A Detail of the Greedy Partitioning Algorithm

In this section, we provide detail for the greedy partitioning algorithm.

In our greedy Go-CART procedure, we start from the coarsest partition $\mathcal{X} = [0,1]^d$ and then compute the decrease of the held-out risk by dyadically splitting each hyperrectangle $\mathcal{A}$ along the dimension $k \in \{1, \dots d\}$. We select the k^* which leads to the maximum drop of the held-out risk. More precisely, let $\mathrm{sl}_k(\mathcal{A})$ be the side length of $\mathcal{A}$ on the dimension k. If $\mathrm{sl}_k(\mathcal{A}) > 2^{-K}$, where $K = \log_2 N$, we dyadically split $\mathcal{A}$ along the dimension k. In this case, let $\mathcal{A}_L^{(k)}$ and $\mathcal{A}_R^{(k)}$ be the two resulted sub-hyperrectangles. The drop of the held-out risk takes the form:

$$\begin{aligned}
&\Delta \widehat{R}^{(k)}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}}) \\
&= \quad \widehat{R}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}}) - \widehat{R}_{\mathrm{out}}(\mathcal{A}_L^{(k)}, \widehat{\mu}_{\mathcal{A}_L^{(k)}}, \widehat{\Omega}_{\mathcal{A}_L^{(k)}}) - \widehat{R}_{\mathrm{out}}(\mathcal{A}_R^{(k)}, \widehat{\mu}_{\mathcal{A}_R^{(k)}}, \widehat{\Omega}_{\mathcal{A}_R^{(k)}}). \qquad (8)
\end{aligned}$$

Note that if splitting any dimension k of $\mathcal{A}$ leads to an increase of the risk, we set a boolean variable $S(\mathcal{A}) = \mathrm{FALSE}$ which indicates that the partition element $\mathcal{A}$ should no longer be split and hence $\mathcal{A}$ should be an partition element of $\Pi(T)$. The greedy Go-CART as presented in Algorithm 1 recursively applies the previous procedure to split each partition element until all the partition elements cannot be further split. Note that we also record the dyadic partition tree structure in the implementations.

Algorithm 1 Greedy Dyadic Partitioning Tree Learning Algorithm

```
Input: training data $\{x_i, y_i\}_{i=1}^{n_1}$, held-out data validation $\{x_i', y_i'\}_{i=1}^{n_2}$, and an integer $K$
  Start from the $\mathcal{X} = [0,1]^d$. Set the boolean variable $S(\mathcal{X}) = \mathrm{TRUE}$ and estimate $\widehat{\mu}_{\mathcal{X}}, \widehat{\Omega}_{\mathcal{X}}$
  while exists a hyperrectangle $\mathcal{A}$ such that $S(\mathcal{A}) = \mathrm{TRUE}$ do
    for all dimension $k \in \{1, \dots d\}$ do
      if $\mathrm{sl}_k(\mathcal{A}) \geq 2^{-K+1}$ then
        Calculate $\Delta \widehat{R}^{(k)}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}})$ according to (8)
      else
        Set $\Delta \widehat{R}^{(k)}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}}) = -\infty$
      end if
    end for
    Determine the best splitting dimension $k^* = \arg\max_{k \in \{1,\dots,d\}} \Delta \widehat{R}^{(k)}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}})$
    if $\Delta \widehat{R}^{(k^*)}_{\mathrm{out}}(\mathcal{A}, \widehat{\mu}_{\mathcal{A}}, \widehat{\Omega}_{\mathcal{A}}) > 0$ then
      Dyadically split $\mathcal{A}$ along the dimension $k^*$ which leads to two new hyperrectangles $\mathcal{A}_L^{(k^*)}$ and $\mathcal{A}_R^{(k^*)}$.
      Estimate $\widehat{\mu}_{\mathcal{A}_L^{(k^*)}}, \widehat{\Omega}_{\mathcal{A}_L^{(k^*)}}, \widehat{\mu}_{\mathcal{A}_R^{(k^*)}}, \widehat{\Omega}_{\mathcal{A}_R^{(k^*)}}$ and set $S(\mathcal{A}_L^{(k^*)}) = S(\mathcal{A}_R^{(k^*)}) = \mathrm{TRUE}$
    else
      Set $S(\mathcal{A}) = \mathrm{FALSE}$ and put $\mathcal{A}$ into the final partition set
    end if
  end while
Output: the obtained partition set $\Pi(\widehat{T}) = \{\mathcal{X}_j\}_{j=1}^{m_{\widehat{T}}}$ and the corresponding DPT $\widehat{T}$ with the estimated $\widehat{\mu}_{\mathcal{X}_j}$, $\widehat{\Omega}_{\mathcal{X}_j}$ for each partition element $\mathcal{X}_j$.
```

B Proofs of Technical Results

Proof of Theorem 1. For any $T \in \mathcal{T}_N$, we denote

$$S_{j,n} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{\mathcal{X}_j})(y_i - \mu_{\mathcal{X}_j})^T \cdot I(x_i \in \mathcal{X}_j) \qquad (9)$$

$$\bar{S}_j = \mathbb{E}(Y - \mu_{\mathcal{X}_j})(Y - \mu_{\mathcal{X}_j})^T \cdot I(X \in \mathcal{X}_j). \qquad (10)$$

We then have

$$
\begin{aligned}
&\left|R(T,\mu_T,\Omega_T)-\widehat{R}(T,\mu_T,\Omega_T)\right| \\
\leq\;& \left|\sum_{j=1}^{m}\operatorname{tr}\left[\Omega_{\mathcal{X}_j}\left(S_{j,n}-\bar{S}_j\right)\right]\right| + \left|\sum_{j=1}^{m}\log\left|\Omega_{\mathcal{X}_j}\right|\cdot\left[\frac{1}{n}\sum_{i=1}^{n}I(x_i\in\mathcal{X}_j)-\mathbb{E}I(X\in\mathcal{X}_j)\right]\right| \quad (11)\\
\leq\;& \underbrace{\sum_{j=1}^{m}\|\Omega_{\mathcal{X}_j}\|_1\cdot\left\|S_{j,n}-\bar{S}_j\right\|_\infty}_{A_1} + \underbrace{\sum_{j=1}^{m}\left|\log\left|\Omega_{\mathcal{X}_j}\right|\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}I(x_i\in\mathcal{X}_j)-\mathbb{E}I(X\in\mathcal{X}_j)\right|}_{A_2}. \quad (12)
\end{aligned}
$$

We now analyze the terms A_1 and A_2 separately.

For A_2, using the Hoeffding's inequality, for $\epsilon>0$, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}I(x_i\in\mathcal{X}_j)-\mathbb{E}I(X\in\mathcal{X}_j)\right|>\epsilon\right)\leq 2\exp\left(-2n\epsilon^2\right), \tag{13}$$

which implies that,

$$\mathbb{P}\left(\sup_{T\in\mathcal{T}_N}\left|\frac{1}{n}\sum_{i=1}^{n}I(x_i\in\mathcal{X}_j)-\mathbb{E}I(X\in\mathcal{X}_j)\right|/\epsilon_T>1\right)\leq 2\sum_{T\in\mathcal{T}_N}\exp\left(-2n\epsilon_T^2\right), \tag{14}$$

where ϵ_T means ϵ is a function of T. For any $\delta\in(0,1)$, we have, with probability at least $1-\delta/4$,

$$\forall T\in\mathcal{T}_N,\ \left|\frac{1}{n}\sum_{i=1}^{n}I(x_i\in\mathcal{X}_j)-\mathbb{E}I(X\in\mathcal{X}_j)\right|\leq\sqrt{\frac{[[T]]\log 2+\log(8/\delta)}{2n}} \tag{15}$$

where $[[T]]>0$ is the prefix code of T given in (4).

From Assumption 1, since $\Omega_{\mathcal{X}_j}\in\Lambda_j$, we have that

$$\max_{1\leq j\leq m_T}\log\left|\Omega_{\mathcal{X}_j}\right|\leq L_n \tag{16}$$

Therefore, with probability at least $1-\delta/4$,

$$A_2\leq L_n m_T\sqrt{\frac{[[T]]\log 2+\log(8/\delta)}{2n}}. \tag{17}$$

Next, we analyze the term A_1. It's obvious that

$$\max_{1\leq j\leq m_T}\|\Omega_{\mathcal{X}_j}\|_1\leq L_n. \tag{18}$$

We only need to bound the term $\left\|S_{j,n}-\bar{S}_j\right\|_\infty$. By Assumption 2 and the union bound, we have, for any $\epsilon>0$,

$$
\begin{aligned}
&\mathbb{P}\left(\left\|S_{j,n}-\bar{S}_j\right\|_\infty>\epsilon\right)\\
\leq\;& \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}y_iy_i^TI(x_i\in\mathcal{X}_j)-\mathbb{E}\big[YY^TI(X\in\mathcal{X}_j)\big]\right\|_\infty>\frac{\epsilon}{4}\right) \quad (19)\\
&+\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}y_i\mu_{\mathcal{X}_j}^TI(x_i\in\mathcal{X}_j)-\mathbb{E}\big[Y\mu_{\mathcal{X}_j}^TI(X\in\mathcal{X}_j)\big]\right\|_\infty>\frac{\epsilon}{4}\right) \quad (20)\\
&+\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mu_{\mathcal{X}_j}y_i^TI(x_i\in\mathcal{X}_j)-\mathbb{E}\big[\mu_{\mathcal{X}_j}Y^TI(X\in\mathcal{X}_j)\big]\right\|_\infty>\frac{\epsilon}{4}\right) \quad (21)\\
&+\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mu_{\mathcal{X}_j}\mu_{\mathcal{X}_j}^TI(x_i\in\mathcal{X}_j)-\mathbb{E}\big[\mu_{\mathcal{X}_j}\mu_{\mathcal{X}_j}^TI(X\in\mathcal{X}_j)\big]\right\|_\infty>\frac{\epsilon}{4}\right). \quad (22)
\end{aligned}
$$

Using the fact that $\|\mu\|_\infty \leq B$ and the Assumption 2, we can apply Bernstein's exponential inequality on (19), (20), and (21). Also, since the indicator function is bounded, we can apply Hoeffding's inequality on (22). We then obtain:

$$\begin{aligned}
&\mathbb{P}\left(\left\|S_{j,n} - \bar{S}_j\right\|_\infty > \epsilon\right) && (23)\\
&\quad \leq 2p^2 \exp\left(-\frac{1}{32}\left(\frac{n\epsilon^2}{v_2 + M_2\epsilon}\right)\right) + 4p^2 \exp\left(-\frac{1}{32B^2}\left(\frac{n\epsilon^2}{v_1 + M_1\epsilon}\right)\right) + 2p^2 \exp\left(-\frac{2n\epsilon^2}{B^4}\right).
\end{aligned}$$

Therefore, for any $\delta \in (0, 1/4)$, we have, for any $\epsilon \to 0$ as n goes to infinity, with probability at least $1 - \delta/4$:

$$\begin{aligned}
\forall T \in \mathcal{T}_N,\ \left\|S_{j,n} - \bar{S}_j\right\|_\infty &\leq (8\sqrt{v_2}) \cdot \sqrt{\frac{[[T]]\log 2 + 2\log p + \log(24/\delta)}{n}} && (24)\\
&+ (8B\sqrt{v_1}) \cdot \sqrt{\frac{[[T]]\log 2 + 2\log p + \log(48/\delta)}{n}} && (25)\\
&+ B^2 \cdot \sqrt{\frac{[[T]]\log 2 + 2\log p + \log(24/\delta)}{2n}} && (26)
\end{aligned}$$

Combined with (18), we get that

$$A_1 \leq C_1 L_n m_T \sqrt{\frac{[[T]]\log 2 + 2\log p + \log(48/\delta)}{n}}. \tag{27}$$

where $C_1 = 8\sqrt{v_2} + 8B\sqrt{v_1} + B^2$.

Since the above analysis holds uniformly over the whole space of $\mathcal{T}_N$, when choosing

$$\text{pen}(T) = (C_1 + 1)L_n m_T \sqrt{\frac{[[T]]\log 2 + 2\log p + \log(48/\delta)}{n}}, \tag{28}$$

we then get, with probability at least $1 - \delta/2$,

$$\sup_{T\in\mathcal{T}_N, \mu_j\in M_j, \Omega_j\in\Lambda_j} \left|R(T, \mu_T, \Omega_T) - \widehat{R}(T, \mu_T, \Omega_T)\right| \leq \text{pen}(T) \tag{29}$$

for large enough n.

Given the uniform deviation inequality in (29), we have, for large enough n: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\begin{aligned}
R(\widehat{T}, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) &\leq \widehat{R}(\widehat{T}, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) + \text{pen}(\widehat{T}) && (30)\\
&= \inf_{T\in\mathcal{T}_N, \mu_{\mathcal{X}_j}\in M_j, \Omega_{\mathcal{X}_j}\in\Lambda_j} \left\{\widehat{R}(T, \mu_T, \Omega_T) + \text{pen}(T)\right\} && (31)\\
&\leq \inf_{T\in\mathcal{T}_N} \left\{\widehat{R}(T, \mu_T^*, \Omega_T^*) + \text{pen}(T)\right\} && (32)\\
&\leq \inf_{T\in\mathcal{T}_N} \left\{R(T, \mu_T^*, \Omega_T^*) + 2\text{pen}(T)\right\} && (33)\\
&= \inf_{T\in\mathcal{T}_N} \left\{\inf_{\mu_{\mathcal{X}_j}\in M_j, \Omega_{\mathcal{X}_j}\in\Lambda_j} (R(T, \mu_T, \Omega_T) + 2\text{pen}(T)\right\}. && (34)
\end{aligned}$$

The desired result of the theorem follows by subtracting R^* from both sides. □

Proof of Theorem 2. From (29), we have, for large enough n, on the dataset $\mathcal{D}_1$, with probability at least $1 - \delta/4$

$$\sup_{T\in\mathcal{T}_N, \mu_j\in M_j, \Omega_j\in\Lambda_j} \left|R(T, \mu_T, \Omega_T) - \widehat{R}(T, \mu_T, \Omega_T)\right| \leq \phi_n(T). \tag{35}$$

Follow the same line of analysis, we can also get, on the validation dataset $\mathcal{D}_2$, with probability at least $1 - \frac{\delta}{4}$.

$$\sup_{T\in\mathcal{T}_N} \left|R(T, \widehat{\mu}_T, \widehat{\Omega}_T) - \widehat{R}_{\text{out}}(T, \widehat{\mu}_T, \widehat{\Omega}_T)\right| \leq \phi_n(T) \tag{36}$$

for large enough n. Where $\widehat{\mu}_T, \widehat{\Omega}_T$ are as defined in (7).

Using the fact that

$$\widehat{T} = \mathrm{argmin}_{T\in\mathcal{T}_N} \widehat{R}_{\mathrm{out}}(T, \widehat{\mu}_T, \widehat{\Omega}_T), \tag{37}$$

we have, for large enough n: for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned}
R(\widehat{T}, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) &\leq \widehat{R}_{\mathrm{out}}(\widehat{T}, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) + \phi_n(\widehat{T}) && (38)\\
&= \inf_{T\in\mathcal{T}_N} \widehat{R}_{\mathrm{out}}(T, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) + \phi_n(\widehat{T}) && (39)\\
&\leq \inf_{T\in\mathcal{T}_N} \Big\{ R(T, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) + \phi_n(T) \Big\} + \phi_n(\widehat{T}) && (40)\\
&\leq \inf_{T\in\mathcal{T}_N} \Big\{ \widehat{R}(T, \widehat{\mu}_{\widehat{T}}, \widehat{\Omega}_{\widehat{T}}) + \phi_n(T) + \phi_n(T) \Big\} + \phi_n(\widehat{T}) && (41)\\
&= \inf_{T\in\mathcal{T}_N} \left\{ 3\phi_n(T) + \inf_{\mu_{\mathcal{X}_j}\in M_j, \Omega_{\mathcal{X}_j}\in\Lambda_j} R(T, \mu_T, \Omega_T) \right\} + \phi_n(\widehat{T}).
\end{aligned}$$

The result follows by subtracting R^* on both sides. □

Proof of Theorem 3. For any $T \in \mathcal{T}_N$, $\Pi(T^*) \not\subseteq \Pi(T)$, there must exists a subregion $\mathcal{X}' \in \Pi(T)$ such that there does not exist any $\mathcal{A} \in \Pi(T^*)$ which makes $\mathcal{X}' \subset \mathcal{A}$. In this case, we can find a minimal class of disjoint subregions $\{\widetilde{\mathcal{X}}_1, \ldots, \widetilde{\mathcal{X}}_{k'}\} \in \Pi(T^*)$, such that

$$\mathcal{X}' \subset \cup_{i=1}^{k'} \widetilde{\mathcal{X}}_i, \tag{42}$$

where $k' \geq 2$. We define $\mathcal{X}_i^* = \widetilde{X}_i \cap \mathcal{X}'$ for $i = 1, \ldots, k'$. Then we have

$$\mathcal{X}' = \cup_{i=1}^{k'} \mathcal{X}_i^*. \tag{43}$$

Let $\{\mu^*_{\mathcal{X}_j^*}, \Omega^*_{\mathcal{X}_j^*}\}_{j=1}^{k'}$ be the corresponding true parameters on $\widetilde{\mathcal{X}}_1, \ldots, \widetilde{\mathcal{X}}_{k'}$. We denote $R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*})$ to be the risk of $\mu^*_{T^*}$ and $\Omega^*_{T^*}$ on the subregion $\mathcal{X}'$, then

$$\begin{aligned}
R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*}) &= \sum_{j=1}^{k'} \mathbb{E}\Bigg[\Big(\mathrm{tr}\left[\Omega^*_{\mathcal{X}_j^*}\left((Y-\mu^*_{\mathcal{X}_j^*})(Y-\mu^*_{\mathcal{X}_j^*})^T\right)\right] - \log|\Omega^*_{\mathcal{X}_j^*}|\Big) \cdot I(X\in\mathcal{X}_j^*)\Bigg]\\
&= p\mathbb{P}\left(X\in\mathcal{X}'\right) - \sum_{j=1}^{k'} \mathbb{P}\left(X\in\mathcal{X}_j^*\right)\log|\Omega^*_{\mathcal{X}_j^*}|. && (44)
\end{aligned}$$

Since the DPT T does not further partition $\mathcal{X}'$, we have, for any $\mu_T, \Omega_T \in \mathcal{M}_T$:

$$\begin{aligned}
R(\mathcal{X}', \mu_T, \Omega_T) &= \sum_{j=1}^{k'} \mathbb{E}\Bigg[\Big(\mathrm{tr}\left[\Omega_T\left((Y-\mu_T)(Y-\mu_T)^T\right)\right] - \log|\Omega_T|\Big) \cdot I(X\in\mathcal{X}_j^*)\Bigg]\\
&= \sum_{j=1}^{k'} \mathbb{E}\Bigg[\Big(\mathrm{tr}\left[\Omega_T\left((Y-\mu_T)(Y-\mu_T)^T\right)\right]\Big) \cdot I(X\in\mathcal{X}_j^*)\Bigg] - \mathbb{P}(X\in\mathcal{X}')\log|\Omega_T|.
\end{aligned}$$

Since

$$\begin{aligned}
(Y-\mu_T)(Y-\mu_T)^T = (Y-\mu^*_{\mathcal{X}_j^*})(Y-\mu^*_{\mathcal{X}_j^*})^T + (Y-\mu^*_{\mathcal{X}_j^*})(\mu^*_{\mathcal{X}_j^*}-\mu_T)^T\\
+ (\mu^*_{\mathcal{X}_j^*}-\mu_T)(Y-\mu^*_{\mathcal{X}_j^*})^T + (\mu^*_{\mathcal{X}_j^*}-\mu_T)(\mu^*_{\mathcal{X}_j^*}-\mu_T)^T. && (45)
\end{aligned}$$

This implies that

$$\begin{aligned}
&\sum_{j=1}^{k'} \mathbb{E}\Bigg[\Big(\mathrm{tr}\left[\Omega_T\left((Y-\mu_T)(Y-\mu_T)^T\right)\right]\Big) \cdot I(X\in\mathcal{X}_j^*)\Bigg]\\
&= \sum_{j=1}^{k'} \mathbb{P}\left(X\in\mathcal{X}_j^*\right)\left[\mathrm{tr}(\Omega_T(\Omega_j^*)^{-1}) + \mathrm{tr}(\Omega_T(\mu^*_{\mathcal{X}_j^*}-\mu_T)(\mu^*_{\mathcal{X}_j^*}-\mu_T)^T)\right]. && (46)
\end{aligned}$$

Using the fact that

$$R(\mathcal{X}', \mu_T, \Omega_T) \geq \max\{R(\mathcal{X}', \mu^*_{T^*}, \Omega_T), R(\mathcal{X}', \mu_T, \Omega^*_{T^*})\}, \tag{47}$$

We consider the two cases on the R.H.S. separately.

Case 1: The μ's are different.

we know that

$$\begin{aligned}
&\inf_{\mu_T, \Omega_T \in \mathcal{M}_T} R(\mathcal{X}', \mu_T, \Omega_T) - R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*}) \qquad (48)\\
&\quad\geq \inf_{\mu_T} R(\mathcal{X}', \mu_T, \Omega^*_{T^*}) - R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*})\\
&\quad= \inf_{\mu_T} \sum_{j=1}^{k'} \mathbb{P}\left(X \in \mathcal{X}^*_j\right) (\mu^*_{\mathcal{X}^*_j} - \mu_T)^T \Omega^*_{\mathcal{X}^*_j} (\mu^*_{\mathcal{X}^*_j} - \mu_T)\\
&\quad\geq c_1 c_2 \inf_{\mu_T} \sum_{j=1}^{k'} \|\mu^*_{\mathcal{X}^*_j} - \mu_T\|_2^2
\end{aligned}$$

where the last inequality follows from that fact that $\rho_{\min}(\Omega^*_{\mathcal{X}^*_j}) \geq c_1, \mathbb{P}\left(X \in \mathcal{X}^*_j\right) \geq c_2$. It's easy to see that a lower bound of the last term is achieved at $\bar{\mu}_T$,

$$\bar{\mu}_T = \frac{1}{k'} \sum_{j=1}^{k'} \mu^*_{\mathcal{X}^*_j}. \tag{49}$$

Furthermore, for any two DPTs T and T', if $\Pi(T) \subset \Pi(T')$. it's obvious that

$$\inf_{\mu_T, \Omega_T \in \mathcal{M}_T} R(T, \mu_T, \Omega_T) \geq \inf_{\mu_{T'}, \Omega_{T'} \in \mathcal{M}_{T'}} R(T', \mu_{T'}, \Omega_{T'}). \tag{50}$$

Therefore, in the sequel, without loss of generality, we only need to consider the case $k' = 2$.

The result of this case then follows from the fact that

$$\sum_{j=1}^{2} \|\mu^*_{\mathcal{X}^*_j} - \bar{\mu}_T\|_2^2 = \frac{1}{2} \|\mu_{\mathcal{X}^*_1} - \mu_{\mathcal{X}^*_2}\|_2^2 \geq \frac{c_3}{2}. \tag{51}$$

Case 2: The Ω's are different.

In this case, we have

$$\inf_{\mu_T, \Omega_T \in \mathcal{M}_T} R(\mathcal{X}', \mu_T, \Omega_T) - R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*}) \geq \inf_{\Omega_T} R(\mathcal{X}', \mu^*_{T^*}, \Omega_T) - R(\mathcal{X}', \mu^*_{T^*}, \Omega^*_{T^*})$$

$$\begin{aligned}
&= \inf_{\Omega_T} \sum_{j=1}^{k'} \mathbb{P}\left(X \in \mathcal{X}^*_j\right) \left(\mathrm{tr}\left[\Omega^{-1}_{\mathcal{X}^*_j} (\Omega_T - \Omega^*_{\mathcal{X}^*_j})\right] - \left(\log|\Omega_T| - \log|\Omega^*_{\mathcal{X}^*_j}|\right)\right) && (52)\\
&\geq c_2 \inf_{\Omega_T} \sum_{j=1}^{k'} \left(\mathrm{tr}\left[\Omega^{-1}_{\mathcal{X}^*_j} (\Omega_T - \Omega^*_{\mathcal{X}^*_j})\right] - \left(\log|\Omega_T| - \log|\Omega^*_{\mathcal{X}^*_j}|\right)\right) && (53)\\
&\geq c_2 \inf_{\Sigma_T} \sum_{j=1}^{k'} \left(\mathrm{tr}\left[\Sigma^*_{\mathcal{X}^*_j} (\Sigma_T^{-1} - \Omega^*_{\mathcal{X}^*_j})\right] + \log \frac{|\Sigma_T|}{|\Sigma^*_{\mathcal{X}^*_j}|}\right) && (54)\\
&= c_2 \inf_{\Sigma_T} \sum_{j=1}^{k'} \left(\mathrm{tr}\left(\Sigma^*_{\mathcal{X}^*_j} \Sigma_T^{-1}\right) + \log \frac{|\Sigma_T|}{|\Sigma^*_{\mathcal{X}^*_j}|} - p\right) && (55)
\end{aligned}$$

where $\Sigma_T = \Omega_T^{-1}$

As discussed before, we only need to consider the case $k' = 2$, a lower bound of the last term is achieved at

$$\bar{\Sigma}_T = \frac{\Sigma_{\mathcal{X}^*_1} + \Sigma_{\mathcal{X}^*_2}}{2} \tag{56}$$

Plug-in $\bar{\Sigma}_T$, we get

$$\inf_{\Sigma_T}\sum_{j=1}^{2}\left(\mathrm{tr}\left(\Sigma^*_{\mathcal{X}_j^*}\Sigma_T^{-1}\right)+\log\frac{|\Sigma_T|}{|\Sigma^*_{\mathcal{X}_j^*}|}-p\right)\geq\sum_{j=1}^{2}\left(\mathrm{tr}\left(\Sigma^*_{\mathcal{X}_j^*}\bar{\Sigma}_T^{-1}\right)+\log\frac{|\bar{\Sigma}_T|}{|\Sigma^*_{\mathcal{X}_j^*}|}-p\right)$$

$$= \mathrm{tr}\left((2\bar{\Sigma}_T-\Sigma_{\mathcal{X}_2^*})\bar{\Sigma}_T^{-1}\right)+\log\frac{|\bar{\Sigma}_T|}{|\Sigma_{\mathcal{X}_1^*}|}-p+\mathrm{tr}\left(\Sigma_{\mathcal{X}_2^*}\bar{\Sigma}_T^{-1}\right)+\log\frac{|\bar{\Sigma}_T|}{|\Sigma_{\mathcal{X}_2^*}|}-p \tag{57}$$

$$= \log\frac{|\bar{\Sigma}_T|}{|\Sigma_{\mathcal{X}_1^*}|}+\log\frac{|\bar{\Sigma}_T|}{|\Sigma_{\mathcal{X}_2^*}|} \tag{58}$$

$$= 2\log\left|\frac{\Sigma_{\mathcal{X}_1^*}+\Sigma_{\mathcal{X}_2^*}}{2}\right|-\log|\Sigma_{\mathcal{X}_1^*}|-\log|\Sigma_{\mathcal{X}_2^*}| \tag{59}$$

$$\geq c_4. \tag{60}$$

where the last inequality follows from the given assumption.

Therefore, we have

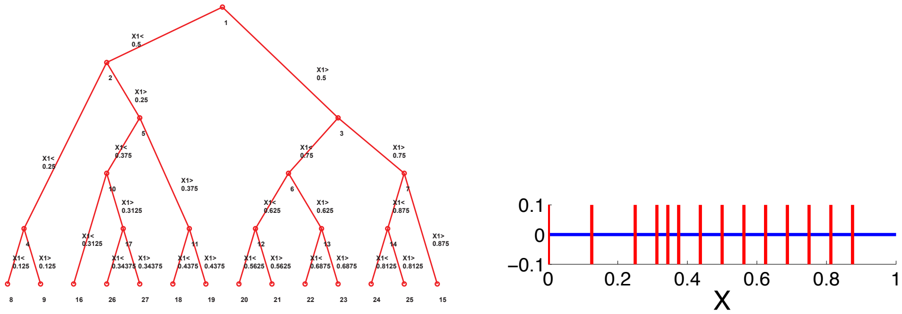
$$\inf_{\mu_T,\Omega_T\in\mathcal{M}_T} R(\mathcal{X}',\mu_T,\Omega_T)-R(\mathcal{X}',\mu^*_{T^*},\Omega^*_{T^*})\geq c_2c_4. \tag{61}$$

The desired result of theorem is obtained by combining the above discussed two cases. □

C More Simulations

To further demonstrate the recovery quality of our method, in this section we simulate data where the ground true conditional covariance matrix is continuous in X. We compare the graphs estimated by our method to the single graph obtained by applying the glasso directly to the entire dataset.

C.1 Chain Structure



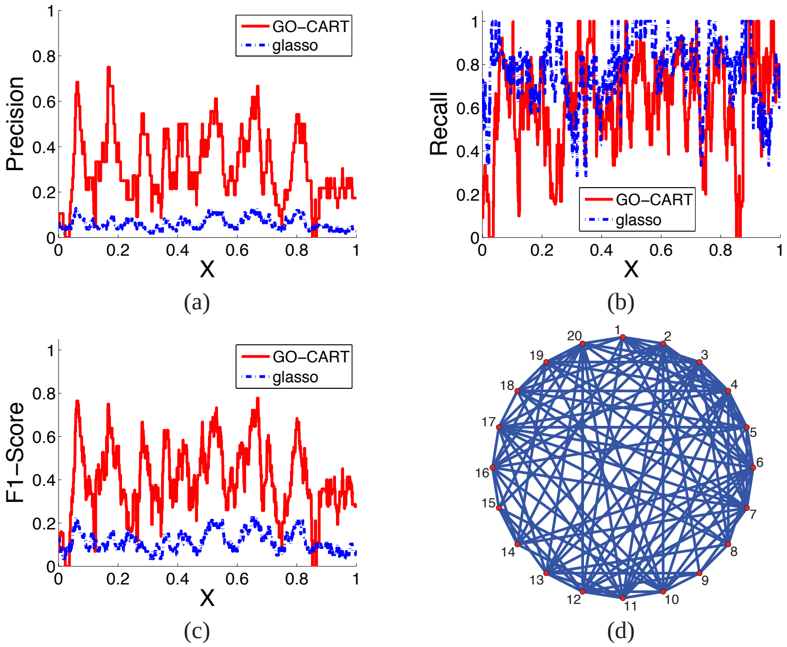
(a) (b)

Figure 3: (a) Learned tree structure; (b) Corresponding partitions

In this subsection, we consider the case where X lies on a one dimensional chain. More precisely, we generate n equally spaced points $x_1,\ldots,x_n\in\mathbb{R}$ with $n=10,000$ on $[0,1]$. We generate an Erdös-Rényi random graph $G^1=(V^1,E^1)$ with the number of vertices $p=20$, the number of edges $|E|=10$ and the maximum node degree to be 4 as the basis. Then we simulate the output $y_1,\ldots,y_n]\in\mathbb{R}^p$ as follows:

1. From $t=2$ to T, we construct the graph $G^t=(V^t,E^t)$ as follows: (a) with probability 0.05, remove one edge from G^{t-1} and (b) with probability 0.05, add one edge to the graph generated in (a). We make sure that the total number of edges is between 5 and 15; and maximum node degree is still 4.
2. For each graph G^t, generate the inverse covariance matrix Ω^t:

$$\Omega^t(i,j)=\begin{cases}1 & \text{if } i=j,\\ 0.245 & \text{if } (i,j)\in E^t,\\ 0 & \text{otherwise},\end{cases}$$



(a) (b)

(c) (d)

Figure 4: Comparison of our algorithm with glasso (a) Precision; (b) Recall; (c) F1-score; (d) Estimated graph by applying glasso on the entire dataset

where 0.245 guarantees the positive definiteness of Ω^t when the maximum degree is 4.

3. For each t, we sample y_t from a multivariate Gaussian distribution with mean $\mu = (0, \dots, 0) \in \mathbb{R}^p$ and covariance matrix $\Sigma^t = (\Omega^t)^{-1}$:

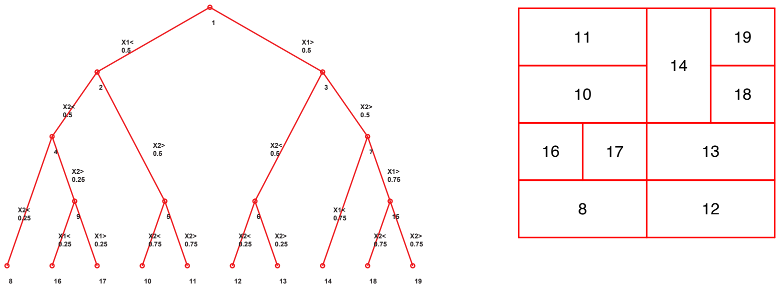
$$y_t \sim N(\mu, \Sigma^t),$$

In addition, we generate an equal-sized held-out dataset in the same manner as described above, using the same μ and Σ^t. We apply our greedy algorithm to learn the dyadic tree structure and corresponding inverse covariance matrices. These are presented in Figure 3.

To examine the recovery quality of the underlying graph structure, we compare our estimated graphs to the one estimated by directly applying glasso to the entire dataset. We present the comparison of precision, recall and F1-score in Figure 4 (a), (b) and (c) respectively. As we can see, out method achieves much higher precision and F1-Score. As for recall, glasso is even slightly better than us because the graphs estimated by glasso on the entire data is very dense as shown in 4 (d). The dense graphs lead to to fewer false negatives (thus large recall values) but many false positives (thus small precision values).

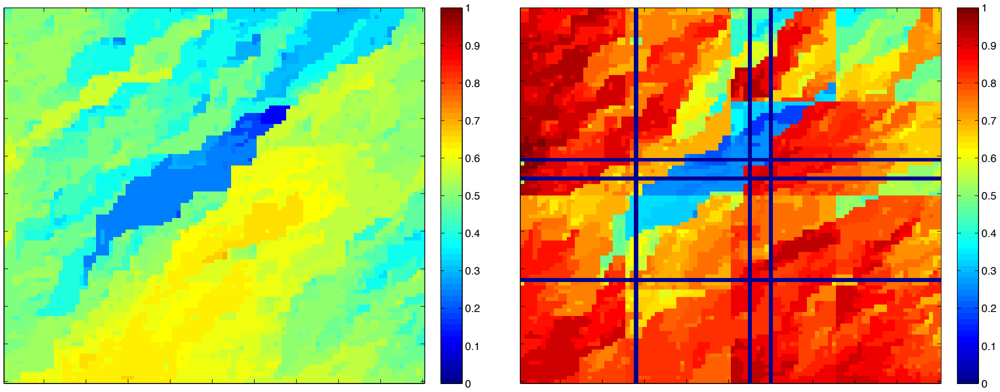
C.2 Two-way Grid Structure

In this section, we apply Go-CART to a two dimensional design X. The underlying graph structures and Y are generated in manner similar to that used in the previous section. In particular, we generate equally spaced $x_1, \dots, x_n \in \mathbb{R}^2$ with $n = 10,000$ on a unit two-way grid $[0, 1]^2$. We generate an Erdös-Rényi random graph $G^{1,1} = (V^{1,1}, E^{1,1})$ with the number of vertices $p = 20$, the number of edges $|E| = 10$ and the maximum node degree to be 4 then construct the graphs for each x along diagonals. More precisely, for each pair of i, j, where $1 \leq i \leq 100$ and $1 \leq j \leq 100$, we randomly select either $G^{i-1,j}$ (if it exists) or $G^{i,j-1}$ (if it exists) with equal probability as the basis graph. Then, we construct the graph $G^{i,j} = (V^{i,j}, E^{i,j})$ by removing one edge and adding one edge with probability 0.05 based on the selected basis graph and taking care that the number of edge is between 5 and 15 and the maximum degree is still 4. With the underlying graph structures, we generate the covariance matrix and output Y in the same way as in the last section.



(a) (b)

Figure 5: (a) Learned tree structure; (b) Learned partitions where the labels correspond to the index of the leaf node in (a)



(a) (b)

Figure 6: (a) Color map of F1-score via applying glasso on the entire dataset; (b) Color map of F1-score learned by our method. Red pixels indicate large values (approaching 1) and blue pixels indicate small values (approaching 0) as shown in the color bar.

We apply the greedy algorithm to learn the dyadic tree structure and corresponding inverse covariance matrices, which are presented in Figure 5. We plot the F1-score obtained by glasso on the entire data as compared to our method in Figure 6. As we can see, for most x, our method achieves significantly higher F1-score than directly applying glasso. Note that since the graphs around the middle part of the diagonal (line connecting $[0, 1]$ and $[1, 0]$) have the most variability, the F1-scores for both methods are relatively low in this region.