1 Proofs of Theorems and Lemmas

THEOREM 4.1. Let $\mathcal{C}$ be a collection of itemsets and let $\mathcal{D}$ be a dataset. Let d be the maximum integer for which there are at least d transactions $\tau_1, \ldots, \tau_d \in \mathcal{D}$ such that the set $\{\tau_1, \ldots, \tau_d\}$ is an antichain, and each τ_i, $1 \leq i \leq d$ contains at least 2^{d-1} itemsets from $\mathcal{C}$. Then $\mathsf{EVC}(\mathcal{R}(\mathcal{C}), \mathcal{D}) \leq d$.

Proof. The antichain requirement guarantees that the set of transactions considered in the computation of d could indeed theoretically be shattered. Assume that a subset $\mathcal{F}$ of $\mathcal{D}$ contains two transactions τ' and τ'' such that $\tau' \subseteq \tau''$. Any itemset from $\mathcal{C}$ appearing in τ' would also appear in τ'', so there would not be any itemset $A \in \mathcal{C}$ such that $\tau'' \in T(A) \cap F$ but $\tau' \notin T(A) \cap \mathcal{F}$, which would imply that $\mathcal{F}$ can not be shattered. Hence sets that are not antichains should not be considered. This has the net effect of potentially resulting in a lower d, i.e., in a stricter upper bound to $\mathsf{EVC}(\mathcal{R}(\mathcal{C}), \mathcal{D})$.

Let now $\ell > d$ and consider a set $\mathcal{L}$ of ℓ transactions from $\mathcal{D}$ that is an antichain. Assume that $\mathcal{L}$ is shattered by $\mathcal{R}(\mathcal{C})$. Let τ be a transaction in $\mathcal{L}$. The transactions τ belongs to $2^{\ell-1}$ subsets of L. Let $\mathcal{K} \subseteq \mathcal{L}$ be one of these subsets. Since $\mathcal{L}$ is shattered, there exists an itemset $A \in \mathcal{C}$ such that $T(A) \cap \mathcal{L} = \mathcal{K}$. From this and the fact that $t \in \mathcal{K}$, we have that $\tau \in T(A)$ or equivalently that $A \subseteq \tau$. Given that all the subsets $\mathcal{K} \subseteq \mathcal{L}$ containing τ are different, then also all the $T(A)$'s such that $T(A) \cap \mathcal{L} = \mathcal{K}$ should be different, which in turn implies that all the itemsets A should be different and that they should all appear in τ. There are $2^{\ell-1}$ subsets $\mathcal{K}$ of $\mathcal{L}$ containing τ, therefore τ must contain at least $2^{\ell-1}$ itemsets from $\mathcal{C}$, and this holds for all ℓ transactions in $\mathcal{L}$. This is a contradiction because $\ell > d$ and d is the maximum integer for which there are at least d transactions containing at least 2^{d-1} itemsets from $\mathcal{C}$. Hence $\mathcal{L}$ cannot be shattered and the thesis follows.

LEMMA 4.1. Let j be the minimum integer for which $b_i \leq L_i$. Then $\mathsf{EVC}(\mathcal{C}, \mathcal{D}) \leq b_j$.

Proof. If $b_j \leq L_j$, then there are at least b_j transactions which can contain 2^{b_j-1} itemsets from $\mathcal{C}$ and this is the maximum b_i for which it happens, because the sequence $b_1, b_2, \ldots, b_w$ is sorted in decreasing order, given that the sequence $q_1, q_2, \ldots, q_w$ is. Then b_j satisfies the conditions of Thm. 4.1. Hence $\mathsf{EVC}(\mathcal{C}, \mathcal{D}) \leq b_j$.

LEMMA 5.1. Let $\mathcal{Y}$ be the set of maximal antichains in $\mathcal{F}$. If $\mathcal{D}$ is an ε_1-approximation to $(\mathcal{R}(2^I), \pi)$, then

1. $\max_{\mathcal{A} \in \mathcal{Y}} \mathsf{EVC}(\mathcal{R}(\mathcal{A}), \mathcal{D}) \geq \mathsf{EVC}(\mathcal{R}(\mathcal{B}), \mathcal{D})$, and
2. $\max_{\mathcal{A} \in \mathcal{Y}} \mathsf{VC}(\mathcal{R}(\mathcal{A})) \geq \mathsf{VC}(\mathcal{R}(\mathcal{B}))$.

Proof. Given that $\mathcal{D}$ is an ε_1-approximation to $(\mathcal{R}(2^I), \pi)$, then $\mathsf{TFI}(\pi, I, \theta) \subseteq \mathcal{G} \cup \mathcal{C}_1$. From this and the definition of negative border and of $\mathcal{F}$, we have that $\mathcal{B}) \subseteq \mathcal{F}$. Since $\mathcal{B}$ is a maximal antichain, then $\mathcal{B} \in \mathcal{Y}$. Hence the thesis.

THEOREM 5.1. With probability at least $1 - \delta$, $\mathsf{FI}(\mathcal{D}, I, \hat{\theta})$ contains no false positives:

$$\Pr\left(\mathsf{FI}(\mathcal{D}, I, \hat{\theta}) \subseteq \mathsf{TFI}(\pi, I, \theta)\right) \geq 1 - \delta \ .$$

Proof. Consider the two events E_1="$\mathcal{D}$ is an ε_1-approximation for $(\mathcal{R}(2^I), \pi)$" and E_2 ="$\mathcal{D}$ is an ε_2-approximation for $(\mathcal{R}(\mathcal{B}), \pi)$". From the above discussion and the definition of δ_1 and δ_2 it follows that the event $\mathsf{E} = \mathsf{E}_1 \cap \mathsf{E}_1$ occurs with probability at least $1 - \delta$. Suppose from now on that indeed E occurs.

Since E_1 occurs, then Lemma 5.1 holds, and the bounds we compute by solving the modified SUKP problems are indeed bounds to $\mathsf{VC}(\mathcal{R}(\mathcal{B}))$ and $\mathsf{EVC}(\mathcal{R}(\mathcal{B}, \mathcal{D}))$. Since E_2 also occurs, then for any $A \in \mathcal{B}$ we have $|t_\pi(A) - f_{\mathcal{D}}(A)| \leq \varepsilon_2$, but given that $t_\pi(A) < \theta$ because the elements of $\mathcal{B}$ are not TFIs, then we have $f_{\mathcal{D}}(A) < \theta + \varepsilon_2$. Because of the antimonotonicity property of the frequency and the definition of $\mathcal{B}$, this holds for any itemset that is not in $\mathsf{TFI}(\pi, I, \theta)$. Hence, the only itemsets that can have a frequency in $\mathcal{D}$ at least $\hat{\theta} = \theta + \varepsilon_2$ are the TFIs, so $\mathsf{FI}(\mathcal{D}, I, \hat{\theta}) \subseteq \mathsf{TFI}(\pi, I, \theta)$, which concludes our proof.