{D}}(M) = M$ .  $\square$

Když už víme, že báze vždy existuje, přesuneme pozornost na její velikost vzhledem k počtu FZ. Z předchozího textu vyplývá, že se budeme snažit najít bázi ekvivalentní, ale co nejmenší.

První ideou je odstranit z teorie nějaké FZ tak, že je pořád bází. Pokud už nejde zmenšit je tzv. neredundantní.

**Definice 5.** Teorie  $T$  je neredundantní báze relace  $\mathcal{D}$ , pokud  $T$  je báze  $\mathcal{D}$  a pro každou  $T' \subset T$  platí, že  $T'$  není báze  $\mathcal{D}$ .

Tato definice má i ekvivalentní formulaci, která vede na jednoduchý algoritmus transformace báze na bázi neredundantní.

**Věta 6** (o charakterizaci neredundantní báze). *Teorie  $T$  je neredundantní báze relace  $\mathcal{D}$  právě, když  $T$  je báze a žádná  $A \Rightarrow B \in T$  sémanticky neplyne z  $T \setminus \{A \Rightarrow B\}$ .*

*Důkaz.* " $\Rightarrow$ ": Vezmeme libovolnou  $A \Rightarrow B \in T$ . Z předpokladu, že  $T$  je báze, vyplývá, že  $T \setminus \{A \Rightarrow B\}$  není báze, a tedy není ekvivalentní  $T$ . Pak existuje model  $T \setminus \{A \Rightarrow B\}$ , který není modelem  $T$ , a tedy není v něm pravdivá  $A \Rightarrow B$ , což znamená, že  $T \setminus \{A \Rightarrow B\} \not\models A \Rightarrow B$ .

" $\Leftarrow$ ": Vezmeme libovolnou  $T' \subset T$ . Pak nutně existuje  $A \Rightarrow B \in T$  tak, že  $A \Rightarrow B \notin T'$ , a díky předpokladu máme  $T' \not\models A \Rightarrow B$ . Tudíž  $T$  a  $T'$  nejsou sémanticky ekvivalentní.  $\square$

K tomu, abychom definovali konkrétní neredundantní bázi využijeme následující množinu.

**Definice 6.** Pro relaci  $\mathcal{D}$  nad  $R$  uvažujeme množinu  $\mathcal{P}_{\mathcal{D}} \subseteq 2^R$ , která je definovaná předpisem:

$$\mathcal{P}_{\mathcal{D}} = \{P \neq C_{\mathcal{D}}(P) \mid \forall Q \in \mathcal{P}_{\mathcal{D}} : \text{pokud } Q \subset P, \text{ pak } C_{\mathcal{D}}(Q) \subseteq P\}.$$

Prvkům z této množiny se někdy říká pseudo-uzávěry. Koresponduje to s tím, že to nejsou sice uzavřené množiny, ale mají uzavěrovou vlastnost vzhledem ke všem ostatním prvkům z této množiny.

**Definice 7.** GD bázi  $\mathcal{D}$  nazveme teorií definovanou následujícím předpisem:

$$GD(\mathcal{D}) = \{P \Rightarrow C_{\mathcal{D}}(P) \mid P \in \mathcal{P}_{\mathcal{D}}\}.$$

**Poznámka.** Teorie je nazvaná podle francouzských vědců Guigues a Duquenne, kteří ji prvně definovali v kontextu formální konceptuální analýzy.

**Věta 7.** *GD báze relace  $\mathcal{D}$  je bázi  $\mathcal{D}$ .*

*Důkaz.*  $GD(\mathcal{D})$  je báze právě, když  $C_{\mathcal{D}}$  a  $[\dots]_{GD(\mathcal{D})}$  mají stejné pevné body, tedy  $M = C_{\mathcal{D}}(M)$  právě, když  $M \in \mathcal{M}_{GD(\mathcal{D})}$  pro každou  $M \subseteq R$ .

" $\Rightarrow$ ": Víme, že  $T' = \{A \Rightarrow C_{\mathcal{D}}(A) \mid A \subseteq R\}$  je báze a vidíme, že  $T \subseteq T'$ , z čehož vyplývá, že  $\mathcal{M}_{T'} \subseteq \mathcal{M}_T$ . Necht'  $M = C_{\mathcal{D}}(M)$ . Pak  $M \in \mathcal{M}_{T'}$ , a tedy z předchozího  $M \in \mathcal{M}_T$ .

" $\Leftarrow$ ": Necht'  $M \in \mathcal{M}_T$ , pak  $\mathcal{D}_M \models P \Rightarrow C_{\mathcal{D}}(P)$  pro každou  $P \in \mathcal{P}_{\mathcal{D}}$ , což znamená, že pokud  $P \subseteq M$ , pak  $C_{\mathcal{D}}(P) \subseteq M$ . Nyní sporem dokážeme, že  $M = C_{\mathcal{D}}(M)$ . Necht' tedy  $M \neq C_{\mathcal{D}}(M)$ . Pak díky předchozímu je  $M \in \mathcal{P}_{\mathcal{D}}$ , a tedy z předpokladu plyne, že  $\mathcal{D}_M \models M \Rightarrow C_{\mathcal{D}}(M)$ . Jelikož ale  $M \subseteq M$ , pak  $C_{\mathcal{D}}(M) \subseteq M$ . Navíc z extenzivity uzavěrového operátoru  $C_{\mathcal{D}}$  máme  $M \subseteq C_{\mathcal{D}}(M)$ , což dohromady dává  $M = C_{\mathcal{D}}(M)$ .  $\square$

**Věta 8.** *GD báze relace  $\mathcal{D}$  je neredundantní bázi  $\mathcal{D}$ .*

*Důkaz.* Vezmeme libovolnou  $T \subset GD(\mathcal{D})$  a ukážeme, že  $T$  není báze. Díky předpokladu existuje  $P \Rightarrow C_{\mathcal{D}}(P) \in GD(\mathcal{D})$ , která není v  $T$ . Z definice  $\mathcal{P}_{\mathcal{D}}$  je vidět, že  $\mathcal{D}_P$  je modelem  $T$ . Jelikož však není modelem  $GD(\mathcal{D})$ , nemohou být  $T$  a  $GD(\mathcal{D})$  sémanticky ekvivalentní, což dohromady s tím, že  $GD(\mathcal{D})$  je báze  $\mathcal{D}$  dává, že  $T$  není báze  $\mathcal{D}$ .  $\square$

Jako motivaci pro zbytek sekce vezmeme následující příklad.

**Příklad 1.** Mějme  $R = \{a, b, c\}$  a teorie  $T_1 = \{\{a\} \Rightarrow \{b, c\}\}$  a  $T_2 = \{\{a\} \Rightarrow \{b\}, \{a\} \Rightarrow \{c\}\}$ . Při bližším prozkoumání zjistíme, že  $T_1 \equiv T_2$  a navíc, že jsou obě neredundantní, ale  $|T_1| < |T_2|$ .

Z příkladu je patrné, že kompaktnější verze teorií nelze získat pouze odstraněním redundantních FZ. Jelikož chceme najít teorii s nejmenší kardinalitou, definujeme následující pojem.

**Definice 8.** Pokud teorie  $T$  je báze  $\mathcal{D}$  a pro libovolnou  $T'$ , která je také bází  $\mathcal{D}$ , platí, že  $|T| \leq |T'|$ , pak se  $T$  nazývá minimální báze  $\mathcal{D}$ .

Ukazuje se, že GD báze je také minimální bází, ale abychom byli schopni to potvrdit, je potřeba prozkoumat vzájemné vlastnosti mezi množinou  $\mathcal{P}_{\mathcal{D}}$  a operátorem  $C_{\mathcal{D}}$ .

**Poznámka.** Fakt, že  $|T| \leq |T'|$  intuitivně chápeme, ale přesnou matematickou definicí je, že existuje injektivní zobrazení  $f : T \rightarrow T'$ . Pro připomenutí, zobrazení  $f$  je injektivní, jestliže pro každé  $x_1, x_2 \in T$  platí, že pokud  $f(x_1) = f(x_2)$ , pak  $x_1 = x_2$ , nebo-li neexistují dva prvky, které se zobrazí na to samé, a tedy pro konečné množiny platí, že  $T$  má menší nebo stejný počet prvků jako  $T'$ . Této definice využijeme u důkazu, že GD báze je minimální.

Jednou ze základních vlastností pseudo-uzávěrů je, že přidáme-li jeden z nich do uzávěrového systému generovaným operátorem  $C_{\mathcal{D}}$ , pak je výsledná množina zase uzávěrovým systémem.

**Věta 9.** Pokud  $P \in \mathcal{P}_{\mathcal{D}}$  a  $A \subseteq R$  tak, že  $P \not\subseteq C_{\mathcal{D}}(A)$ , pak  $C_{\mathcal{D}}(A) \cap P = C_{\mathcal{D}}(C_{\mathcal{D}}(A) \cap P)$ .

*Důkaz.* Předpokládejme, že  $P \not\subseteq C_{\mathcal{D}}(A)$  a tedy  $P \not\subseteq C_{\mathcal{D}}(A) \cap P$ . Nyní stačí ukázat, že  $\mathcal{D}_{C_{\mathcal{D}}(A) \cap P}$  je modelem  $T$ , z čehož pak plyne  $C_{\mathcal{D}}(A) \cap P \in \mathcal{M}_T$  a tedy, že  $C_{\mathcal{D}}(A) \cap P = C_{\mathcal{D}}(C_{\mathcal{D}}(A) \cap P)$ .

Vezmeme libovolnou  $Q \in \mathcal{P}_{\mathcal{D}}$  tak, že  $Q \subseteq C_{\mathcal{D}}(A) \cap P$  a prokážeme, že  $C_{\mathcal{D}}(Q) \subseteq C_{\mathcal{D}}(A) \cap P$ . K tomu nám postačí fakt, že  $C_{\mathcal{D}}(Q)$  je podmnožinou obou množin. Z předpokladů  $Q \subseteq C_{\mathcal{D}}(A) \cap P$  a  $P \not\subseteq C_{\mathcal{D}}(A)$  nutně plyne, že  $Q \subset P$ , a tedy z definice  $\mathcal{P}_{\mathcal{D}}$  pak  $C_{\mathcal{D}}(Q) \subseteq P$ . Druhý fakt je už jednoduchý, jelikož z předpokladu nutně plyne  $Q \subseteq C_{\mathcal{D}}(A)$  a tedy využitím monotonie a idempotence operátoru  $C_{\mathcal{D}}$  dostaneme  $C_{\mathcal{D}}(Q) \subseteq C_{\mathcal{D}}(C_{\mathcal{D}}(A)) = C_{\mathcal{D}}(A)$ .  $\square$

Díky této vlastnosti můžeme dát do korespondence FZ GD báze a libovolné jiné báze. Přesněji, zaručí nám existenci injektivního zobrazení z množiny  $\mathcal{P}_{\mathcal{D}}$  do libovolné báze.

**Věta 10.** Nechť  $T$  je báze  $\mathcal{D}$ . Potom pro každou  $P \in \mathcal{P}_{\mathcal{D}}$  existuje  $A \Rightarrow B \in T$  taková, že  $C_{\mathcal{D}}(A) = C_{\mathcal{D}}(P)$  a  $\mathcal{D}_P \not\models A \Rightarrow B$ .

*Důkaz.* Nechť  $P \in \mathcal{P}_{\mathcal{D}}$ . Z toho plyne, že  $P \neq C_{\mathcal{D}}(P)$  a jelikož  $T$  je báze, máme taky  $P \notin \mathcal{M}_T$ , tedy existuje  $A \Rightarrow B \in T$  tak, že  $\mathcal{D}_P \not\models A \Rightarrow B$ . Nyní ukážeme, že  $C_{\mathcal{D}}(A) = C_{\mathcal{D}}(P)$ .

" $\subseteq$ ": Jestliže  $\mathcal{D}_P \not\models A \Rightarrow B$ , pak nutně  $A \subseteq P$  a zbytek plyne z monotonie operátoru  $C_{\mathcal{D}}$ .

" $\supseteq$ ": Stačí dokázat, že  $P \subseteq C_{\mathcal{D}}(A)$ , protože zbytek plyne z monotonie a idempotence operátoru  $C_{\mathcal{D}}$ . Tvrzení dokážeme sporem, tedy předpokládejme, že platí  $P \not\subseteq C_{\mathcal{D}}(A)$ . Jelikož  $A \Rightarrow B \in T$ , pak  $T \models A \Rightarrow B$ , dále pak  $B \subseteq C_{\mathcal{D}}(A) = [A]_T$ , protože  $T$  je báze. Dalším je, že  $B \not\subseteq P$ , protože  $\mathcal{D}_P \not\models A \Rightarrow B$ . Dohromady to znamená, že  $C_{\mathcal{D}}(A) \not\subseteq P$ , a když k tomu přidáme ještě předpoklad  $P \not\subseteq C_{\mathcal{D}}(A)$  zjistíme, že  $C_{\mathcal{D}}(A) \cap P \subset C_{\mathcal{D}}(A)$ . Ze stejného předpokladu máme  $A \subseteq P$ , tedy  $C_{\mathcal{D}}(A) \subseteq C_{\mathcal{D}}(P)$ . Navíc  $A \subseteq C_{\mathcal{D}}(A)$ , což s předchozím dává  $A \subseteq C_{\mathcal{D}}(A) \cap P$ . Monotonií dostaneme  $C_{\mathcal{D}}(A) \subseteq C_{\mathcal{D}}(C_{\mathcal{D}}(A) \cap P)$  a díky předchozí větě  $C_{\mathcal{D}}(A) \subseteq C_{\mathcal{D}}(A) \cap P$ , což je ale v rozporu s tím, že  $C_{\mathcal{D}}(A) \cap P \subset C_{\mathcal{D}}(A)$ .  $\square$

Další vlastnost koresponduje s tou, kterou jsme nazvali základní. Přidáme-li do uzávěrového systému některé dva pseudo-uzávěry, zůstane pořád uzávěrovým systémem.

**Věta 11.** *Pokud  $P_1, P_2 \in \mathcal{P}_{\mathcal{D}}$ ,  $P_1 \not\subseteq P_2$  a  $P_2 \not\subseteq P_1$ , pak  $C_{\mathcal{D}}(P_1 \cap P_2) = P_1 \cap P_2$ .*

*Důkaz.* Nejprve položme

$$\begin{aligned} T_1 &= GD(\mathcal{D}) \setminus \{P_1 \Rightarrow C_{\mathcal{D}}(P_1)\}, \\ T_2 &= GD(\mathcal{D}) \setminus \{P_2 \Rightarrow C_{\mathcal{D}}(P_2)\}. \end{aligned}$$

Pak z definice  $\mathcal{P}_{\mathcal{D}}$  máme, že  $\mathcal{D}_{P_1}$  je modelem  $T_1$  a  $\mathcal{D}_{P_2}$  je modelem  $T_2$ . Tím spíš jsou pak obě relace modely  $T_1 \cap T_2$ . Tím pádem i  $\mathcal{D}_{P_1 \cap P_2}$  musí být model  $T_1 \cap T_2$ , protože  $\mathcal{M}_{T_1 \cap T_2}$  je uzávěrový systém. Navíc  $\mathcal{D}_{P_1 \cap P_2}$  je i modelem  $T_1$ , protože z předpokladu  $P_2 \not\subseteq P_1$  plyne  $P_2 \not\subseteq P_1 \cap P_2$ . To samé platí i pro  $T_2$ , tedy  $\mathcal{D}_{P_1 \cap P_2}$  je modelem  $T_1 \cup T_2 = GD(\mathcal{D})$ . To znamená, že  $P_1 \cap P_2 = [P_1 \cap P_2]_{GD(\mathcal{D})}$ , a jelikož  $GD(\mathcal{D})$  je navíc báze  $\mathcal{D}$ , máme  $[P_1 \cap P_2]_{GD(\mathcal{D})} = C_{\mathcal{D}}(P_1 \cap P_2)$ . Dohromady tedy  $C_{\mathcal{D}}(P_1 \cap P_2) = P_1 \cap P_2$ .  $\square$

Než přistoupíme k důkazu, že GD báze je minimální vzhledem k počtu FZ, je potřeba se zamyslet. Použitím předchozích vlastností můžeme dokázat následující tvrzení.

**Věta 12.** *GD báze  $\mathcal{D}$  je minimální báze  $\mathcal{D}$ .*

*Důkaz.* K prokázání tvrzení nám stačí pro libovolnou bázi  $\mathcal{D}$ , označme ji  $T$ , najít injektivní zobrazení  $f : \mathcal{P}_{\mathcal{D}} \rightarrow T$ . Z předchozích tvrzení víme, že pro každou  $P \in \mathcal{P}_{\mathcal{D}}$  existuje  $A \Rightarrow B \in T$  tak, že platí  $C_{\mathcal{D}}(A) = C_{\mathcal{D}}(P)$  a  $\mathcal{D}_P \not\models A \Rightarrow B$ . Položme tedy  $f(P)$  rovno takovéto  $A \Rightarrow B$  a ukážeme, že se jedná o injektivní zobrazení.

Nechť  $P_1$  a  $P_2$  jsou prvky  $\mathcal{P}_{\mathcal{D}}$  takové, že  $f(P_1) = f(P_2)$ . Z předchozí věty jsou tyto obrazy podle  $f$  rovny  $A \Rightarrow B \in T$  a platí, že  $C_{\mathcal{D}}(P_1) = C_{\mathcal{D}}(A) = C_{\mathcal{D}}(P_2)$ . Nyní není možné, aby  $P_1 \subset P_2$ . Vskutku, kdyby to tak bylo, dle definice  $\mathcal{P}_{\mathcal{D}}$  bychom měli  $C_{\mathcal{D}}(P_1) \subseteq P_2$  a  $P_2 \neq C_{\mathcal{D}}(P_2)$ , což spolu s extenzivitou  $C_{\mathcal{D}}$  dává  $C_{\mathcal{D}}(P_1) \subset C_{\mathcal{D}}(P_2)$ . To je však v rozporu s předchozím. Stejně tak nemůže

nastat  $P_2 \subset P_1$ . Navíc také nemůže nastat  $P_2 \neq P_1$ , jinak bychom měli spor s  $C_{\mathcal{D}}(P_1) = C_{\mathcal{D}}(A)$ , protože dle předchozí věty by pak  $C_{\mathcal{D}}(P_1 \cap P_2) = P_1 \cap P_2$  a spolu s definicí  $f$  bychom měli  $A \subseteq P_1$  a  $A \subseteq P_2$ , tedy  $A \subseteq P_1 \cap P_2$ , což by pak použitím monotonie operátoru  $C_{\mathcal{D}}$  dalo  $C_{\mathcal{D}}(A) \subseteq P_1 \cap P_2 \subset C_{\mathcal{D}}(P_1)$ . Jediná možnost je tedy, že  $P_1 = P_2$ , čímž jsme prokázali injektivitu  $f$ .  $\square$

Jedinnou otázkou teď je, jak takovouto bázi najít. K tomu nám bude sloužit jemně upravený operátor  $\dots_{\mathcal{T}}^{\infty}$ .

**Definice 9.** Pro  $M \subseteq R$  zavedeme následující posloupnost podmnožin  $R$ :

$$\begin{aligned} M_{GD(\mathcal{D})}^{(0)} &= M, \\ M_{GD(\mathcal{D})}^{(i+1)} &= \bigcup \{B \mid A \Rightarrow B \in GD(\mathcal{D}) \text{ a } A \subset M_{GD(\mathcal{D})}^{(i)}\}, \\ M_{GD(\mathcal{D})}^{(\infty)} &= \bigcup_{i=0}^{\infty} M_{GD(\mathcal{D})}^{(i)}. \end{aligned}$$

I v tomto případě je jasné, že  $M_{GD(\mathcal{D})}^{(\infty)}$  je konečná, protože  $R$  je konečná a díky tomu se růst posloupnosti někdy zastaví.

**Cvičení 1.** 1. Dokažte, že  $\dots_{GD(\mathcal{D})}^{(\infty)}$  je uzávěrový operátor na  $R$ .

2. Pro každou množinu  $M \subseteq R = \{a, b, c, d, e, f, g, h\}$  a teorii

$$\begin{aligned} T &= \{\{a, b\} \Rightarrow \{c\}, \\ &\quad \{b\} \Rightarrow \{d\}, \\ &\quad \{c, d\} \Rightarrow \{e\}, \\ &\quad \{c, e\} \Rightarrow \{g, h\}, \\ &\quad \{g\} \Rightarrow \{a\}\} \end{aligned}$$

vypočtete  $M_{GD(\mathcal{D})}^{(\infty)}$ .

3. Dokažte, že platí  $M_{GD(\mathcal{D})}^{(\infty)} = M$  právě, když platí buď  $M = C_{\mathcal{D}}(M)$  nebo  $M \in \mathcal{P}_{\mathcal{D}}$ .

4. Naprogramujte operátor  $\dots_{GD(\mathcal{D})}^{(\infty)}$ . (hint: inspirujte se algoritmem CLOSURE)

5. Můžete zkusit naprogramovat generování  $GD(\mathcal{D})$ . (hint: jednoduše brute-force procházení všech podmnožin  $R$ )

Nyní je však na místě se ptát, jak nám ale pomůže s výpočtem  $GD(\mathcal{D})$ , když ji potřebujeme znát, abychom mohli spočítat pevné body operátoru  $\dots_{GD(\mathcal{D})}^{(\infty)}$ . Trik je ve využití  $\subset$ , které se vyskytuje u výpočtu, nebo-li k výpočtu je potřeba znát pouze všechny ostré podmnožiny. Stačí tedy hledat pevné body našeho operátoru v pořadí, které je obohacením relace podmnožinovitosti.

Nyní zavedeme úplné uspořádání na podmnožinách  $R$  a k tomu účelu seřadíme prvky  $R$  a pro jednoduchost je stotožníme s čísly, tedy  $R = \{1, 2, \dots, n\}$ .

**Definice 10.** Pro  $A, B \subseteq R$  a  $i \in R$  položíme  $A <_i B$ , pokud platí, že  $i \in B \setminus A$  a  $A \cap \{1, \dots, i-1\} = B \cap \{1, \dots, i-1\}$ . Navíc  $A < B$  pokud existuje  $i \in R$  tak, že  $A <_i B$ .

**Poznámka.** Všimněme si, že pokud  $A \subset B$ , pak nutně  $A < B$ . Z toho pak plyne, že  $\emptyset$  je vždy nejmenším prvkem v tomto uspořádání.

Tato relace nám definuje lexikografické uspořádání, které je úplné, tedy pro každé dvě  $A, B \in R$  máme buď  $A < B$  nebo  $B < A$ . V tomto pořadí, pak budeme hledat pevné body operátoru  $\dots_{GD(\mathcal{D})}^{(\infty)}$ .

**Příklad 2.** Mějme  $R = \{1, 2, 3, 4, 5, 6\}$  a uvažujme množiny

$$\{1\}, \{2\}, \{2, 3\}, \{3, 4, 5\}, \{3, 6\}, \{1, 4, 5\}.$$

Dle definice můžeme tyto množiny uspořádat následovně:

$$\{3, 6\} <_4 \{3, 4, 5\} <_2 \{2\} <_3 \{2, 3\} <_1 \{1\} <_4 \{1, 4, 5\}.$$

Nyní definujeme operátor s jehož pomocí, pak budeme hledat pevné body uzávěrového operátoru. Algoritmu, který pak vzejde z tvrzení o tomto operátoru, se říká **NextClosure**.

**Poznámka.** Připomeňme, že pevné body uzávěrového operátoru jsou ty množiny, které se operátorem zobrazí samy na sebe.

**Definice 11.** Pro  $A \subseteq R$ ,  $i \in R$  a uzávěrový operátor  $c$  položíme

$$A \oplus_c i = c((A \cap \{1, \dots, i-1\}) \cup \{i\}).$$

**Příklad 3.** Mějme  $R = \{a, b, c, d, e\}$  a uvažujme teorii

$$\begin{aligned} T &= \{\emptyset \Rightarrow \{e\}, \\ &\quad \{a, b, e\} \Rightarrow \{a, b, d, e\}, \\ &\quad \{c, d, e\} \Rightarrow \{a, b, c, d, e\}\}. \end{aligned}$$

Navíc budeme uvažovat lexikografické uspořádání na řetězcích, abychom mohli využít předchozí definici. To znamená, že  $a < b < c < d < e$ . Pak podle předchozí definice máme

$$\begin{aligned} \{c, d\} \oplus_{[\dots]_T} a &= [(\{c, d\} \cap \emptyset) \cup \{a\}]_T = [\{a\}]_T = \{a, e\} \\ \{a, d\} \oplus_{[\dots]_T} c &= [(\{a, d\} \cap \{a, b\}) \cup \{c\}]_T = [\{a, c\}]_T = \{a, c, e\} \\ \{a, b, d\} \oplus_{[\dots]_T} d &= [(\{a, b, d\} \cap \{a, b, c\}) \cup \{d\}]_T = [\{a, b, d\}]_T = \{a, b, d, e\} \\ \emptyset \oplus_{[\dots]_T} d &= [(\emptyset \cap \{a, b, c\}) \cup \{d\}]_T = [\{d\}]_T = \{d, e\} \end{aligned}$$

**Věta 13.** Nejmenší pevný bod  $A^+$  uzávěrového operátoru  $c$  na  $R$ , který je větší než  $A \subseteq R$  vzhledem k uspořádání  $<$ , je dán předpisem

$$B^+ = B \oplus_c i,$$

kde  $i$  je největší číslo takové, že  $B <_i B \oplus_c i$ .



Důkaz přeskóčíme a zaměříme se na význam předchozí věty. Abychom našli všechny pevné body libovolného uzávěrového operátoru stačí začít s prázdnou množinou a postupně hledat nejmenší větší uzávěry, což je vlastně to, co dělá algoritmus **NextClosure**.

**Příklad 4.** Uvažujme stejnou teorii a uspořádání jako v předchozím příkladu. Pak algoritmus **NextClosure** bude počítat následující:

$$\begin{aligned}
& \emptyset \oplus_{[\dots]_T} e = [(\emptyset \cap \{a, b, c, d\}) \cup \{e\}]_T = [\{e\}]_T = \{e\} \\
& \quad (\emptyset <_e \{e\}) \\
& \text{NextClosure}([\dots]_T, \emptyset) = \{e\} \\
& \{e\} \oplus_{[\dots]_T} d = [(\{e\} \cap \{a, b, c\}) \cup \{d\}]_T = [\{d\}]_T = \{d, e\} \\
& \quad (\{e\} <_d \{d, e\}) \\
& \text{NextClosure}([\dots]_T, \{e\}) = \{d, e\} \\
& \{d, e\} \oplus_{[\dots]_T} c = [(\{d, e\} \cap \{a, b\}) \cup \{c\}]_T = [\{c\}]_T = \{c, e\} \\
& \quad (\{d, e\} <_c \{c, e\}) \\
& \text{NextClosure}([\dots]_T, \{d, e\}) = \{c, e\} \\
& \{c, e\} \oplus_{[\dots]_T} d = [(\{c, e\} \cap \{a, b, c\}) \cup \{d\}]_T = [\{c, d\}]_T = \{a, b, c, d, e\} \\
& \quad (\{c, e\} \not<_d \{a, b, c, d, e\}) \\
& \{c, e\} \oplus_{[\dots]_T} b = [(\{c, e\} \cap \{a\}) \cup \{b\}]_T = [\{b\}]_T = \{b, e\} \\
& \quad (\{c, e\} <_b \{b, e\}) \\
& \text{NextClosure}([\dots]_T, \{c, e\}) = \{b, e\} \\
& \{b, e\} \oplus_{[\dots]_T} d = [(\{b, e\} \cap \{a, b, c\}) \cup \{d\}]_T = [\{b, d\}]_T = \{b, d, e\} \\
& \quad (\{b, e\} <_d \{d\}) \\
& \text{NextClosure}([\dots]_T, \{b, e\}) = \{d\} \\
& \{b, d, e\} \oplus_{[\dots]_T} c = [(\{b, d, e\} \cap \{a, b\}) \cup \{c\}]_T = [\{b, c\}]_T = \{b, c, e\} \\
& \quad (\{b, d, e\} <_c \{b, c, e\}) \\
& \text{NextClosure}([\dots]_T, \{b, d, e\}) = \{b, c, e\} \\
& \{b, c, e\} \oplus_{[\dots]_T} d = [(\{b, c, e\} \cap \{a, b, c\}) \cup \{d\}]_T = [\{b, c, d\}]_T = \{a, b, c, d, e\} \\
& \quad (\{b, c, e\} \not<_d \{a, b, c, d, e\}) \\
& \{b, c, e\} \oplus_{[\dots]_T} a = [(\{b, c, e\} \cap \emptyset) \cup \{a\}]_T = [\{a\}]_T = \{a, e\} \\
& \quad (\{b, c, e\} <_a \{a, e\}) \\
& \text{NextClosure}([\dots]_T, \{b, c, e\}) = \{a, e\} \\
& \quad \vdots
\end{aligned}$$

Jak je vidět, vždy hledáme největší možný atribut, který lze „přičíst“ k množině tak, aby byl výsledek větší.

Jako cvičení můžete pokračovat, dokud nedojdete k  $R$ .

My budeme chtít hledat pevné body operátoru  $\dots_{GD(\mathcal{D})}^{(\infty)}$ . Jelikož ale předem neznáme  $GD(\mathcal{D})$ , musíme ji postupně budovat. Jak už jsme si řekli, k výpočtu

uzávěru potřebujeme vždy pouze FZ z  $GD(\mathcal{D})$  s předpokladem ostře menším než daná množina. Navíc **NextClosure** počítá uzávěry v pořadí obohacujícím podmnožinovitost a tím pádem při každém kroku už můžeme znát všechny tyto předpoklady.

Algoritmus **MinBase** budeme formulovat jako posloupnost teorií a podmnožin  $R$ . Položíme  $M_0 = \emptyset$  a  $T_0 = \emptyset$  a další prvky posloupnosti definujeme následovně:

$$T_{n+1} = \begin{cases} T_n, & \text{pokud } M_n = C_{\mathcal{D}}(M_n) \text{ a} \\ T_n \cup \{M_n \Rightarrow C_{\mathcal{D}}(M_n)\} & \text{jinak.} \end{cases}$$

$$M_{n+1} = \text{NextClosure}(\dots_{T_{n+1}}^{(\infty)}, M_n)$$

Pro takhle sestavenou posloupnost platí, že je neklesající a navíc pro  $i \in \mathbf{N}_0$ , pro které platí, že  $M_i = R$  máme  $T_i = GD(\mathcal{D})$ .

**Příklad 5.** Uvažujme relaci  $D$  danou následující tabulkou:

a	b	c	d	e
1	2	2	2	1
3	2	5	1	1
1	1	1	1	1
2	1	4	3	1
2	1	3	3	1

Navíc uvažujme lexikografické uspořádání na atributech. Pak by **MinBase** počítal následující:

$$\begin{aligned} T_0 &= \emptyset \\ M_0 &= \emptyset \\ T_1 &= \{\emptyset \Rightarrow \{e\}\}, & \text{protože } \emptyset \neq C_{\mathcal{D}}(\emptyset) = \{e\} \\ M_1 &= \text{NextClosure}(\dots_{T_1}^{(\infty)}, \emptyset) = \{e\} \\ T_2 &= \{\emptyset \Rightarrow \{e\}\}, & \text{protože } \{e\} = C_{\mathcal{D}}(\{e\}) = \{e\} \\ M_2 &= \text{NextClosure}(\dots_{T_2}^{(\infty)}, \{e\}) = \{d, e\} \\ T_3 &= \{\emptyset \Rightarrow \{e\}\}, & \text{protože } \{d, e\} = C_{\mathcal{D}}(\{d, e\}) = \{d, e\} \\ M_3 &= \text{NextClosure}(\dots_{T_3}^{(\infty)}, \{d, e\}) = \{c, e\} \\ T_4 &= \{\emptyset \Rightarrow \{e\}\}, & \text{protože } \{c, e\} = C_{\mathcal{D}}(\{c, e\}) = \{c, e\} \\ M_4 &= \text{NextClosure}(\dots_{T_4}^{(\infty)}, \{c, e\}) = \{c, d, e\} \\ T_5 &= \{\emptyset \Rightarrow \{e\}, \{c, d, e\} \Rightarrow \{a, b, c, d, e\}\}, & \text{protože } \{c, d, e\} \neq C_{\mathcal{D}}(\{c, d, e\}) = \{a, b, c, d, e\} \\ M_5 &= \text{NextClosure}(\dots_{T_5}^{(\infty)}, \{c, d, e\}) = \{b, e\} \\ &\vdots \end{aligned}$$

Jako cvičení můžete pokračovat, dokud nedojdete k  $M_i = R$ .

**Cvičení 2.** Naprogramujte algoritmus **MinBase**.