A ANALYSIS OF STABLEDR

In this section, we show the following performance guarantee for STABLEDR.

Theorem A.1. *For a training data $\mathcal{L}$, a depth bound B, and a parameter $\epsilon > 0$, let $\omega = \text{STABLEDR}(\mathcal{L}, B, \epsilon)$. Then, we have $\mathbf{E}_\omega\left[\text{opt}_{\omega,B}(\mathcal{L})\right] \geq (1-\epsilon)\text{opt}_B(\mathcal{L})$. Moreover, we have $\sum_{i=1}^n d_{\text{TV}}(\omega, \omega^{(i)}) = O\left(\frac{\log|\Omega|}{\epsilon}\right)$, where $\omega^{(i)} = A(\mathcal{L}^{(i)}, B, \epsilon)$, and $d_{\text{TV}}(\omega, \omega^{(i)})$ denotes the total variation distance between (the distributions of) ω and $\omega^{(i)}$.*

The first inequality claims that we can achieve a nearly optimal total score using the output decision rule ω. The second inequality claims that the distribution of ω does not change significantly when a data point is removed from the training data. Theorem A.1 is obtained by combining Lemmas A.2 and A.4.

A.1 APPROXIMATION GUARANTEE

First, we show that the selected decision rule does not much deteriorate the total score of an optimal decision tree.

Lemma A.2. *Let $\omega = \text{STABLEDR}(\mathcal{L}, B, \epsilon)$. Then, we have*

$$\mathop{\mathbf{E}}_{\omega}[\text{opt}_{\omega,B}(\mathcal{L})] \geq (1-\epsilon)\text{opt}_B(\mathcal{L}).$$

Proof. For any $c > 0$, we have

$$\begin{aligned}
&\Pr[\text{opt}_{\omega,B}(\mathcal{L}) \leq \text{opt}_B(\mathcal{L}) - c] \\
&= \frac{\sum_{\psi\in\Omega:\text{opt}_{\psi,B}(\mathcal{L})\leq\text{opt}_B(\mathcal{L})-c} \exp(\lambda\cdot\text{opt}_{\psi,B}(\mathcal{L}))}{\sum_{\psi\in\Omega}\exp(\lambda\cdot\text{opt}_{\psi,B}(\mathcal{L}))} \\
&\leq \frac{|\Omega|\exp(\lambda\cdot(\text{opt}_B(\mathcal{L})-c))}{\sum_{\psi\in\Omega}\exp(\lambda\cdot\text{opt}_{\psi,B}(\mathcal{L}))} \\
&\leq \frac{|\Omega|\exp(\lambda\cdot(\text{opt}_B(\mathcal{L})-c))}{\exp(\lambda\cdot\text{opt}_B(\mathcal{L}))} \leq |\Omega|\exp(-\lambda c).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbf{E}[\text{opt}_{\omega,B}(\mathcal{L})] \\
&\geq \Pr[\text{opt}_{\omega,B}(\mathcal{L}) \leq \text{opt}_B(\mathcal{L}) - c]\cdot 0 \\
&\quad + \Pr[\text{opt}_{\omega,B}(\mathcal{L}) > \text{opt}_B(\mathcal{L}) - c]\cdot(\text{opt}_B(\mathcal{L}) - c) \\
&\geq (1 - |\Omega|\exp(-\lambda c))\cdot(\text{opt}_B(\mathcal{L}) - c) \\
&\geq \text{opt}_B(\mathcal{L}) - |\Omega|\exp(-\lambda c) - c.
\end{aligned}$$

By setting $c = \log|\Omega|/\lambda$ and the choice of λ, the claim holds. □

A.2 AVERAGE SENSITIVITY

Next, we analyze the average sensitivity of STABLEDR. For notational simplicity, we write opt_ω and $\text{opt}_\omega^{(i)}$ to denote $\text{opt}_{\omega,B}(\mathcal{L})$ and $\text{opt}_{\omega,B}(\mathcal{L}^{(i)})$, respectively. The following lemma is useful for our analysis.

Lemma A.3. *For any decision rule $\omega \in \Omega$, we have*

$$\sum_{i=1}^n \left(\text{opt}_\omega - \text{opt}_\omega^{(i)}\right) \leq \text{opt}_\omega.$$

Similarly, we have

$$\sum_{i=1}^n \left(\text{opt}_B(\mathcal{L}) - \text{opt}_B(\mathcal{L}^{(i)})\right) \leq \text{opt}_B(\mathcal{L}).$$

Proof. We first consider the first statement. Let ϕ be the optimal decision that attains opt_ω. Note that ϕ has depth B and the root node of ϕ has the decision rule ω. Then, we have

$$\sum_{i=1}^{n}\left(\mathrm{opt}_\omega - \mathrm{opt}_\omega^{(i)}\right) \le \sum_{i=1}^{n}(s(\phi,\mathcal{L}) - s(\phi,\mathcal{L}^{(i)})) = \sum_{i=1}^{n} 1[\phi(x_i)=y]$$
$$= s(\phi,\mathcal{L}) = \mathrm{opt}_\omega.$$

The second statement follows by a similar argument. □

Lemma A.4. *Let*

$$\omega = \textsc{StableDR}(\mathcal{L}, B, \epsilon),$$
$$\omega^{(i)} = \textsc{StableDR}(\mathcal{L}^{(i)}, B, \epsilon).$$

Then, we have

$$\sum_{i=1}^{n} d_{\mathrm{TV}}(\omega,\omega^{(i)}) = O\left(\frac{\log|\Omega|}{\epsilon}\right).$$

Proof. Notice that

$$\sum_{i=1}^{n} d_{\mathrm{TV}}(\omega,\omega^{(i)}) = \sum_{i=1}^{n}\sum_{\psi\in\Omega} \max\Big\{0, \Pr[\omega=\psi] - \Pr[\omega^{(i)}=\psi]\Big\}.$$

Let $\lambda^{(i)}$ be λ used in $\textsc{StableDR}(\mathcal{L}^{(i)}, B, \epsilon)$. Then we have

$$\begin{aligned}
&\max\Big\{0, \Pr[\omega=\psi] - \Pr[\omega^{(i)}=\psi]\Big\} \\
&= \max\left\{0, \frac{\exp(\lambda\cdot\mathrm{opt}_\psi)}{\sum_{\psi'\in\Omega}\exp(\lambda\cdot\mathrm{opt}_{\psi'})} - \frac{\exp(\lambda^{(i)}\cdot\mathrm{opt}_\psi^{(i)})}{\sum_{\psi'\in\Omega}\exp(\lambda^{(i)}\cdot\mathrm{opt}_{\psi'}^{(i)})}\right\} \\
&\le \frac{\exp(\lambda\cdot\mathrm{opt}_\psi) - \exp(\lambda\cdot\mathrm{opt}_\psi^{(i)})}{\sum_{\psi'\in\Omega}\exp(\lambda\cdot\mathrm{opt}_{\psi'})} \\
&\quad + \max\left\{0, \frac{\exp(\lambda\cdot\mathrm{opt}_\psi^{(i)})}{\sum_{\psi'\in\Omega}\exp(\lambda\cdot\mathrm{opt}_{\psi'})} - \frac{\exp(\lambda^{(i)}\cdot\mathrm{opt}_\psi^{(i)})}{\sum_{\psi'\in\Omega}\exp(\lambda^{(i)}\cdot\mathrm{opt}_{\psi'}^{(i)})}\right\},
\end{aligned} \tag{4}$$

where the equality is from the design of the algorithm and the inequality is from the following inequality

$$\max\{0, b-a\} \le (b-x) + \max\{0, x-a\}$$

which holds for any $x \le b$.

Let $A_{i,\psi}$ and $B_{i,\psi}$ denote the first and second terms, respectively, of (4). The following two claims bound the sums of the first and the second terms over i and ψ.

Claim A.5.

$$\sum_{i=1}^{n}\sum_{\psi\in\Omega} A_{i,\psi} \le \lambda\cdot\mathrm{opt}_B(\mathcal{L}).$$

Claim A.6.

$$\sum_{i=1}^{n}\sum_{\psi\in\Omega} B_{i,\psi} \le O(\lambda\cdot\mathrm{opt}_B(\mathcal{L})).$$

Before proving these claims, we first complete the proof of the lemma assuming them. Combining (4) and the two claims above, we have

$$\sum_{i=1}^{n} d_{\mathrm{TV}}(\omega,\omega^{(i)}) \le O(\lambda\cdot\mathrm{opt}_B(\mathcal{L})) + 1 = O\left(\frac{\log|\Omega|}{\epsilon}\right).$$ □

Theorem A.1 follows by combining Lemmas A.2 and A.4.

Proof of Claim A.5. We have

$$\begin{aligned}
&\frac{\exp(\lambda \cdot \mathrm{opt}_\psi) - \exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})} = \Pr[\omega = \psi] \left(1 - \frac{\exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\exp(\lambda \cdot \mathrm{opt}_\psi)}\right) \\
&= \Pr[\omega = \psi] \left(1 - \exp\left(-\lambda \cdot (\mathrm{opt}_\psi - \mathrm{opt}_\psi^{(i)})\right)\right) \\
&\leq \lambda \cdot \Pr[\omega = \psi](\mathrm{opt}_\psi - \mathrm{opt}_\psi^{(i)}),
\end{aligned}$$

where the inequality is from $1 - e^{-x} \leq x$ for any $x \in \mathbb{R}$. Therefore, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\psi \in \Omega} A_{i,\psi} \\
&= \sum_{i=1}^{n} \sum_{\psi \in \Omega} \frac{\exp(\lambda \cdot \mathrm{opt}_\psi) - \exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})} \\
&\leq \lambda \sum_{i=1}^{n} \sum_{\psi \in \Omega} \Pr[\omega = \psi](\mathrm{opt}_\psi - \mathrm{opt}_\psi^{(i)}) \\
&\leq \lambda \sum_{\psi \in \Omega} \Pr[\omega = \psi] \mathrm{opt}_\psi \qquad \text{(by Lemma A.3)} \\
&\leq \lambda \cdot \mathrm{opt}_B(\mathcal{L}).
\end{aligned}$$

as desired. □

Proof of Claim A.6. We first note that

$$\begin{aligned}
&B_{i,\psi} \\
&= \max\left\{0, \frac{\exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})} - \frac{\exp(\lambda^{(i)} \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda^{(i)} \cdot \mathrm{opt}_{\psi'}^{(i)})}\right\} \\
&\leq \max\left\{0, \frac{\exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})} - \frac{\exp(\lambda^{(i)} \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})}\right\} \\
&= \max\left\{0, \frac{\exp(\lambda \cdot \mathrm{opt}_\psi^{(i)})}{\sum_{\psi' \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi'})} \left(1 - \exp\left(-\mathrm{opt}_\psi^{(i)}(\lambda - \lambda^{(i)})\right)\right)\right\} \\
&\leq \max\left\{0, \Pr[\omega = \psi] \mathrm{opt}_\psi^{(i)}(\lambda - \lambda^{(i)})\right\} \\
&\leq \mathrm{opt}_B(\mathcal{L}) \Pr[\omega = \psi] |\lambda - \lambda^{(i)}|.
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \left|\lambda - \lambda^{(i)}\right| \leq \lambda \sum_{i=1}^{n} \left| \frac{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})} - \frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L}^{(i)})}}{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})}} \right| \\
&\leq \lambda \sum_{i=1}^{n} \max\left\{ \frac{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})} - \frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L}^{(i)})}}{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})}}, \frac{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L}^{(i)})} - \frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})}}{\frac{\log|\Omega|}{\mathrm{opt}_B(\mathcal{L})}} \right\} \\
&\leq \lambda \sum_{i=1}^{n} \max\left\{ \frac{\log|\Omega| - \log|\Omega|}{\log|\Omega|}, \frac{\frac{1}{\mathrm{opt}_B(\mathcal{L}^{(i)})} - \frac{1}{\mathrm{opt}_B(\mathcal{L})}}{\frac{1}{\mathrm{opt}_B(\mathcal{L})}} \right\} \\
&\leq \lambda \sum_{i=1}^{n} \frac{\log|\Omega| - \log|\Omega|}{\log|\Omega|} + \lambda \sum_{i=1}^{n} \frac{\mathrm{opt}_B(\mathcal{L}) - \mathrm{opt}_B(\mathcal{L}^{(i)})}{\mathrm{opt}_B(\mathcal{L}^{(i)})}
\end{aligned}$$

$$\leq \lambda \sum_{i=1}^{n} \frac{\log|\Omega| - \log|\Omega|}{\log|\Omega|} + 2\lambda \sum_{i=1}^{n} \frac{\mathrm{opt}_B(\mathcal{L}) - \mathrm{opt}_B(\mathcal{L}^{(i)})}{\mathrm{opt}_B(\mathcal{L})} \qquad \text{(by } \mathrm{opt}_B(\mathcal{L}^{(i)}) \geq 1\text{)}$$

$$= O(\lambda), \qquad \text{(by Lemma A.3)}$$

Combining the two inequalities above, we obtain

$$\sum_{i=1}^{n} \sum_{\psi \in \Omega} B_{i,\psi} \leq \sum_{i=1}^{n} \sum_{\psi \in \Omega} \mathrm{opt}_B \Pr[\omega = \psi] |\lambda - \lambda^{(i)}| = O\left(\lambda \cdot \mathrm{opt}_B(\mathcal{L})\right). \qquad \square$$

B Analysis of StableDT

In this section, we analyze StableDT and prove Theorem 4.1

Proof of the first claim of Theorem 4.1. Let $\mathcal{L}_0$ be the input training data (so that we can use $\mathcal{L}$ to denote other sets).

We prove the following by backward induction on depth.

$$\mathbf{E}[s(\text{StableDT'}(\mathcal{L}, B, \epsilon, d), \mathcal{L})] \geq (1-\epsilon)^{B-d} \mathrm{opt}_B(\mathcal{L}).$$

Then, the statement holds by setting $d = 0$.

The claim clearly holds when $d = B$ because we output the optimal label.

Suppose that the claim holds for depth more than d. Consider a particular call StableDT'$(\mathcal{L}, B, \epsilon, d)$, and let ϕ denote the output decision tree, let ω be the decision rule used in the root node of ϕ, and let $\mathcal{L}_L^\omega$ and $\mathcal{L}_R^\omega$ denote the two training datas obtained from $\mathcal{L}$ by splitting it according to ω. Note that these are random variables. Then, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\phi}[s(\phi, \mathcal{L})] &= \sum_{\psi \in \Omega} \Pr[\omega = \psi] \Big(\mathbf{E}[s(\text{StableDT'}(\mathcal{L}_L^\psi, B, \epsilon, d+1), \mathcal{L}_L^\psi)] \\
&\qquad + \mathbf{E}[s(\text{StableDT'}(\mathcal{L}_R^\psi, B, \epsilon, d+1), \mathcal{L}_R^\psi)] \Big) \\
&\geq \sum_{\psi \in \Omega} \Pr[\omega = \psi] \Big((1-\epsilon)^{B-d-1} \mathrm{opt}_{B-d-1}(\mathcal{L}_L^\psi) \\
&\qquad + (1-\epsilon)^{B-d-1} \mathrm{opt}_{B-d-1}(\mathcal{L}_R^\psi) \Big) \\
&\geq (1-\epsilon)^{B-d-1} \sum_{\psi \in \Omega} \Pr[\omega = \psi] \Big(\mathrm{opt}_{\psi, B-d}(\mathcal{L}) \Big) \\
&\geq (1-\epsilon)^{B-d} \sum_{\psi \in \Omega} \Pr[\omega = \psi] (1-\epsilon) \mathrm{opt}_{B-d}(\mathcal{L}) \\
&\geq (1-\epsilon)^{B-d} \mathrm{opt}_{B-d}(\mathcal{L}),
\end{aligned}$$

where the first inequality is based on the induction hypothesis and the second to last inequality is based on Theorem A.1. $\square$

Proof of the second claim of Theorem 4.1. For notational simplicity, we drop the arguments B and ϵ when calling StableDT'$(\mathcal{L}, B, \epsilon, d)$, because they are fixed in this proof. Additionally, we write StableDT' instead of StableDT'.

Let $\mathcal{L}_0 = ((x_1, y_1), \ldots, (x_n, y_n))$ be the input training data (so that we can use $\mathcal{L}$ to denote other sets). For a subset $\mathcal{L}$ of $\mathcal{L}_0$ and $i \in \{1, 2, \ldots, n\}$, let $\mathcal{L}^{(i)} := \mathcal{L} \setminus \{(x_i, y_i)\}$.

For $0 \leq d \leq B$, let $\mathcal{L}_{d,1}, \ldots, \mathcal{L}_{d,2^d}$ be the sets on which StableDT' is called at depth d (if the number of sets on which StableDT' is called at depth d is less than 2^d, we append empty sets). We can order them so that StableDT'$(\mathcal{L}_{d,j}, d)$ calls StableDT'$(\mathcal{L}_{d+1,2j-1}, d+1)$ and StableDT'$(\mathcal{L}_{d+1,2j}, d+1)$ (if StableDT'$(\mathcal{L}_{d,j}, d)$ does not make recursive calls, we set $\mathcal{L}_{d+1,2j-1} = \mathcal{L}_{d+1,2j} = \emptyset$).

For fixed $\{\mathcal{L}_{B,j}\}_j$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{2^d} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}_{B,j}, B), \textsc{StableDT'}(\mathcal{L}^{(i)}_{B,j}, B)) \\ \leq \sum_{i=1}^{n}\sum_{j=1}^{2^d} 1[(x_i, y_i) \in \mathcal{L}_{B,j}] = \sum_{i=1}^{n} 1[(x_i, y_i) \in \mathcal{L}] = |\mathcal{L}|$$

because the output changes only when $(x_i, y_i) \in \mathcal{L}$ and the output change is bounded by one.

Let $0 \leq d < B$. Let $\omega_{d,j}$ and $\omega^{(i)}_{d,j}$ be the ω values used in $\textsc{StableDT'}(\mathcal{L}_{d,j}, d)$ and $\textsc{StableDT'}(\mathcal{L}^{(i)}_{d,j}, d)$, respectively. Note that they are random variables. For a rule ω, Let $\mathcal{L}^{\omega}_{d+1,2j-1}$ and $\mathcal{L}^{\omega}_{d+1,2j}$ be the two sets obtained by partitioning $\mathcal{L}_{d,j}$ according to ω. Then for fixed $\{\mathcal{L}_{d,j}\}_j$, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{2^d} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}_{d,j}, d), \textsc{StableDT'}(\mathcal{L}^{(i)}_{d,j}, d)) \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{2^d} \Big(d_{\mathrm{TV}}(\omega_{d,j}, \omega^{(i)}_{d,j}) \cdot 2^{B-d} \\
&\quad + \mathop{\mathbf{E}}_{\omega_{d,j}} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}^{\omega_{d,j}}_{d+1,2j-1}, d+1), \textsc{StableDT'}(\mathcal{L}^{\omega_{d,j},(i)}_{d+1,2j-1}, d+1)) \\
&\quad + \mathop{\mathbf{E}}_{\omega_{d,j}} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}^{\omega_{d,j}}_{d+1,2j}, d+1), \textsc{StableDT'}(\mathcal{L}^{\omega_{d,j},(i)}_{d+1,2j}, d+1)) \Big) \\
&\leq C \cdot 2^{B-d} \sum_{j=1}^{2^d} \frac{\log|\Omega|}{\epsilon} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{2^{d+1}} \mathop{\mathbf{E}}_{\omega_{d,j}} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}^{\omega_{d,j}}_{d+1,j}, d+1), \textsc{StableDT'}(\mathcal{L}^{\omega_{d,j},(i)}_{d+1,j}, d+1)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemma A.1)} \\
&\leq C \cdot 2^{B} \frac{\log|\Omega|}{\epsilon} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{2^{d+1}} \mathop{\mathbf{E}}_{\omega_{d,j}} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}^{\omega_{d,j}}_{d+1,j}, d+1), \textsc{StableDT'}(\mathcal{L}^{\omega_{d,j},(i)}_{d+1,j}, d+1)),
\end{aligned}$$

where $C > 0$ is some universal constant. By backward induction, we obtain for any d and fixed $\{\mathcal{L}_{d,j}\}_j$

$$\sum_{i=1}^{n}\sum_{j=1}^{2^d} d_{\mathrm{EM}}(\textsc{StableDT'}(\mathcal{L}_{d,j}, d), \textsc{StableDT'}(\mathcal{L}^{(i)}_{d,j}, d)) \\ \leq C \cdot 2^{B}(B-d)\frac{\log|\Omega|}{\epsilon} + n$$

for every $0 \leq d \leq B$. By setting $d = 0$, we obtain the claim. □

C Missing proofs of Section 5

In this section, we prove Theorem 5.1. We discuss modifications to StableDR and StableDT in Sections C.1 and C.2, respectively.

Algorithm 4:

Procedure SEEDEDSTABLEDR($\mathcal{L}, B, \epsilon, \pi$)
  $\lambda \leftarrow \frac{2 \log |\Omega|}{\epsilon \cdot \mathrm{opt}_B(\mathcal{L})}$;
  while `true` **do**
    Sample $\omega \in \Omega$ uniformly at random using π;
    Sample $\tau \in [0, 1]$ uniformly at random using π;
    p_ω be the probability of choosing ω as given in STABLEDR;
    if $p_\omega > \tau$ **then return** ω;
Procedure SEEDEDSTABLEDT'($\mathcal{L}, B, \epsilon, d, j, \pi$)
  if $|\mathcal{L}| \leq 1$ *or* $d = B$ **then**
    return an optimal label for $\mathcal{L}$.
  $\omega \leftarrow$ SEEDEDSTABLEDR($\mathcal{L}, B, \epsilon, \pi$);
  Partition $\mathcal{L}$ into $\mathcal{L}_L \cup \mathcal{L}_R$ according to ω;
  $\pi_L \leftarrow (\pi_1, \pi_3, \ldots)$ and $\pi_R \leftarrow (\pi_2, \pi_4, \ldots)$;
  $\phi_L \leftarrow$ SEEDEDSTABLEDT'($\mathcal{L}_L, B, \epsilon, d + 1, 2j, \pi_L$);
  $\phi_R \leftarrow$ SEEDEDSTABLEDT'($\mathcal{L}_R, B, \epsilon, d + 1, 2j + 1, \pi_R$);
  Let ϕ_ω be the decision tree such that the root node t has rule ω and the left and right children of t are ϕ_L and ϕ_R, respectively;
  return ϕ_ω.
Procedure SEEDEDSTABLEDT($\mathcal{L}, B, \epsilon, \pi$)
  return SEEDEDSTABLEDT'($\mathcal{L}, B, \epsilon, 0, 1, \pi$).

C.1 DECISION RULE SELECTION

In STABLEDR, we sampled a rule $\omega \in \Omega$ by the exponential mechanism McSherry & Talwar (2007). To bound the expected deterministic average sensitivity over random bits, we perform the following rejection sampling. We first sample a rule $\omega \in \Omega$ and threshold $\tau \in [0, 1]$ uniformly at random by using π. If the threshold τ is more than the probability p_ω that we sample ω in the exponential mechanism, then we output ω. Otherwise, we repeat the same process again. The details are given as SEEDEDSTABLEDR in Algorithm 4.

The following lemma shows that the distributions of STABLEDR and DERANDOMIZEDSTABLEDR are the same and the derandomized average sensitivity of the latter can be bounded from above by the average sensitivity of the former.

Lemma C.1. *Let*

$$\begin{aligned}
\omega &= \text{STABLEDR'}(\mathcal{L}, B, \epsilon), \\
\omega^{(i)} &= \text{STABLEDR'}(\mathcal{L}^{(i)}, B, \epsilon), \\
\omega_\pi &= \text{DERANDOMIZEDSTABLEDR'}(\mathcal{L}, B, \epsilon, \pi), \\
\omega_\pi^{(i)} &= \text{DERANDOMIZEDSTABLEDR'}(\mathcal{L}^{(i)}, B, \epsilon, \pi).
\end{aligned}$$

Then, the distribution of ω and that of ω_π over π are the same. Moreover for any $i \in \{1, 2, \ldots, n\}$, we have

$$\Pr_\pi[\omega_\pi \neq \omega_\pi^{(i)}] \leq 2 d_{\mathrm{TV}}(\omega, \omega^{(i)}).$$

Proof. The first claim is clear from the design of DERANDOMIZEDSTABLEDR.

Now we see the second claim. Let $Z = \sum_{\psi \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi, B}(\mathcal{L}))$ and let $Z^{(i)} = \sum_{\psi \in \Omega} \exp(\lambda \cdot \mathrm{opt}_{\psi, B}(\mathcal{L}^{(i)}))$. For $\omega \in \Omega$, we let $p_\omega = \exp(\lambda \cdot \mathrm{opt}_{\omega, B}(\mathcal{L}))/Z$. For $\omega \in \Omega$, we let $p_\omega^{(i)} = \exp(\lambda \cdot \mathrm{opt}_{\omega, B}(\mathcal{L}^{(i)}))/Z^{(i)}$, and for $\omega \in \Omega \setminus \Omega$, we let $p_\omega^{(i)} = 0$. Then, we have

$$\Pr_\pi[\omega_\pi \neq \omega_\pi^{(i)}] \leq \sum_{\psi \in \Omega} \Pr_\tau[\min\{p_\psi, p_\psi^{(i)}\} < \tau < \max\{p_\psi, p_\psi^{(i)}\}]$$

$$= \sum_{\psi \in \Omega} |p_\psi - p_\psi^{(i)}| = 2 d_{\mathrm{TV}}(\omega, \omega^{(i)}). \qquad \square$$

By the analysis of StableDR and Lemma C.1, we obtain the following:

Theorem C.2. *Let* $\omega_\pi = \mathrm{SeededStableDR}(\mathcal{L}, B, \epsilon, \pi)$. *We have* $\mathbf{E}_\pi[\mathrm{opt}_{\omega_\pi, B}(\mathcal{L})] \geq (1 - \epsilon)\mathrm{opt}_B(\mathcal{L})$. *Moreover for* $\omega_\pi^{(i)} = \mathrm{SeededStableDR}(\mathcal{L}^{(i)}, B, \epsilon, \pi)$, *we have* $\mathbf{E}_\pi \left[\sum_{i=1}^n d_{\mathrm{DT}}(\omega_\pi, \omega_\pi^{(i)}) \right] = O\left(\frac{\log |\Omega|}{\epsilon} \right)$.

C.2 Decision Tree Construction

We now explain the modification to StableDT. Let $\mathcal{L}_{d,1}, \ldots, \mathcal{L}_{d,2^d}$ be the sets on which our algorithm is called at depth d as defined in the proof in Section B. Then, we want to make sure that the same random bits are used when processing particular $\mathcal{L}_{d,j}$ no matter whether the input training data is $\mathcal{L}$ or $\mathcal{L}^{(i)}$ $(1 \leq i \leq n)$. To this end, at each node in the decision tree, we split the random bits $\pi = (\pi_1, \pi_2, \ldots)$ into $\pi_L = (\pi_1, \pi_3, \ldots)$ and $\pi_R = (\pi_2, \pi_4, \ldots)$, and then pass π_L and π_R on to the nodes for $\mathcal{L}_{d+1,2j}$ and $\mathcal{L}_{d+1,2j+1}$, respectively. See Algorithm 4 for details.

We replace Theorem A.1 with Theorem C.2 in the proof of Theorem 4.1, and we obtain Theorem 5.1.

D Additional Results

D.1 Detailed Results in Section 7.2

In Section 7.2, we reported the trends of average sensitivity and accuracies over ϵ on small datasets, breast cancer and diabetes. Here, we show the detailed results (i) with error bars, and (ii) with a relaxed version of the tree distance. For (i), in addition to the average results, we also show their variations. More specifically, we report the 25 and 75 percentiles of the results over 10 random realizations of the sampled training data. For (ii), we adopt a relaxed version of the tree distance in Algorithm 5. In the original tree distance in Algorithm 2, we regarded that two trees are (completely) different when their top rules are different (Line 6). In the relaxed version in Algorithm 5, we regard that two trees are completely different only when the features used in the top rules are different. With this relaxation, we regard two subtrees with similar top rules such as $\omega : u \mapsto 1[u_1 \leq 1.0]$ and $\omega' : u \mapsto 1[u_1 \leq 1.01]$ as identical.

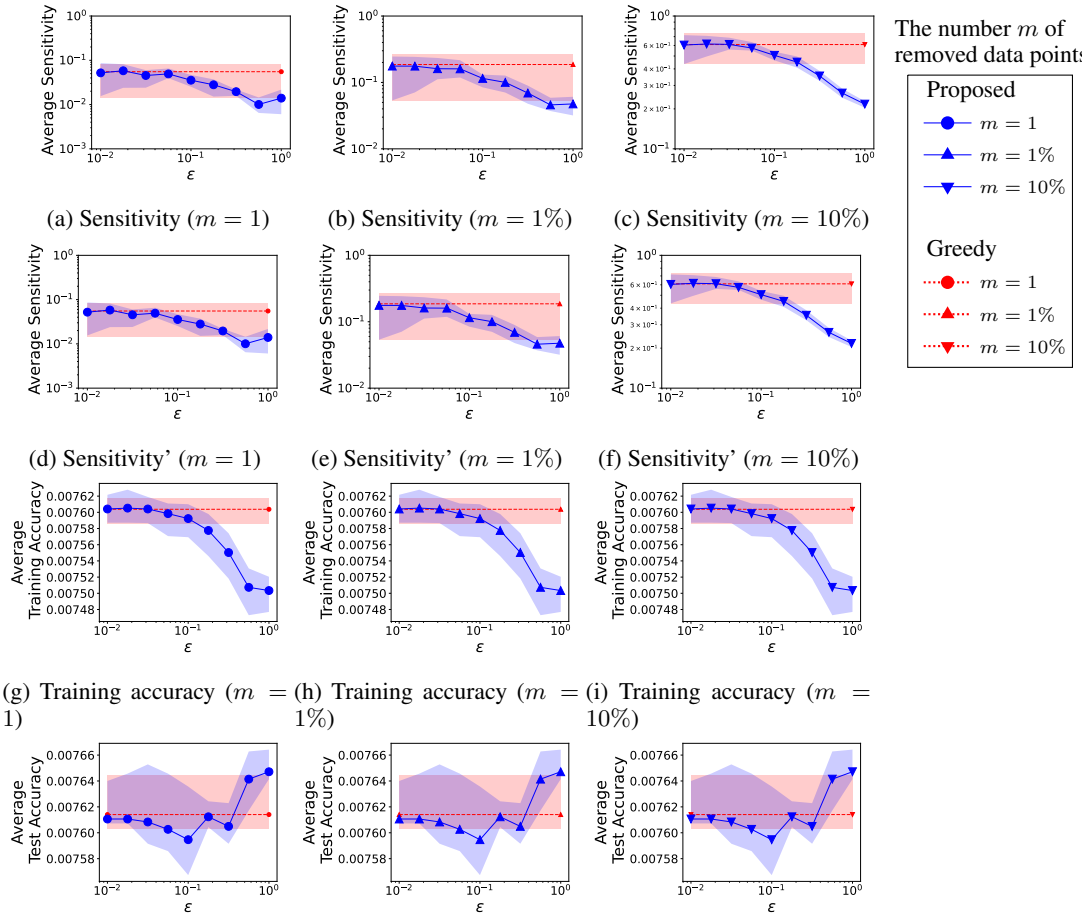
Algorithm 5:

1 **Procedure** Distance'(ϕ, ϕ')
2 Let t and t' be the root nodes of ϕ and ϕ', respectively;
3 Let ω.`feature` be the feature used in ω;
4 **if** *both* t *and* t' *are leaves* **then**
5 **return** 0 if $y_t = y_{t'}$ and 2 otherwise.
6 **else if** *either* t *or* t' *is a leaf* **then return** $|\phi| + |\phi'|$;
7 **else if** ω_t.`feature` $\neq \omega_{t'}$.`feature` **then return** $|\phi| + |\phi'|$;
8 **else**
9 Let ϕ_L, ϕ_R be the decision trees rooted at the left and right children of t, respectively;
10 Let ϕ'_L, ϕ'_R be the decision trees rooted at the left and right children of t', respectively;
11 **return** Distance'(ϕ_L, ϕ'_L) + Distance'(ϕ_R, ϕ'_R).

Figures 5 and 6 show the detailed results on breast cancer and diabetes, respectively. In the figures, we show the 25 and 75 percentiles using colored shades. The figures named Sensitivity and Sensitivity' are the average sensitivity computed using the original distance and the relaxed distance, respectively.

The figures confirm that the decrease of the average sensitivity for $\epsilon > 0.1$ will be sufficiently significant, in particular for the number of data removal $m = 1\%$ and 10%. The figures on Sensitivity and Sensitivity' also confirm that the average sensitivity measured by using the original tree distance

and the relaxed tree distance are almost identical, implying the choice of the tree distance will only have negligible impacts to the results.



(a) Sensitivity ($m = 1$) (b) Sensitivity ($m = 1\%$) (c) Sensitivity ($m = 10\%$)

(d) Sensitivity' ($m = 1$) (e) Sensitivity' ($m = 1\%$) (f) Sensitivity' ($m = 10\%$)

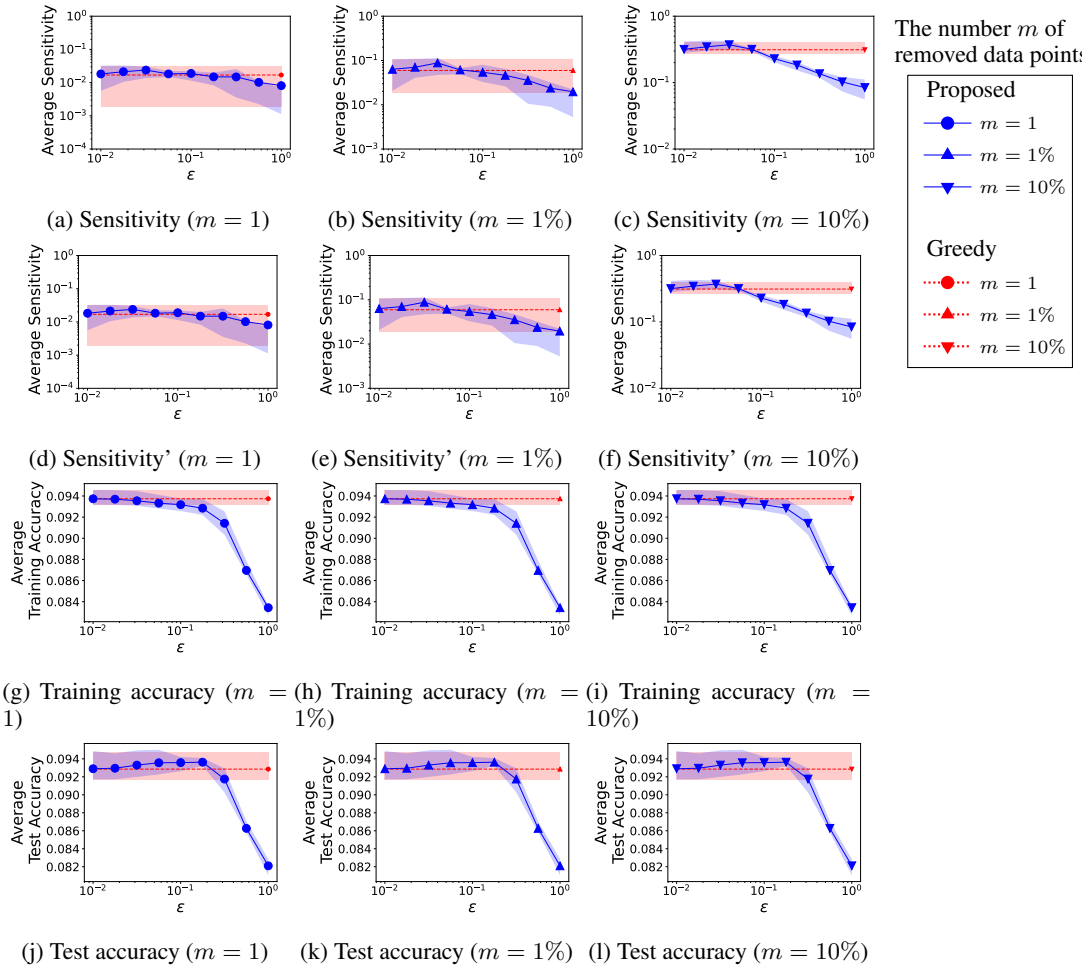
(g) Training accuracy ($m = 1$) (h) Training accuracy ($m = 1\%$) (i) Training accuracy ($m = 10\%$)

(j) Test accuracy ($m = 1$) (k) Test accuracy ($m = 1\%$) (l) Test accuracy ($m = 10\%$)

Figure 5: Detailed results on average sensitivity and accuracy of the trained trees over different ϵ on breast cancer. The figures named Sensitivity and Sensitivity' are the average sensitivity computed using the original distance and the relaxed distance, respectively.

D.2 Test Accuracy

For the experiments in Section 7, Figure 7 shows the trade-off curves between average sensitivity and test accuracy when ϵ is changed.



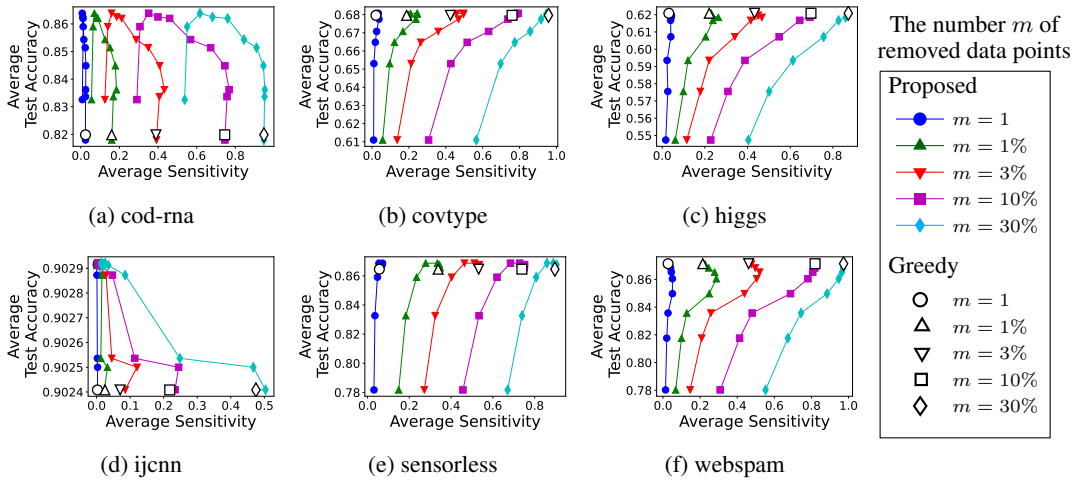
(a) Sensitivity ($m = 1$) (b) Sensitivity ($m = 1\%$) (c) Sensitivity ($m = 10\%$)

(d) Sensitivity' ($m = 1$) (e) Sensitivity' ($m = 1\%$) (f) Sensitivity' ($m = 10\%$)

(g) Training accuracy ($m = 1$) (h) Training accuracy ($m = 1\%$) (i) Training accuracy ($m = 10\%$)

(j) Test accuracy ($m = 1$) (k) Test accuracy ($m = 1\%$) (l) Test accuracy ($m = 10\%$)

Figure 6: Detailed results on average sensitivity and accuracy of the trained trees over different ϵ on diabetes. The figures named Sensitivity and Sensitivity' are the average sensitivity computed using the original distance and the relaxed distance, respectively.



(a) cod-rna (b) covtype (c) higgs

(d) ijcnn (e) sensorless (f) webspam

Figure 7: Trade-off curves between average sensitivity and test accuracy when ϵ is changed. We varied the number of training data points to be removed from one to 30% of the sampled training data. White markers denote the results for the greedy tree learning.