Learning (from) Deep Hierarchical Structure among Features: Supplementary Material

Proof for Theorem 1

Proof. We prove the convexity via the first-order condition. For any $\bar{w} \in \mathbb{R}$ and $\bar{\mathbf{z}} \geq \mathbf{0}$, We can easily compute $\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial w} = \frac{2\bar{w}}{\prod_{i=1}^m \bar{z}_i^{\theta_i}}$ and $\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial z_j} = -\frac{\theta_j \bar{w}^2}{\bar{z}_j \prod_{i=1}^m \bar{z}_i^{\theta_i}}$. Then for any $\hat{w}, \bar{w} \in \mathbb{R}$ and $\hat{\mathbf{z}}, \bar{\mathbf{z}} \geq \mathbf{0}$, we have

$$
\begin{aligned}
& f(\hat{w},\hat{\mathbf{z}}) - f(\bar{w},\bar{\mathbf{z}}) - (\hat{w}-\bar{w})\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial w} - \sum_{i=1}^m (\hat{z}_i - \bar{z}_i)\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial z_i} \\
=& \frac{\hat{w}^2}{\prod_{i=1}^m \hat{z}_i^{\theta_i}} - \frac{\bar{w}^2}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} - \frac{2\bar{w}(\hat{w}-\bar{w})}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} + \frac{\bar{w}^2 \sum_{j=1}^m \theta_j (\frac{\hat{z}_j}{\bar{z}_j} - 1)}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} \\
=& \frac{1}{\prod_{i=1}^m \hat{z}_i^{\theta_i}} \hat{w}^2 - \frac{2}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} \bar{w}\hat{w} + \frac{\sum_{j=1}^m \theta_j \frac{\hat{z}_j}{\bar{z}_j}}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} \bar{w}^2,
\end{aligned}
$$

where the second equality holds due to the property of $\boldsymbol{\theta}$ that $\sum_{i=1}^m \theta_i = 1$. Since $\ln(\cdot)$ is concave and $\boldsymbol{\theta}$ satisfies that $\theta_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m \theta_i = 1$, we can have the following inequalities:

$$
\begin{aligned}
& \sum_{i=1}^m \theta_i \ln \frac{\hat{z}_i}{\bar{z}_i} \leq \ln \left(\sum_{j=1}^m \frac{\hat{z}_j}{\bar{z}_j} \theta_j \right) \\
\Rightarrow & \sum_{i=1}^m \theta_i \ln \hat{z}_i \leq \sum_{i=1}^m \theta_i \ln \bar{z}_i + \ln \left(\sum_{j=1}^m \frac{\hat{z}_j}{\bar{z}_j} \theta_j \right) \\
\Rightarrow & \prod_{i=1}^m \hat{z}_i^{\theta_i} \leq \left(\prod_{i=1}^m \bar{z}_i^{\theta_i} \right) \left(\sum_{j=1}^m \frac{\hat{z}_j}{\bar{z}_j} \theta_j \right) \\
\Rightarrow & \frac{\prod_{i=1}^m \hat{z}_i^{\theta_i}}{\left(\prod_{i=1}^m \bar{z}_i^{\theta_i} \right)^2} \leq \frac{\sum_{j=1}^m \frac{\hat{z}_j}{\bar{z}_j} \theta_j}{\prod_{i=1}^m \bar{z}_i^{\theta_i}}.
\end{aligned}
$$

Then we can have

$$
\begin{aligned}
& f(\hat{w},\hat{\mathbf{z}}) - f(\bar{w},\bar{\mathbf{z}}) - (\hat{w}-\bar{w})\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial w} - \sum_{i=1}^m (\hat{z}_i - \bar{z}_i)\frac{\partial f(\bar{w},\bar{\mathbf{z}})}{\partial z_i} \\
=& \frac{\left(\hat{w} - \frac{\prod_{i=1}^m \hat{z}_i^{\theta_i}}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} \bar{w} \right)^2}{\prod_{i=1}^m \hat{z}_i^{\theta_i}} + \left(\frac{\sum_{j=1}^m \theta_j \frac{\hat{z}_j}{\bar{z}_j}}{\prod_{i=1}^m \bar{z}_i^{\theta_i}} - \frac{\prod_{i=1}^m \hat{z}_i^{\theta_i}}{\left(\prod_{i=1}^m \bar{z}_i^{\theta_i} \right)^2} \right) \bar{w}^2 \\
\geq & 0,
\end{aligned}
$$

in which we reach the conclusion. □

Proof for Theorem 2

Proof. It is obvious that the first and third terms of the objective function in problem (1) as well as the linear constraints are convex. Moreover, based on Theorem 1, each summand in the second term of the objective function in problem (1) is jointly convex with respect to $\mathbf{w}$ and $\boldsymbol{\sigma}$, making the whole problem convex. □

Proof for Theorem 3

Proof. We first rewrite problem (1) where $m = 3$ as

$$
\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\sigma}} & \frac{1}{n} \sum_{i=1}^n l(\mathbf{x}_i, y_i, \mathbf{w}) + \frac{\lambda_1}{2} \sum_{i=1}^{s_1} \sum_{j \in \mathcal{C}_i^1} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} \\
& + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2 \\
\text{s.t.} & \sum_{i=1}^{s_i} d_i^{(1)} \sigma_{1,i}^{(1)} = 1, \sum_{i=1}^{s_1} \sum_{j \in \mathcal{C}_i^1} d_j^{(2)} \sigma_{i,j}^{(2)} = 1, \sum_{j=1}^{s_2} \sum_{k \in \mathcal{C}_j^2} \sigma_{j,k}^{(3)} = 1 \\
& \sigma_{1,i}^{(1)} \geq 0 \, \forall i, \; \sigma_{i,j}^{(2)} \geq 0 \, \forall i,j, \; \sigma_{j,k}^{(3)} \geq 0 \, \forall j,k. \qquad (13)
\end{aligned}
$$

The Lagrangian of problem (13) is

$$
\begin{aligned}
\mathcal{L} = & \frac{1}{n} \sum_{i=1}^n l(\mathbf{x}_i, y_i, \mathbf{w}) + \frac{\lambda_1}{2} \sum_{i=1}^{s_1} \sum_{j \in \mathcal{C}_i^1} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} \\
& + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2 + \nu_1 (\sum_{i=1}^{s_i} d_i^{(1)} \sigma_{1,i}^{(1)} - 1) + \nu_2 (\sum_{i=1}^{s_1} \sum_{j \in \mathcal{C}_i^1} d_j^{(2)} \sigma_{i,j}^{(2)} - 1) \\
& + \nu_3 (\sum_{j=1}^{s_2} \sum_{k \in \mathcal{C}_j^2} \sigma_{j,k}^{(3)} - 1) - \sum_{i=1}^{s_i} \varepsilon_{1,i}^{(1)} \sigma_{1,i}^{(1)} - \sum_{i=1}^{s_i} \sum_{j \in \mathcal{C}_i^1} \varepsilon_{i,j}^{(2)} \sigma_{i,j}^{(2)} \\
& - \sum_{j=1}^{s_2} \sum_{k \in \mathcal{C}_j^2} \varepsilon_{j,k}^{(3)} \sigma_{j,k}^{(3)}.
\end{aligned}
$$

The optimality conditions for $\sigma_{1,i}^{(1)}$, $\sigma_{i,j}^{(2)}$ and $\sigma_{j,k}^{(3)}$ are

$$
\frac{\partial \mathcal{L}}{\partial \sigma_{1,i}^{(1)}} = 0 \Longrightarrow d_i^{(1)} \nu_1 - \varepsilon_{1,i}^{(1)} - \frac{\lambda_1}{6} \sum_{j \in \mathcal{C}_i^1} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{4}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} = 0 \qquad (14)
$$

$$
\frac{\partial \mathcal{L}}{\partial \sigma_{i,j}^{(2)}} = 0 \Longrightarrow d_j^{(2)} \nu_2 - \varepsilon_{i,j}^{(2)} - \frac{\lambda_1}{6} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{4}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} = 0 \qquad (15)
$$

$$
\frac{\partial \mathcal{L}}{\partial \sigma_{j,k}^{(3)}} = 0 \Longrightarrow \nu_3 - \varepsilon_{j,k}^{(3)} - \frac{\lambda_1 w_k^2}{6(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{4}{3}}} = 0. \qquad (16)
$$

Multiplying Eqs. (14), (15) and (16) by $\sigma_{1,i}^{(1)}$, $\sigma_{i,j}^{(2)}$ and $\sigma_{j,k}^{(3)}$ respectively gives

$$
d_i^{(1)} \nu_1 \sigma_{1,i}^{(1)} - \frac{\lambda_1}{6} \sum_{j \in \mathcal{C}_i^1} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} = 0 \qquad (17)
$$

$$
d_j^{(2)} \nu_2 \sigma_{i,j}^{(2)} - \frac{\lambda_1}{6} \sum_{k \in \mathcal{C}_j^2} \frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} = 0 \qquad (18)
$$

$$
\nu_3 \sigma_{j,k}^{(3)} - \frac{\lambda_1 w_k^2}{6(\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{1}{3}}} = 0, \qquad (19)
$$

where we use the KKT conditions that $\varepsilon_{1,i}^{(1)} \sigma_{1,i}^{(1)} = 0$, $\varepsilon_{i,j}^{(2)} \sigma_{i,j}^{(2)} = 0$ and $\varepsilon_{j,k}^{(3)} \sigma_{j,k}^{(3)} = 0$. Based on Eq. (19), we have

$$
\lambda_1 w_k^2 = 6\nu_3 (\sigma_{1,i}^{(1)})^{\frac{1}{3}} (\sigma_{i,j}^{(2)})^{\frac{1}{3}} (\sigma_{j,k}^{(3)})^{\frac{4}{3}}.
$$

By plugging this equation into Eqs. (17) and (18), we can get

$$d_i^{(1)}\nu_1\sigma_{1,i}^{(1)} - \nu_3 \sum_{j\in\mathcal{C}_i^1}\sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = 0 \tag{20}$$

$$d_j^{(2)}\nu_2\sigma_{i,j}^{(2)} - \nu_3 \sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = 0. \tag{21}$$

By taking sum over i on Eq. (20), we can get $\nu_1 = \nu_3$ since $\sum_{i=1}^{s_i} d_i^{(1)}\sigma_{1,i}^{(1)} = 1$ and $\sum_{i=1}^{s_1}\sum_{j\in\mathcal{C}_i^1}\sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = \sum_{j=1}^{s_2}\sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = 1$ due to linear equality constraints in problem (13). By taking sum over i and j on Eq. (21), we have $\nu_2 = \nu_3$ since $\sum_{i=1}^{s_1}\sum_{j\in\mathcal{C}_i^1} d_j^{(2)}\sigma_{i,j}^{(2)} = 1$ and $\sum_{i=1}^{s_1}\sum_{j\in\mathcal{C}_i^1}\sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = \sum_{j=1}^{s_2}\sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = 1$ due to linear equality constraints in problem (13). By combining those two results, we have $\nu_1 = \nu_2 = \nu_3$ and hence we use ν to denote them. Based on Eqs. (17-19), we have $\nu > 0$. Based on Eq. (19), we have

$$\sigma_{j,k}^{(3)} = \frac{\lambda_1^{\frac{3}{4}}|w_k|^{\frac{3}{2}}}{6^{\frac{3}{4}}\nu^{\frac{3}{4}}(\sigma_{1,i}^{(1)})^{\frac{1}{4}}(\sigma_{i,j}^{(2)})^{\frac{1}{4}}}. \tag{22}$$

Then based on Eqs. (18), (19) and (22), since $\nu_2 = \nu_3 = \nu > 0$, we have

$$d_j^{(2)}\sigma_{i,j}^{(2)} = \sum_{k\in\mathcal{C}_j^2}\sigma_{j,k}^{(3)} = \frac{\lambda_1^{\frac{3}{4}}}{6^{\frac{3}{4}}\nu^{\frac{3}{4}}(\sigma_{1,i}^{(1)})^{\frac{1}{4}}(\sigma_{i,j}^{(2)})^{\frac{1}{4}}}\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}},$$

leading to

$$\sigma_{i,j}^{(2)} = \frac{\lambda_1^{\frac{3}{5}}}{6^{\frac{3}{5}}\nu^{\frac{3}{5}}(d_j^{(2)})^{\frac{4}{5}}(\sigma_{1,i}^{(1)})^{\frac{1}{5}}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}. \tag{23}$$

Based on Eqs. (17), (18) and (23), since $\nu_1 = \nu_2 = \nu > 0$, we can get

$$\begin{aligned} d_i^{(1)}\sigma_{1,i}^{(1)} &= \sum_{j\in\mathcal{C}_i^1} d_j^{(2)}\sigma_{i,j}^{(2)} \\ &= \frac{\lambda_1^{\frac{3}{5}}}{6^{\frac{3}{5}}\nu^{\frac{3}{5}}(\sigma_{1,i}^{(1)})^{\frac{1}{5}}}\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}, \end{aligned}$$

leading to

$$\sigma_{1,i}^{(1)} = \frac{\lambda_1^{\frac{1}{2}}}{6^{\frac{1}{2}}\nu^{\frac{1}{2}}(d_i^{(1)})^{\frac{5}{6}}}\left(\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{5}{6}}. \tag{24}$$

By plugging Eq. (24) into the first linear equality constraint in problem (13), we can get

$$\nu = \frac{\lambda_1}{6}\left(\sum_{i=1}^{s_1}(d_i^{(1)})^{\frac{1}{6}}\left(\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{5}{6}}\right)^2. \tag{25}$$

Then based on Eqs. (22-25), we can get

$$\begin{aligned} &(\sigma_{1,i}^{(1)})^{\frac{1}{3}}(\sigma_{i,j}^{(2)})^{\frac{1}{3}}(\sigma_{j,k}^{(3)})^{\frac{1}{3}} \\ =&\frac{\lambda_1^{\frac{1}{4}}|w_k|^{\frac{1}{2}}(\sigma_{1,i}^{(1)})^{\frac{1}{3}}(\sigma_{i,j}^{(2)})^{\frac{1}{3}}}{6^{\frac{1}{4}}\nu^{\frac{1}{4}}(\sigma_{1,i}^{(1)})^{\frac{1}{12}}(\sigma_{i,j}^{(2)})^{\frac{1}{12}}} \\ =&\frac{\lambda_1^{\frac{1}{4}}|w_k|^{\frac{1}{2}}}{6^{\frac{1}{4}}\nu^{\frac{1}{4}}}(\sigma_{1,i}^{(1)})^{\frac{1}{4}}(\sigma_{i,j}^{(2)})^{\frac{1}{4}} \\ =&\frac{\lambda_1^{\frac{2}{5}}(\sigma_{1,i}^{(1)})^{\frac{1}{5}}|w_k|^{\frac{1}{2}}}{6^{\frac{2}{5}}\nu^{\frac{2}{5}}(d_j^{(2)})^{\frac{1}{5}}}\left(\sum_{k'\in\mathcal{C}_j^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{1}{5}} \\ =&\frac{\lambda_1^{\frac{1}{2}}|w_k|^{\frac{1}{2}}}{6^{\frac{1}{2}}\nu^{\frac{1}{2}}(d_i^{(1)})^{\frac{1}{6}}(d_j^{(2)})^{\frac{1}{5}}}\left(\sum_{k'\in\mathcal{C}_j^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{1}{5}}\times \\ &\left(\sum_{j'\in\mathcal{C}_i^1}(d_{j'}^{(2)})^{\frac{1}{5}}\left(\sum_{k'\in\mathcal{C}_{j'}^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{1}{6}} \end{aligned}$$

and

$$\begin{aligned} &\sum_{i=1}^{s_1}\sum_{j\in\mathcal{C}_i^1}\sum_{k\in\mathcal{C}_j^2}\frac{w_k^2}{(\sigma_{1,i}^{(1)})^{\frac{1}{3}}(\sigma_{i,j}^{(2)})^{\frac{1}{3}}(\sigma_{j,k}^{(3)})^{\frac{1}{3}}} \\ =&\frac{6^{\frac{1}{2}}\nu^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}}}\sum_{i=1}^{s_1}\frac{(d_i^{(1)})^{\frac{1}{6}}}{\left(\sum_{j'\in\mathcal{C}_i^1}(d_{j'}^{(2)})^{\frac{1}{5}}\left(\sum_{k'\in\mathcal{C}_{j'}^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{1}{6}}}\times \\ &\sum_{j\in\mathcal{C}_i^1}\frac{(d_j^{(2)})^{\frac{1}{5}}}{\left(\sum_{k'\in\mathcal{C}_j^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{1}{5}}}\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}} \\ =&\frac{6^{\frac{1}{2}}\nu^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}}}\sum_{i=1}^{s_1}\frac{(d_i^{(1)})^{\frac{1}{6}}}{\left(\sum_{j'\in\mathcal{C}_i^1}(d_{j'}^{(2)})^{\frac{1}{5}}\left(\sum_{k'\in\mathcal{C}_{j'}^2}|w_{k'}|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{1}{6}}}\times \\ &\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}} \\ =&\frac{6^{\frac{1}{2}}\nu^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}}}\sum_{i=1}^{s_1}(d_i^{(1)})^{\frac{1}{6}}\left(\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{5}{6}} \\ =&\left(\sum_{i=1}^{s_1}(d_i^{(1)})^{\frac{1}{6}}\left(\sum_{j\in\mathcal{C}_i^1}(d_j^{(2)})^{\frac{1}{5}}\left(\sum_{k\in\mathcal{C}_j^2}|w_k|^{\frac{3}{2}}\right)^{\frac{4}{5}}\right)^{\frac{5}{6}}\right)^2. \end{aligned}$$

By plugging this equation into the objective function of problem (13), we reach the conclusion. □

The FISTA and GIST Algorithms

The detailed procedures for the FISTA and GIST algorithms are shown in Algorithms 1 and 2.

Algorithm 1 The FISTA Algorithm

Initialize r_0, $\alpha > 1$;
Set $\boldsymbol{\phi}^{(0)}$ as the initial value for variable $\boldsymbol{\phi}$;
$\boldsymbol{\psi}^{(1)} := \boldsymbol{\phi}^{(0)}$;
$k := 1$;
$t_1 := 1$;
while not converged **do**
  Find the smallest nonnegative integers i_k such that with $\hat{r} = \alpha^{i_k} r_{k-1}$, $F(q_{\hat{r}}(\boldsymbol{\psi}^{(k)})) \leq Q_{\hat{r}}(q_{\hat{r}}(\boldsymbol{\psi}^{(k)}), \boldsymbol{\psi}^{(k)})$;
  $r_k := \alpha^{i_k} r_{k-1}$;
  $\boldsymbol{\phi}^{(k)} := q_{r_k}(\boldsymbol{\psi}^{(k)})$;
  $t_{k+1} := \frac{1+\sqrt{1+4t_k^2}}{2}$;
  $\boldsymbol{\psi}^{(k+1)} := \boldsymbol{\phi}^{(k)} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\boldsymbol{\phi}^{(k)} - \boldsymbol{\phi}^{(k-1)})$
  $k := k + 1$;
end while

Algorithm 2 The GIST Algorithm

Choose $\eta > 1$, r_0, $\sigma \in (0, 1)$;
Set $\boldsymbol{\phi}^{(0)}$ as the initial value for variable $\boldsymbol{\phi}$;
$k := 0$;
repeat
  $r_{k+1} := r_k$;
  repeat
    $\boldsymbol{\phi}^{(k+1)} := \arg\min_{\boldsymbol{\phi}} H_{r_{k+1}}(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$;
    $r_{k+1} := \eta r_{k+1}$;
  until $F(\boldsymbol{\phi}^{(k+1)}) \leq F(\boldsymbol{\phi}^{(k)}) - \frac{r_{k+1}\sigma}{2}\|\boldsymbol{\phi}^{(k+1)} - \boldsymbol{\phi}^{(k)}\|_F^2$
  $k := k + 1$;
until Some convergence criterion is satisfied

Optimization Procedure for Problem (4)

Note that the only coupling in the variables of problem (4) comes from the equality constraint. The Lagrangian of problem (4) with respect to the equality constraint is given by

$$\mathcal{L}(\boldsymbol{\rho}, \gamma) = \|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|_2^2 - 2\gamma(\mathbf{a}^T\boldsymbol{\rho} - 1).$$

By setting the derivative of $\mathcal{L}$ with respect to each ρ_i to 0, we can see that the minimum is reached when $\rho_i = \hat{\rho}_i + a_i\gamma$ where a_i is the ith entry of $\mathbf{a}$. Since each ρ_i is required to be nonnegative and $\mathcal{L}(\boldsymbol{\rho}, \gamma)$ is a quadratic function of $\boldsymbol{\rho}$, the optimal solution for ρ_i is given by

$$\rho_i = \max\left(0, \hat{\rho}_i + a_i\gamma\right). \tag{26}$$

Plugging the optimal solution of ρ_i into $\mathcal{L}(\boldsymbol{\rho}, \gamma)$, we can obtain the dual problem as

$$\min_{\gamma} \sum_{\gamma \geq -\hat{\rho}_i/a_i} (a_i^2\gamma^2 + 2a_i\hat{\rho}_i\gamma) - \sum_{\gamma < -\hat{\rho}_i/a_i} \hat{\rho}_i^2 - 2\gamma. \tag{27}$$

Obviously, the objective function of problem (27) is a piecewise linear or quadratic function over regions determined by the sequence $\{-\frac{\hat{\rho}_i}{a_i}\}$. The main idea of our method is to determine the functional form of problem (27) over each region, then compute the local optimum over each region which has an analytical solution, and finally obtain the global optimum by comparing all the local optima. So the main problem is to determine the coefficients of problem (27) over each region efficiently. Without loss of the generality, we assume $\hat{\rho}_1/a_1 \geq \hat{\rho}_2/a_2 \geq \ldots \geq \hat{\rho}_s/a_s$ where s is the length of $\mathbf{a}$. If this is not the case, we can sort the sequence $\{\hat{\rho}_i/a_i\}$. When $\gamma \in [-\hat{\rho}_s/a_s, +\infty)$, the objective function of problem (27) is $c_2\lambda^2 + c_1\lambda + c_0$, where $c_2 = \sum_{i=1}^s a_i^2$, $c_1 = 2(\sum_{i=1}^s a_i\hat{\rho}_i - 1)$, and $c_0 = 0$, and it has an analytical solution as $\gamma = \max(-\frac{\hat{\rho}_s}{a_s}, -\frac{c_1}{2c_2})$. When $\gamma \in (-\infty, -\hat{\rho}_1/a_1)$, problem (27) has no well-defined solution since the objective function becomes $-2\gamma - \sum_{i=1}^s \hat{\rho}_i^2$. So we only need to consider the situation where $\gamma \in [-\hat{\rho}_1/a_1, -\hat{\rho}_s/a_s)$. We summarize the algorithm for solving problem (27) in Algorithm 3. This algorithm needs to scan the sequence $\{-\hat{\rho}_i/a_i\}$ only once which costs $O(s)$. So the complexity of the whole algorithm is at most $O(s \ln s)$, which is more efficient than existing QP solvers.

After determining the optimal γ, we can obtain the solution for ρ_i via Eq. (26).

Algorithm 3 Algorithm for problem (27)

Sort $\{\hat{\rho}_i/a_i\}$ if needed;
$c_0 := 0$; % coefficient for constant term
$c_1 := 2(\sum_{i=1}^s a_i\hat{\rho}_i - 1)$; % coefficient for linear term
$c_2 := \sum_{i=1}^s a_i^2$; % coefficient for quadratic term
$\gamma := \max(-\frac{\hat{\rho}_s}{a_s}, -\frac{c_1}{2c_2})$;
$f := c_0 + c_1\gamma + c_2\gamma^2$; % value of current minimum
for $i = s$ to 2 **do**
  % Determine the coefficients over $[-\hat{\rho}_{i-1}/a_{i-1}, -\hat{\rho}_i/a_i)$;
  $c_0 := c_0 - \hat{\rho}_i^2$;
  $c_1 := c_1 - 2a_i\hat{\rho}_i$;
  $c_2 := c_2 - a_i^2$;
  $\gamma_0 := \min(-\frac{\hat{\rho}_i}{a_i}, \max(-\frac{\hat{\rho}_{i-1}}{a_{i-1}}, -\frac{c_1}{2c_2}))$;
  $f_0 := c_0 + c_1\gamma_0 + c_2\gamma_0^2$;
  if $f_0 < f$ **then**
    $\gamma := \gamma_0$;
    $f := f_0$;
  end if
end for

Optimization Procedure for Problem (5)

Unfortunately, problem (5) is non-convex with respect to all variables. To see this, we present a case that $f(w, \mathbf{z}, \boldsymbol{\theta}) = \frac{w^2}{\prod_{i=1}^m z_i^{\theta_i}}$ is not jointly convex with respect to $w \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\theta}$, where w and $\{z_i\}$ are required to be positive, and $\boldsymbol{\theta}$ is required to satisfy $\theta_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m \theta_i = 1$. By considering a case that $d = 1$ and $m = 2$ and defining $\hat{w} = -48$, $\bar{w} = -74$, $\hat{\mathbf{z}} = (60\ 86)^T$, $\bar{\mathbf{z}} = (99\ 93)^T$, $\hat{\boldsymbol{\theta}} = (0.0006\ 0.9994)^T$, $\bar{\boldsymbol{\theta}} = (0.8286\ 0.1714)^T$,

and $\alpha = 0.4095$, we have

$$\begin{aligned}
&\alpha f(\hat{w}, \hat{\mathbf{z}}, \hat{\boldsymbol{\theta}}) + (1-\alpha) f(\bar{w}, \bar{\mathbf{z}}, \bar{\boldsymbol{\theta}}) \\
=&0.4095 \times 26.7968 + (1 - 0.4095) \times 55.9091 \\
<&46.3551 \\
=&f(\alpha\hat{w} + (1-\alpha)\bar{w}, \alpha\hat{\mathbf{z}} + (1-\alpha)\bar{\mathbf{z}}, \alpha\hat{\boldsymbol{\theta}} + (1-\alpha)\bar{\boldsymbol{\theta}}),
\end{aligned}$$

which implies that problem (5) is non-convex.

In order to solve problem (5), we use the GIST algorithm. With the abuse of notations, we use a vector $\boldsymbol{\phi}$ to denote the concatenation of $\mathbf{w}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\theta}$. We define the set of constraints on $\boldsymbol{\phi}$ as $\mathcal{S}_\phi = \{\boldsymbol{\phi} | \sum_{j=1}^{s_i} d_j^{(i)} \sigma^{(i)}_{\mathcal{F}_j^i, j} = 1 \ \forall i \in [m],\ \sigma^{(i)}_{\mathcal{F}_j^i, j} \geq 0 \ \forall i, j, \sum_{j=1}^{m} \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} = 1 \ \forall i \in [d],\ \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} \geq 0 \ \forall i, j.\}$. We define $f(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{x}_i, y_i, \mathbf{w}) + \frac{\lambda_1}{2}\sum_{i=1}^{d} \frac{w_i^2}{\prod_{j=1}^{m} \left(\sigma^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j}\right)^{\theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j}}}$ and $g(\boldsymbol{\phi}) = \begin{cases} \frac{\lambda_2}{2}\|\mathbf{w}\|_2^2 & \text{if } \phi \in \mathcal{S}_\phi \\ +\infty, & \text{otherwise} \end{cases}$ as an extended real-value function. Instead of directly minimizing the original composite objective function $F(\boldsymbol{\phi}) = f(\boldsymbol{\phi}) + g(\boldsymbol{\phi})$, the GIST algorithm shown in Algorithm 2 of the supplementary material minimizes a surrogate function: $\min_{\boldsymbol{\phi} \in \mathcal{S}_\phi} H_r(\boldsymbol{\phi}, \hat{\boldsymbol{\phi}}) = g(\boldsymbol{\phi}) + f(\hat{\boldsymbol{\phi}}) + (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^T \bigtriangledown_{\boldsymbol{\phi}} f(\hat{\boldsymbol{\phi}}) + \frac{r}{2}\|\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}\|_2^2$, which can be simplified as

$$\begin{aligned}
\min_{\mathbf{w}, \boldsymbol{\sigma}, \boldsymbol{\theta}} \ & \frac{\lambda_2}{2}\|\mathbf{w}\|_2^2 + \frac{r}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + \frac{r}{2}\|\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}\|_2^2 + \frac{r}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \\
\text{s.t.} \ & \sum_{j=1}^{s_i} d_j^{(i)} \sigma^{(i)}_{\mathcal{F}_j^i, j} = 1 \ \forall i \in [m],\ \sigma^{(i)}_{\mathcal{F}_j^i, j} \geq 0 \ \forall i, j \\
& \sum_{j=1}^{m} \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} = 1 \ \forall i \in [d],\ \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} \geq 0 \ \forall i, j, \qquad (28)
\end{aligned}$$

where r is a step size determined by the GIST algorithm, $\tilde{\mathbf{w}} = \hat{\mathbf{w}} - \frac{1}{r} \bigtriangledown_{\mathbf{w}} f(\hat{\mathbf{w}})$, $\tilde{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}} - \frac{1}{r} \bigtriangledown_{\boldsymbol{\sigma}} f(\hat{\boldsymbol{\sigma}})$, and $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \frac{1}{r} \bigtriangledown_{\boldsymbol{\theta}} f(\hat{\boldsymbol{\theta}})$. The solution of $\mathbf{w}$ in problem (28) is $\mathbf{w} = \frac{r}{\lambda_2 + r}\tilde{\mathbf{w}}$. It is easy to see that $\boldsymbol{\sigma}$ and $\boldsymbol{\theta}$ in problem (28) are independent and the subproblem with respect to $\boldsymbol{\sigma}$ can be decomposed into multiple problems each of which has the same formulation as problem (4), leading to an efficient solution. The subproblem with respect to $\boldsymbol{\theta}$ based on problem (28) is formulated as

$$\begin{aligned}
\min_{\boldsymbol{\theta}} \ & \frac{1}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \\
\text{s.t.} \ & \sum_{j=1}^{m} \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} = 1 \ \forall i \in [d],\ \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} \geq 0 \ \forall i, j. \qquad (29)
\end{aligned}$$

Problem (29) is a QP problem and here we devise an optimization method for problem (29) to achieve the speedup. Since different features share the same $\theta^{(j)}_{k,r}$ when they have the same ancestor in the tree, problem (29) cannot directly reduce to problem (4). By introducing $\vartheta_i^{(j)}$ as a copy of $\theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j}$, we can reformulate problem (29) as

$$\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\vartheta}} \ & \frac{1}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \\
\text{s.t.} \ & \sum_{j=1}^{m} \vartheta_i^{(j)} = 1 \ \forall i \in [d],\ \vartheta_i^{(j)} \geq 0 \ \forall i, j,\ \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} = \vartheta_i^{(j)} \ \forall i, j, \\
& \qquad (30)
\end{aligned}$$

where $\boldsymbol{\vartheta}$ is a vector containing all $\vartheta_i^{(j)}$'s. Because of the linear equality constraints in problem (30), we use the ADMM algorithm (Boyd et al. 2011) to solve it. The augmented Lagrangian function is defined as $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \frac{1}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + \sum_{i=1}^{d}\sum_{j=1}^{m}\left(p_i^{(j)}(\theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} - \vartheta_i^{(j)}) + \frac{\rho}{2}\left(\theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} - \vartheta_i^{(j)}\right)^2\right)$, where $\{p_i^{(j)}\}_{i=1}^m$ act as Lagrangian multipliers and ρ is a penalty parameter. Then we need to solve

$$\min_{\boldsymbol{\theta}, \boldsymbol{\vartheta}} \ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \quad \text{s.t.} \ \sum_{j=1}^{m} \vartheta_i^{(j)} = 1 \ \forall i \in [d],\ \vartheta_i^{(j)} \geq 0 \ \forall i, j. \quad (31)$$

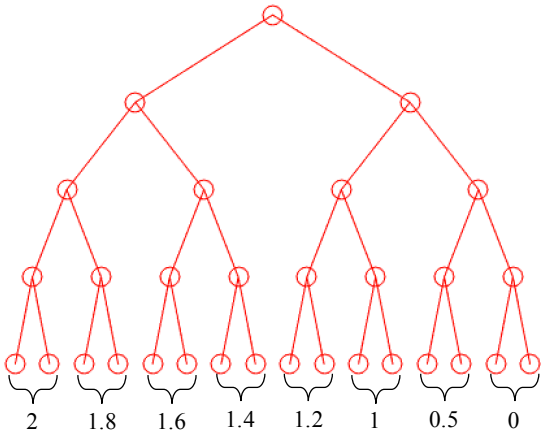
In the ADMM algorithm, problem (31) can be solved alternatively with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$. When $\boldsymbol{\theta}$ is fixed, the problem with respect to $\boldsymbol{\vartheta}_i = (\vartheta_i^{(1)}, \ldots, \vartheta_i^{(m)})^T$ is formulated as

$$\min_{\boldsymbol{\vartheta}_i} \sum_{j=1}^{m}\left(\vartheta_i^{(j)} - \tilde{\vartheta}_i^{(j)}\right)^2 \quad \text{s.t.} \ \sum_{j=1}^{m} \vartheta_i^{(j)} = 1 \ \forall i \in [d],\ \vartheta_i^{(j)} \geq 0 \ \forall i, j,$$

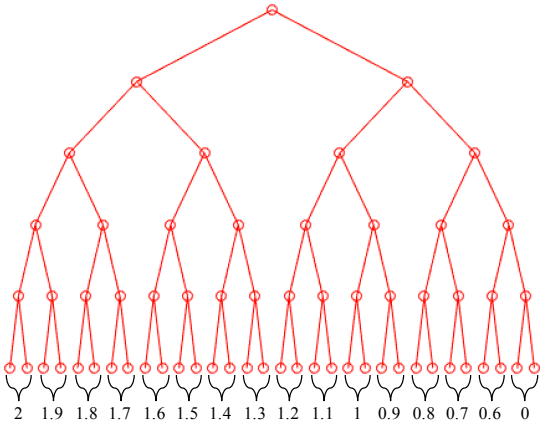
where $\tilde{\vartheta}_i^{(j)} = \theta^{(j)}_{\mathcal{P}_i^{j-1}, \mathcal{P}_i^j} + \frac{p_i^{(j)}}{\rho}$. This problem takes the same form as problem (4), which makes it have an efficient solution. The subproblem with respect to $\boldsymbol{\theta}$ is a QP problem without any constraint and by setting the derivative to zero, we can obtain the analytical solution as $\theta^{(j)}_{k,r} = \frac{1}{1 + \rho d_r^{(j)}}\left(\tilde{\theta}^{(j)}_{k,r} + \sum_{i \in \mathcal{S}^{(j)}_{k,r}}\left(\rho\vartheta_i^{(j)} - p_i^{(j)}\right)\right)$, where $\mathcal{S}^{(j)}_{k,r} = \{i | \mathcal{P}_i^{j-1} = k, \mathcal{P}_i^j = r\}$ with its cardinality equal to $d_r^{(j)}$.

Details in Experiments

The true feature weights for $\mathbf{w}^*$ when m equals 4 and 5 are shown in Fig. 3.



(a) $m = 4$



(b) $m = 5$

Figure 3: Binary trees generated in synthetic data when m equals 4 and 5.