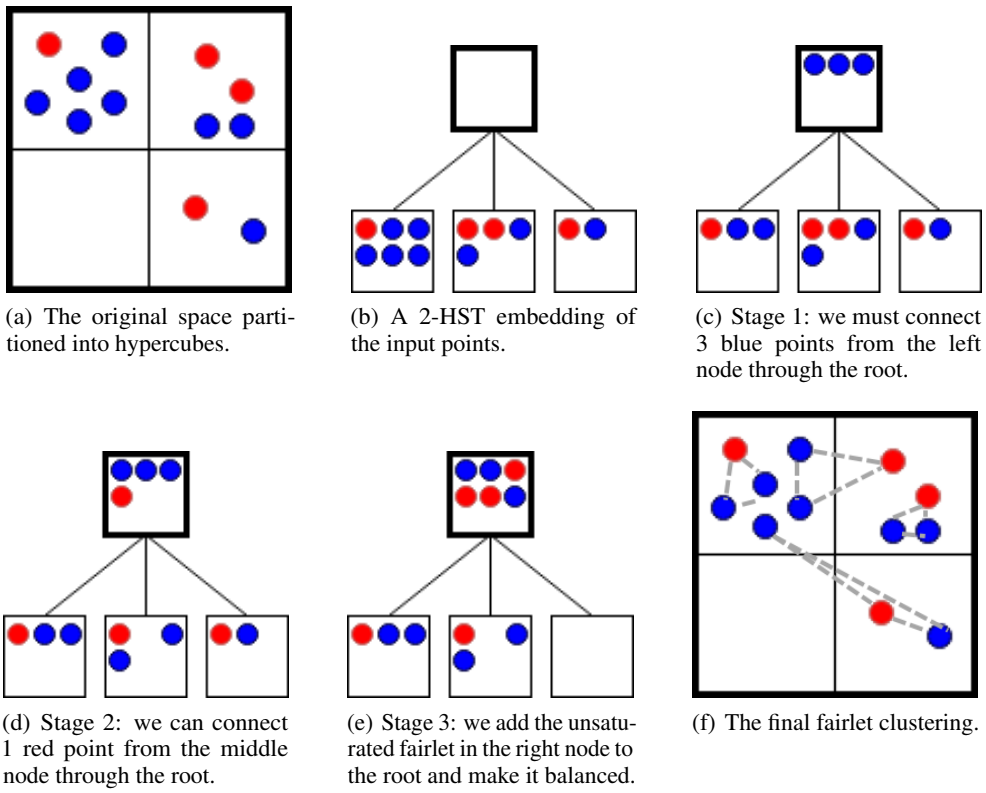
A. Example of Our Fairlet Decomposition



(a) The original space partitioned into hypercubes.

(b) A 2-HST embedding of the input points.

(c) Stage 1: we must connect 3 blue points from the left node through the root.

(d) Stage 2: we can connect 1 red point from the middle node through the root.

(e) Stage 3: we add the unsaturated fairlet in the right node to the root and make it balanced.

(f) The final fairlet clustering.

Figure 2. A run of our algorithm for (1,3)-fairlet decomposition on 8 blue points and 4 red points in $\mathbb{R}^2$. Steps (c)-(e) show the three stages of step 1 in FAIRLETDECOMPOSITION.

B. Missing Proofs

Proof of **Lemma 4.3.** The proof is by induction on height of v in T. The base case is when v is a leaf node in T and the algorithm trivially finds an optimal solution in this case. Suppose that the induction hypothesis holds for all vertices of T at height $h-1$. Here, we show that the statement holds for the vertices of T at height h as well.

Let OPT denote an optimal (r,b)-fairlet decomposition of the points in $T(v)$ with respect to $\mathrm{cost}_{\mathrm{med}}$. Next, we decompose OPT into γ^d+1 parts: $\{\mathrm{OPT}_i\}_{i\in[\gamma^d]}$ and $\mathrm{OPT}_{\mathrm{H}}$. For each $i\in[\gamma^d]$, OPT_i denotes the set of fairlets in OPT whose lca are in $T(v_i)$. Moreover, $\mathrm{OPT}_{\mathrm{H}}$ denotes the set of heavy fairlets with respect to v and $\mathrm{H}_{\mathrm{OPT}}$ denotes the set of heavy points with respect to v in OPT. Lastly, $\mathrm{H}^i_{\mathrm{OPT}} := \mathrm{H}_{\mathrm{OPT}} \cap T(v_i)$ denotes the set of heavy points with respect to v in OPT that are contained in $T(v_i)$.

Let SOL denote the solution returned by FAIRLETDECOMPOSITION(v,r,b). Similarly, we decompose SOL into γ^d+1 parts: $\{\mathrm{SOL}_i\}_{i\in[\gamma^d]}$ and $\mathrm{SOL}_{\mathrm{H}}$. Moreover, $\mathrm{H}_{\mathrm{SOL}}$ denotes the set of heavy points with respect to v in SOL and for each $i\in[\gamma^d]$, $\mathrm{H}^i_{\mathrm{SOL}} := \mathrm{H}_{\mathrm{SOL}} \cap T(v_i)$ denotes the set of heavy points with respect to v in SOL that are contained in $T(v_i)$.

Claim B.1. *For each $i\in[\gamma^d]$, there exists an (r,b)-fairlet decomposition of $P_i \setminus \mathrm{H}^i_{\mathrm{SOL}}$ of cost at most $\mathrm{cost}(\mathrm{OPT}_i) + (|\mathrm{H}^i_{\mathrm{OPT}}| + |\mathrm{H}^i_{\mathrm{SOL}}|\cdot(r+b))\cdot(r^2+b^2)\cdot\gamma^{h-1}$ where P_i is the set of points contained in $T(v_i)$.*

Hence, by the induction hypothesis, for each $i\in[\gamma^d]$,

$$\mathrm{cost}(\mathrm{SOL}_i) \le c\cdot(r^2+b^2)\cdot(\mathrm{cost}(\mathrm{OPT}_i) + (|\mathrm{H}^i_{\mathrm{OPT}}| + |\mathrm{H}^i_{\mathrm{SOL}}|\cdot(r+b))\cdot(r^2+b^2)\cdot\gamma^{h-1}). \tag{4}$$

Next, we bound the cost of SOL by Lemma 4.4 and (4) as follows:

$$\begin{aligned}
\text{cost}(\text{SOL}) &= \text{cost}(\text{SOL}_{\text{H}}) + \sum_{i\in[\gamma^d]} \text{cost}(\text{SOL}_i) \\
&\le \eta_{\text{H}} \cdot (r^2+b^2) \cdot \text{cost}(\text{OPT}_{\text{H}}) + c \cdot (r^2+b^2) \cdot (\frac{\eta_{\text{H}}(r+b)^5}{\gamma} \cdot \text{cost}(\text{OPT}_{\text{H}}) + \sum_{i\in[k^d]} \text{cost}(\text{OPT}_i)) \\
&\le c \cdot (r^2+b^2) \cdot \text{cost}(\text{OPT}) + (\eta_{\text{H}} - \frac{c}{2}) \cdot (r^2+b^2) \cdot \text{cost}(\text{OPT}_{\text{H}}) \quad \triangleright \text{ By setting } \gamma := 2\eta_{\text{H}}(r+b)^5 \\
&\le c \cdot (r^2+b^2) \cdot \text{cost}(\text{OPT}) \triangleright\, c \ge 2\eta_{\text{H}}
\end{aligned}$$

□

Proof of **Claim B.1.** Consider the fairlet decomposition OPT_i on $P_i \setminus \text{H}^i_{\text{OPT}}$. A fairlet $D \in \text{OPT}_i$ is *affected* if it contains a point $p \in \text{H}^i_{\text{SOL}}$.

We define the set of *affected points* as $\overline{P}_i = \text{H}^i_{\text{OPT}} \cup \bigcup_{D\in\overline{\text{OPT}}_i} D$ to denote the union of the points in the affected fairlets (i.e., $\bigcup_{D\in\overline{\text{OPT}}_i} D$) and the set H^i_{OPT} (whose points do not belong to any of fairlets in OPT_i).

Next, we bound the cost of the fairlet decomposition which is constructed by augmenting the set of fairlets $\text{OPT}_i \setminus \overline{\text{OPT}}_i$ with the set of affected points $\overline{P}_i$.

Let Q_0 denote the set of affected points $\overline{P}_i$. We augment the fairlet decomposition in three steps:

Step 1. In this step, we create as many (r,b)-balanced fairlets using the affected points Q_0 only. Note that the contribution of each point involved in such fairlets is $h_T(v_i)$ where $h_T(v_i)$ denotes the distance of v_i from the leaves in $T(v_i)$. Let $Q_1 \subseteq Q_0$ denote the set of affected points that do not join any fairlets at the end of this step. Note that all points in Q_1 are of the same color c.

Step 2: Next, we add as many points of Q_1 as possible to the existing fairlets in $\text{OPT}_i \setminus \overline{\text{OPT}}_i$ while preserving the (r,b)-balanced property. Now the extra cost incurred by each points of Q_1 that joins a fairlet in this step is at most $(r+b)\cdot h_T(v_i)$. Let $Q_2 \subset Q_1$ be the set of points that do not belong to an fairlets by the end of the second phase. Note that at the end of this step, if Q_2 is non-empty, then all fairlets are maximally-balanced c-dominant (a fairlet S is maximally-balanced c-dominant if (1) in S, the number of points of color c are larger than the number of points in color $\overline{c}$, (2) the set S is (r,b)-balanced, and (3) adding a point of color c to S makes it unbalanced).

Step 3: Finally, we show that by mixing the points of at most $b\cdot|Q_2|$ existing fairlets with the set Q_2, we can find an (r,b)-balanced fairlet decomposition of the involved points and the contribution of each such point to the total cost is at most $h_T(v_i)$. Note that since the set of all points we are considering is (r,b)-balanced, not all of the so far constructed fairlets are saturated (i.e., has size exactly $r+b$). In particular, we show that there exists a set of non-saturated fairlets $\mathcal{X}$ of size at most $b\cdot|Q_2|$ whose addition to Q_2 constitutes a (r,b)-balanced set. For each fairlet $D \in \mathcal{X}$,

$$|c_D| < \frac{r}{b}|\overline{c}_D| \Rightarrow b\cdot|c_D| \le r\cdot|\overline{c}_D| - 1,$$

where c_D and $\overline{c}_D$ respectively denotes the set of points of color c and $\overline{c}$ in D. This implies that after picking at most $|Q_2|$ non-saturated fairlets (i.e., the fairlets in $\mathcal{X}$),

$$b\cdot|c_{\mathcal{X}}| \le r\cdot|\overline{c}_{\mathcal{X}}| - b\cdot|Q_2| \Rightarrow b\cdot(|c_{\mathcal{X}}| + |Q_2|) \le r\cdot|\overline{c}_{\mathcal{X}}|,$$

where $c_{\mathcal{X}}$ and $\overline{c}_{\mathcal{X}}$ respectively denotes the set of points of color c and $\overline{c}$ in $\bigcup_{D\in\mathcal{X}} D$. Hence, the set of points $Q_2 \cup \bigcup_{D\in\mathcal{X}} D$ is (r,b)-balanced. Moreover, the cost of this step is at most $|Q_2|\cdot b\cdot(r+b)\cdot h_T(v_i)$.

Altogether, there exists a fairlet decomposition of $P_i \setminus \text{H}^i_{\text{SOL}}$ of cost at most

$$\begin{aligned}
&\text{cost}(\text{OPT}_i) + |Q_0 \setminus Q_1|\cdot h_T(v_i) + |Q_1 \setminus Q_2|\cdot(r+b)\cdot h_T(v_i) + |Q_2|\cdot b\cdot(r+b)\cdot h_T(v_i) \\
&\le \text{cost}(\text{OPT}_i) + |Q_0|\cdot b\cdot(r+b)\cdot h_T(v_i) \\
&\le \text{cost}(\text{OPT}_i) + (|\text{H}^i_{\text{OPT}}| + |\text{H}^i_{\text{SOL}}|\cdot(r+b))\cdot(r^2+b^2)\cdot\gamma^{h-1}
\end{aligned}$$

□

Proof of **Theorem 3.4** For a pointset X, let $\text{OPT}_{k\text{-fair}}(X)$ and $\text{OPT}_{\text{fairlet}}(X)$ respectively denote an optimal (r,b)-fair k-median and an optimal (r,b)-fairlet decomposition of X. It is straightforward to see that for any set of point X, $\text{cost}(\text{OPT}_{\text{fairlet}}(X)) \leq \text{cost}(\text{OPT}_{k\text{-fair}}(X))$ and in particular,

$$\text{cost}(Q) \leq \alpha \cdot \text{cost}(\text{OPT}_{k\text{-fair}}(P)). \tag{5}$$

Let N denote the set of the centers of fairlets in Q. For a set of points X, let $\text{OPT}_{k\text{-median}}(X)$ denotes an optimal k-median clustering of X (note that there is not fairness requirement). Since $C \subseteq P$, the optimal k-median cost of N is smaller than the optimal k-median cost of P. Since $\overline{P}$ contains at most $(r+b)$ copies of each point of N, by assigning all copies of each point $p \in N$ in $\overline{P}$ to the center of p in an optimal k-median clustering of N,

$$\text{cost}(\text{OPT}_{k\text{-median}}(\overline{P})) \leq (r+b) \cdot \text{cost}(\text{OPT}_{k\text{-median}}(N)) \leq (r+b) \cdot \text{cost}(\text{OPT}_{k\text{-median}}(P)). \tag{6}$$

As CLUSTERFAIRLET returns a β-approximate k-median clustering of $\overline{P}$, and by (5)-(6), the cost of the clustering $\mathcal{C}$ constructed by CLUSTERFAIRLET is

Since the distance of each point $p_i \in P$ to the center of its cluster in $\mathcal{C}^*$ is less than the sum of its distance to the center of its fairlet c_i in Q and the distance of c_i to its center in $\mathcal{C}$, we can bound the cost of $\mathcal{C}^*$ in terms of the costs of $\mathcal{C}$ and Q as follows:

$$\begin{aligned}
\text{cost}(\mathcal{C}^*) &\leq \text{cost}(Q) + \text{cost}(\mathcal{C}) \\
&\leq \alpha \cdot \text{cost}(\text{OPT}_{k\text{-fair}}(P)) && \triangleright \text{ By (5)} \\
&\quad + \beta \cdot \text{cost}(\text{OPT}_{k\text{-median}}(\overline{P})) \\
&\leq \alpha \cdot \text{cost}(\text{OPT}_{k\text{-fair}}(P)) \\
&\quad + \beta \cdot (r+b) \cdot \text{cost}(\text{OPT}_{k\text{-median}}(P)) && \triangleright \text{ By (6)} \\
&\leq (\alpha + \beta \cdot (r+b)) \cdot \text{cost}(\text{OPT}_{k\text{-fair}}(P)) && \square
\end{aligned}$$