A Algorithms

The pseudo-code of the greedy algorithm for solving Equation (6) in Wasserstein DRO is illustrated in Algorithm 1.

Algorithm 1 Greedy Algorithm for the Wasserstein Worst-case Risk

Input: $\boldsymbol{W}$, γ, $\boldsymbol{x}^{(i)}$
Output: a solution $\hat{\boldsymbol{x}}$ to Equation (6)
Initialize $\hat{\boldsymbol{x}} = \boldsymbol{x}^{(i)}$
for all $(j, x_j^t) \in [n] \times \mathcal{C}_j$ **do**
  Get a random permutation $\boldsymbol{\pi}$ over $[n]$ with $\pi_1 = j$
  for $k := 2$ **to** n **do**
    $x_{\pi_j}^t \leftarrow \arg\sup_{x_{\pi_k}^t} \ell_{\boldsymbol{W}}(\boldsymbol{x}_{\boldsymbol{\pi}_{[k]}}^t) - \gamma\|\mathcal{E}(\boldsymbol{x}_{\boldsymbol{\pi}_{[k]}}^t) - \mathcal{E}(\boldsymbol{x}_{\boldsymbol{\pi}_{[k]}}^{(i)})\|$
  end for
  if $\boldsymbol{x}^t$ yields a greater objective than $\hat{\boldsymbol{x}}$ **then**
    $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x}^t$
  end if
end for

B Optimization Details

Define

$$\ell_{\boldsymbol{W}}(\boldsymbol{X}) := \frac{1}{2}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal \mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2.$$

The Lagrangian dual problem of the Wasserstein DRO problem is

$$\inf_{\boldsymbol{W}, \gamma \geqslant 0} f(\boldsymbol{W}, \gamma) := \gamma\varepsilon + \frac{1}{m}\sum_{i=1}^{m} \sup_{\boldsymbol{x} \in \mathcal{X}} \ell_{\boldsymbol{W}}(\boldsymbol{x}) - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1.$$

One of its sub-gradients can be computed as

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{E}(\hat{\boldsymbol{x}}_{\bar{r}}^{(i)})\mathcal{E}(\hat{\boldsymbol{x}}_{\bar{r}}^{(i)})^\intercal \boldsymbol{W} - \mathcal{E}(\hat{\boldsymbol{x}}_{\bar{r}}^{(i)})\mathcal{E}(\hat{\boldsymbol{x}}_r^{(i)})^\intercal \in \frac{\partial}{\partial \boldsymbol{W}} f$$

$$\varepsilon - \frac{1}{m}\sum_{i=1}^{m} \|\mathcal{E}(\hat{\boldsymbol{x}}^{(i)}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1 \in \frac{\partial}{\partial \gamma} f.$$

For the DRO problem based on the KL divergence:

$$\inf_{\boldsymbol{W}, \gamma > 0} f(\boldsymbol{W}, \gamma) := \gamma \ln\big[\frac{1}{m}\sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma}\big] + \gamma\varepsilon,$$

a sub-gradient of which can be computed as

$$\frac{\sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma}(\mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)})\mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)})^\intercal \boldsymbol{W} - \mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)})\mathcal{E}(\boldsymbol{x}_r^{(i)})^\intercal)}{\sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma}} \in \frac{\partial}{\partial \boldsymbol{W}} f$$

$$\ln\big(\frac{1}{m}\sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma}\big) - \frac{\sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma} \cdot \ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})}{\gamma \sum_{i \in [m]} e^{\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})/\gamma}} + \epsilon \in \frac{\partial}{\partial \gamma} f.$$

C Technical Proofs

Proposition 11 (NP-hardness of Wasserstein DRO Supremum)**.** *The problem in Equation* (6) *is NP-hard.*

Proof. Recall the MAXQP problem:

$$\sum_{i,j=1}^{n} a_{ij}x_ix_j, \quad \text{s.t. } x_i \in \{-1,1\}\ \forall i.$$

In Equation (6), let $\gamma = 0$, $\mathcal{E}(x_r) = 0$, $\boldsymbol{x}_{\bar{r}}$ correspond to n binary variables taking values in {-1, 1} and $\mathcal{E}(\boldsymbol{x}_{\bar{r}}) = \boldsymbol{x}_{\bar{r}}$. Let $\boldsymbol{W} \in \mathbb{R}^{n\times n^2}$. For all $i, j \in [n]$, let $k = (i-1)n + j$. The k-the column of $\boldsymbol{W}$ satisfies $W_{ik} = 1$, $W_{jk} = a_{ij}/2$ and 0 for the other elements. We have obtained a polynomial-time reduction from an NP-hard problem to Equation (6). □

Proposition 5 (Regularization Equivalence)**.** *Let $\ddot{\boldsymbol{W}} := [\boldsymbol{W}; -\boldsymbol{I}_{\rho_r}]^\intercal \in \mathbb{R}^{\rho_{[n]}\times\rho_r}$ with $\boldsymbol{W}_r = -\boldsymbol{I}_{\rho_r}$. If $\gamma \geqslant \rho_{[n]}\|\ddot{\boldsymbol{W}}\|_F^2$, the Wasserstein distributionally robust regression problem in Equation* (5) *is equivalent to*

$$\inf_{\boldsymbol{W}} \mathbb{E}_{\tilde{\mathbb{P}}_m} \frac{1}{2}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2 + \varepsilon\rho_{[n]}\|\ddot{\boldsymbol{W}}\|_F^2,$$

which subsumes a linear regression approach regularized by the Frobenius norm as a special case.

Proof. Recapitulating on Equation (6):

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{2}\|\mathcal{E}(x_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}})\|_2^2 - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1.$$

Observe that

$$\begin{aligned}\|\mathcal{E}(x_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}})\|_2^2 \triangleq& \|\ddot{\boldsymbol{W}}^\intercal\mathcal{E}(\boldsymbol{x}_{[n]})\|_2^2\\ \leqslant& \|\ddot{\boldsymbol{W}}^\intercal\|_{\infty,2}^2\\ \leqslant& \|\ddot{\boldsymbol{W}}\|_{1,2}^2\\ \leqslant& \rho_{[n]}\|\ddot{\boldsymbol{W}}\|_F^2\\ \leqslant& \gamma.\end{aligned}$$

Therefore, for any $\boldsymbol{x} \neq \boldsymbol{x}^{(i)}$,

$$\begin{aligned}&\frac{1}{2}\|\mathcal{E}(x_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}})\|_2^2 - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1 - (\frac{1}{2}\|\mathcal{E}(x_r^{(i)}) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)})\|_2^2 - \gamma\|\mathcal{E}(\boldsymbol{x}^{(i)}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1)\\ \leqslant&\frac{1}{2}(\|\mathcal{E}(x_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}})\|_2^2 - \|\mathcal{E}(x_r^{(i)}) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)})\|_2^2) - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1\\ \leqslant&\frac{1}{2}(2\gamma) - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|_1\\ \leqslant&\gamma - \gamma\\ =&0,\end{aligned}$$

which implies that the supremum can always be achieved at $\boldsymbol{x} = \boldsymbol{x}^{(i)}$. Minimizing over γ leads to

$$\inf_{\boldsymbol{W}} \mathbb{E}_{\tilde{\mathbb{P}}_m} \frac{1}{2}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2 + \varepsilon\rho_{[n]}\|\ddot{\boldsymbol{W}}\|_F^2.$$

□

Lemma 6. *Suppose Ξ is separable Banach space and fix $\mathbb{P}_0 \in \mathcal{P}(\Xi')$ for some $\Xi' \subseteq \Xi$. Suppose $c : \Xi \to \mathbb{R}_{\geqslant 0}$ is closed convex, k-positively homogeneous. Suppose $f : \Xi \to \mathcal{Y}$ is a mapping in the Lebesgue space of functions with finite first-order moment under $\mathbb{P}_0$ and upper semi-continuous with finite Lipschitz constant $lip_c(f)$. Then for all $\varepsilon \geqslant 0$, the following inequality holds with probability* 1*:*

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0),\mathbb{Q}\in\mathcal{P}(\Xi')} \int f(\xi')\mathbb{Q}(\mathrm{d}\xi') \leqslant \varepsilon lip_c(f) + \int f(\xi')\mathbb{P}_0(\mathrm{d}\xi').$$

Proof. The result follows directly from Theorem 1 in Cranko et al. [2021]:

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0),\mathbb{Q}\in\mathcal{P}(\Xi)}\int f(\xi)\mathbb{Q}(\mathrm{d}\xi)\leqslant\varepsilon\mathrm{lip}_c(f)+\int f(\xi')\mathbb{P}_0(\mathrm{d}\xi').$$

Since $\Xi'\subseteq\Xi$, observe

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0),\mathbb{Q}\in\mathcal{P}(\Xi')}\int f(\xi')\mathbb{Q}(\mathrm{d}\xi')\leqslant\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0),\mathbb{Q}\in\mathcal{P}(\Xi)}\int f(\xi)\mathbb{Q}(\mathrm{d}\xi).$$

□

Lemma 7. *If Assumption 3 holds, for any $\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$, with probability at least $1-2|\mathcal{S}_r|^2\exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$, we have*

$$\Lambda_{min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\geqslant\Lambda_{min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})-4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}}-t.$$

Proof. The minimum eigenvalue of the true covariance matrix $\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}$ satisfies:

$$\begin{aligned}\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})&\triangleq\min_{\|\boldsymbol{v}\|_2=1}\boldsymbol{v}^\intercal\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}\boldsymbol{v}\\&=\min_{\|\boldsymbol{v}\|_2=1}\boldsymbol{v}^\intercal\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}\boldsymbol{v}+\boldsymbol{v}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{v}+\boldsymbol{v}^\intercal(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{v}\\&\leqslant\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})+\boldsymbol{u}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}+\boldsymbol{u}^\intercal(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u},\end{aligned}$$

where $\|\boldsymbol{u}\|_2=1$ is an eigenvector of $\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}$ with minimum eigenvalue.

Therefore, $\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})$ can be lower bounded as follows:

$$\begin{aligned}\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})&\geqslant\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})-\boldsymbol{u}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}-\boldsymbol{u}^\intercal(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}\\&\geqslant\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})-|\boldsymbol{u}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}|-\|(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})\|_F,\end{aligned}$$

due to the fact that

$$\boldsymbol{u}^\intercal\boldsymbol{H}\boldsymbol{u}\leqslant\Lambda_{\max}(\boldsymbol{H})\leqslant\sqrt{\sum_i(\Lambda_i(\boldsymbol{H}))^2}\leqslant\|\boldsymbol{H}\|_{2,2}.$$

We can obtain an upper bound on $|\boldsymbol{u}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}|$ based on Lemma 6:

$$|\boldsymbol{u}^\intercal(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\boldsymbol{u}|\leqslant4|\mathcal{S}_r|^{\frac{1}{2}}\varepsilon,$$

because for function $g(\mathcal{E}(\boldsymbol{x}))\coloneqq\boldsymbol{u}^\intercal\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}\boldsymbol{u}$, it can be shown that for any $\|\mathcal{E}(\boldsymbol{x})-\mathcal{E}(\boldsymbol{x}')\|_1=k$ and some $|\mathcal{S}|=k$,

$$|g(\mathcal{E}(\boldsymbol{x}))-g(\mathcal{E}(\boldsymbol{x}'))|\leqslant\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S}_r}|H_{ik}-H'_{ik}|u_iu_k+|H_{ki}-H'_{ki}|u_ku_i\leqslant4k|\mathcal{S}_r|^{\frac{1}{2}}.$$

Recall that we assume that the encoding schemes take values in $\mathcal{B}=\{-1,0,1\}$. Therefore $\mathrm{lip}_c(g)=4|\mathcal{S}_r|^{\frac{1}{2}}$.

We derive an upper bound of $\|(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})\|_F$ as follows. Consider a random variable and its expectation

$$\begin{aligned}&Z_{ij}\coloneqq(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})_{ij}=\frac{1}{m}\sum_{l=1}^m\mathcal{E}(\boldsymbol{x}^{(l)}_{\tilde{r}})_i\mathcal{E}(\boldsymbol{x}^{(l)}_{\tilde{r}})_j\in[-1/m,1/m]\\&\mathbb{E}_{\mathbb{P}}Z_{ij}=(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})_{ij}.\end{aligned}$$

By Hoeffding's inequality, we observe

$$\mathrm{Prob}(|(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r})_{ij}-(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})_{ij}|\geqslant t)\leqslant2\exp(-\frac{mt^2}{2}),$$

for $t > 0$. Setting $t = \frac{t}{|\mathcal{S}_r|}$ for all $i, j \in \mathcal{S}_r$ and applying the union bound,

$$\text{Prob}(\|(\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}) - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})\|_F \geqslant t) \leqslant 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right). \tag{8}$$

To conclude, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right)$, we have

$$\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}) \geqslant \Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}) - 4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}} - t.$$

□

Lemma 8. *If Assumption 3 and Assumption 4 hold, for any $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$ and $\alpha \in (0, 1]$, with probability at least $1 - \mathcal{O}(\exp\left(-\frac{Cm}{\rho_{max}^2|\mathcal{S}_r|^3} + \log|\mathcal{S}_r^c| + \log|\mathcal{S}_r|\right))$ and $\varepsilon \leqslant \frac{C}{\rho_{max}|\mathcal{S}_r|^{3/2}}$,*

$$\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant 1 - \frac{\alpha}{2},$$

where C only depends on α, $\Lambda_{min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})$.

Proof. We would like to obtain an upper bound for $\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}$. We may write

$$\begin{aligned}
\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} =& \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}] \\
&+ [\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1} \\
&+ [\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}] \\
&+ \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1} \\
\Longrightarrow& \\
\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant& \|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\
&+ \|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \\
&+ \|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\
&+ \|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}.
\end{aligned}$$

By Hoeffding's inequality,

$$\text{Prob}(|(\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r})_{ij} - (\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r})_{ij}| \geqslant t) \leqslant 2\exp\left(-\frac{mt^2}{2}\right),$$

for $t > 0$. Taking $t = \frac{t}{\rho_i|\mathcal{S}_r|}$ and applying the union bound over $i \in \mathbf{Co}_r$, we observe that

$$\begin{aligned}
\text{Prob}(\|\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} \geqslant t) \leqslant& \sum_{i\in\mathbf{Co}_r} 2\rho_i|\mathcal{S}_r|\exp\left(-\frac{mt^2}{2\rho_i^2|\mathcal{S}_r|^2}\right) \\
\leqslant& 2|\mathcal{S}_r^c||\mathcal{S}_r|\exp\left(-\frac{mt^2}{2\rho_{\max}^2|\mathcal{S}_r|^2}\right).
\end{aligned}$$

Similarly, taking $t = \frac{t}{|\mathcal{S}_r|}$,

$$\begin{aligned}
\text{Prob}(\|\!|\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}|\!\|_{\infty,\infty} \geqslant t) \leqslant& \sum_{i\in\mathcal{S}_r}\sum_{j\in\mathcal{S}_r} 2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right) \\
=& 2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right).
\end{aligned}$$

In order to bound $\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}$, for $\mathbb{Q} \neq \tilde{\mathbb{P}}$, consider

$$\begin{aligned}
\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} \leqslant& \|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} + \|\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} \\
\leqslant& \mathbb{E}_{\mathbb{Q}}\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty} + \mathbb{E}_{\tilde{\mathbb{P}}_m}\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty} \\
=& \sup_{\tilde{\mathbb{P}}'_m, \mathbb{Q}'\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)} |\mathbb{E}_{\mathbb{Q}'}\xi_1\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty} - \mathbb{E}_{\tilde{\mathbb{P}}'_m}\xi_2\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty}|,
\end{aligned}$$

where $\mathbb{Q}'$ and $\tilde{\mathbb{P}}'_m$ are probability measures on $\mathcal{X}\times\Xi$ with $\Xi=\{-1,+1\}$ and identical marginals as $\mathbb{Q}$ and $\tilde{\mathbb{P}}_m$ respectively. We assume that $\mathbb{Q}\neq\tilde{\mathbb{P}}$ because otherwise $\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}=0$ holds trivially. In this way, the equality is always achieved by some $\mathbb{Q}',\tilde{\mathbb{P}}'_m$, i.e., setting $\mathbb{Q}'(\mathcal{X},\xi=1)=1$ and $\tilde{\mathbb{P}}'_m(\mathcal{X},\xi=-1)=1$.

Define the transport cost function in the ambiguity set $\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)$ to be $c'((\boldsymbol{X}_1,\xi_1),(\boldsymbol{X}_2,\xi_2)):=\|\mathcal{E}(\boldsymbol{X}_1)-\mathcal{E}(\boldsymbol{X}_2)\|_1$ with zero cost for ξ. Let $g(\boldsymbol{X},\xi):=\xi_1\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty}$. Consider the Lipschitz constants of g:

$$\begin{aligned}\mathrm{lip}_{c'}(g)\leqslant&\sup_{\boldsymbol{X},\xi,\boldsymbol{X}',\xi'}\frac{|g(\boldsymbol{X},\xi)-g(\boldsymbol{X}',\xi')|}{c'((\boldsymbol{X},\xi),(\boldsymbol{X}',\xi'))}\\ \leqslant&\sup_{\boldsymbol{X},\boldsymbol{X}'}\frac{\|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty}+\|\mathcal{E}(\boldsymbol{X}'_{\bar{r}})_{\mathcal{S}_r^c}\mathcal{E}(\boldsymbol{X}'_{\bar{r}})^\intercal_{\mathcal{S}_r}\|_{B,1,\infty}}{\|\mathcal{E}(\boldsymbol{X})-\mathcal{E}(\boldsymbol{X}')\|_1}\\ \leqslant&2\rho_{\max}|\mathcal{S}_r|.\end{aligned}\tag{9}$$

Therefore, by the Kantorovich-Rubinstein theorem [Kantorovich and Rubinshtein, 1958],

$$\begin{aligned}\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}\leqslant&\sup_{\tilde{\mathbb{P}}'_m,\mathbb{Q}'\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)}|\mathbb{E}_{\mathbb{Q}'}g(\boldsymbol{X},\xi)-\mathbb{E}_{\tilde{\mathbb{P}}'_m}g(\boldsymbol{X},\xi)|\\ \leqslant&\sup_{\tilde{\mathbb{P}}'_m,\mathbb{Q}'\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)}\mathrm{lip}_{c'}(g)|\mathbb{E}_{\mathbb{Q}'}g(\boldsymbol{X},\xi)/\mathrm{lip}_{c'}(g)-\mathbb{E}_{\tilde{\mathbb{P}}'_m}g(\boldsymbol{X},\xi)/\mathrm{lip}_{c'}(g)|\\ \leqslant&\sup_{\tilde{\mathbb{P}}'_m,\mathbb{Q}'\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)}\mathrm{lip}_{c'}(g)W_1(\mathbb{Q}',\tilde{\mathbb{P}}'_m)\\ \leqslant&\mathrm{lip}_{c'}(g)\varepsilon\\ \leqslant&2\varepsilon\rho_{\max}|\mathcal{S}_r|.\end{aligned}$$

Similarly,

$$|||\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r\mathcal{S}_r}|||_{\infty,\infty}\leqslant 2\varepsilon|\mathcal{S}_r|.$$

Based on the above two inequalities, we find that

$$\begin{aligned}\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}-\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}\leqslant&\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}-\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}+\|\tilde{\boldsymbol{H}}_{\mathcal{S}_r^c\mathcal{S}_r}-\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}\\ \leqslant&2\varepsilon\rho_{\max}|\mathcal{S}_r|+t,\end{aligned}\tag{10}$$

with probability at least $1-2|\mathcal{S}_r^c||\mathcal{S}_r|\exp\left(-\frac{mt^2}{2\rho^2_{\max}|\mathcal{S}_r|^2}\right)$, and

$$|||\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}|||_{\infty,\infty}\leqslant 2\varepsilon|\mathcal{S}_r|+t,\tag{11}$$

with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right)$.

Based on Equation (8), we also have

$$\|[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}]\|_F\leqslant 2\varepsilon|\mathcal{S}_r|+t,\tag{12}$$

with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right)$.

Next we look at the upper bound on the difference between the inverses of $\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}$ and $\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}$. Observe that

$$\begin{aligned}|||(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}-(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{\infty,\infty}=&|||(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}](\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{\infty,\infty}\\ \leqslant&\sqrt{|\mathcal{S}_r|}|||(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}](\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{2,2}\\ \leqslant&\sqrt{|\mathcal{S}_r|}|||(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{2,2}|||[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}]|||_{2,2}|||(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{2,2}\\ \leqslant&\sqrt{\frac{|\mathcal{S}_r|}{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}}|||[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}]|||_{2,2}|||(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|||_{2,2}.\end{aligned}$$

According to Lemma 7, with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right)$, we have

$$\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}) \geqslant \Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}) - 4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}} - t.$$

Let $t = \frac{1}{2}\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})$ and $\varepsilon \leqslant \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}$. We get that, with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right)$,

$$\begin{aligned}
&\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}) \geqslant \frac{1}{4}\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}) \\
\Longrightarrow\ &|\!|\!|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{2,2} \leqslant \sqrt{\frac{4}{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}}. \qquad (13)
\end{aligned}$$

Set $t = \frac{t\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{4\sqrt{|\mathcal{S}_r|}}$ and $\varepsilon \leqslant \frac{t\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}}$ in Equation (12), we get that, with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3}\right)$,

$$|\!|\!|[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}]|\!|\!|_{2,2} \leqslant \|[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}]\|_F \leqslant \frac{t\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{2\sqrt{|\mathcal{S}_r|}}.$$

Therefore, with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3}\right) - 2|\mathcal{S}_r|^2\exp\left(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right)$ and $\varepsilon \leqslant \min\left(\frac{t\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right)$,

$$|\!|\!|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{\infty,\infty} \leqslant t. \qquad (14)$$

Now we are ready to obtain upper bounds for the four terms recapitulated here:

$$\begin{aligned}
\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant &\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\
&+ \|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \\
&+ \|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\
&+ \|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}.
\end{aligned}$$

We derive the bounds separately.

For the first term, based on Assumption 4, consider

$$\begin{aligned}
&\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\
=&\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}[\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}](\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \\
\leqslant&\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}|\!|\!|\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}|\!|\!|_{\infty,\infty}|\!|\!|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{\infty,\infty} \\
\leqslant&(1-\alpha)|\!|\!|\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} - \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}|\!|\!|_{\infty,\infty}\sqrt{|\mathcal{S}_r|}|\!|\!|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{2,2}.
\end{aligned}$$

Taking $t = \frac{\alpha}{24(1-\alpha)}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}$ and $\varepsilon \leqslant \frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}$ in Equation (11) and adopting Equation (13), we conclude that, with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{m\alpha^2\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{1152(1-\alpha)^2|\mathcal{S}_r|^3}\right) - 2|\mathcal{S}_r|^2\exp\left(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right)$ and $\varepsilon \leqslant \min\left(\frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right)$,

$$\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \leqslant \frac{\alpha}{6}.$$

For the second term, rewrite it as

$$\begin{aligned}
&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \\
\leqslant&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}]\|_{B,1,\infty}|\!|\!|(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{\infty,\infty} \\
\leqslant&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}]\|_{B,1,\infty}\sqrt{|\mathcal{S}_r|}|\!|\!|(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{2,2} \\
\leqslant&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}]\|_{B,1,\infty}\sqrt{\frac{|\mathcal{S}_r|}{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}}.
\end{aligned}$$

Using Equation (10) by setting $t = \frac{\alpha}{12}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}$ and $\varepsilon \leqslant \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}$, we have, with probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r|\exp(-\frac{m\alpha^2\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{288\rho_{\max}^2|\mathcal{S}_r|^3})$ and $\varepsilon \leqslant \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}$,

$$\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant \frac{\alpha}{6}.$$

For the third term, we obtain the upper bound

$$\begin{aligned}&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty}\\ \leqslant&\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}]\|_{B,1,\infty}\|\|[(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|\|_{\infty,\infty}.\end{aligned}$$

Taking $t = \sqrt{\frac{\alpha}{6}}$ in Equation (14). Taking $t = \frac{1}{2}\sqrt{\frac{\alpha}{6}}$ and $2\varepsilon\rho_{\max}|\mathcal{S}_r| \leqslant \frac{1}{2}\sqrt{\frac{\alpha}{6}}$ in Equation (10). We establish the upper bound that, with probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r|\exp(-\frac{m\alpha}{48\rho_{\max}^2|\mathcal{S}_r|^2}) - 2|\mathcal{S}_r|^2\exp(-\frac{m\alpha(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{192|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2\exp(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and $\varepsilon \leqslant \min(\frac{1}{4\rho_{\max}|\mathcal{S}_r|}\sqrt{\frac{\alpha}{6}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|}\sqrt{\frac{\alpha}{6|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}})$,

$$\|[\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r} - \boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1} - (\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \leqslant \frac{\alpha}{6}.$$

For the fourth term, in accordance with Assumption 4,

$$\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant 1 - \alpha.$$

In conclusion, we have shown that, with probability at least $1 - 2|\mathcal{S}_r|^2\exp(-\frac{m\alpha^2\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{1152(1-\alpha)^2|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2\exp(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}) - 2|\mathcal{S}_r^c||\mathcal{S}_r|\exp(-\frac{m\alpha^2\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{288\rho_{\max}^2|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r^c||\mathcal{S}_r|\exp(-\frac{m\alpha}{48\rho_{\max}^2|\mathcal{S}_r|^2}) - 2|\mathcal{S}_r|^2\exp(-\frac{m\alpha(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{192|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2\exp(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and

$$\begin{aligned}\varepsilon \leqslant \min(&\frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}, \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{|\mathcal{S}_r|}},\\ &\frac{1}{4\rho_{\max}|\mathcal{S}_r|}\sqrt{\frac{\alpha}{6}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|}\sqrt{\frac{\alpha}{6|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}),\end{aligned}$$

the mutual incoherence condition holds for any worst-case distributions:

$$\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant 1 - \frac{\alpha}{2}.$$

Simplifying the above expressions, with probability at least $1 - \mathcal{O}(\exp(-\frac{Cm}{\rho_{\max}^2|\mathcal{S}_r|^3} + \log|\mathcal{S}_r^c| + \log|\mathcal{S}_r|))$ and $\varepsilon \leqslant \frac{C}{\rho_{\max}|\mathcal{S}_r|^{3/2}}$,

$$\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \leqslant 1 - \frac{\alpha}{2},$$

where C only depends on α, $\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})$. □

Lemma 12. *If Assumption 1 holds, then for any $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$ and $\alpha \in (0,1]$, with probability at least $1 - |\mathcal{S}_r|\rho_r\exp(-\frac{m\mu^2}{2\sigma^2})$, $\varepsilon \leqslant \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$, we have*

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} \leqslant \frac{\lambda_B^*\alpha}{8(1-\alpha/2)}.$$

With probability at least $1 - |\boldsymbol{Co}_r|\rho_r\exp(-\frac{m\mu^2}{2\sigma^2})$, $\varepsilon \leqslant \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_{max}\rho_r}}{\alpha}$, we have

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\|_{B,2,\infty} \leqslant \frac{\lambda_B^*\alpha}{8}.$$

Proof. We start with $\|\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^{\mathsf{T}}\|\|_{2,\infty}$. After some algeraic manipulation, we find that

$$\begin{aligned}\|\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^{\mathsf{T}}\|\|_{2,\infty} &\leqslant \max_{i\in\mathcal{S}_r}\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_i\boldsymbol{e}\|_2 \\ &\leqslant \max_{i\in\mathcal{S}_r}\sqrt{\rho_r}\max_{j\in\rho_r}|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_i e_j| \\ &\leqslant \max_{i\in\mathcal{S}_r}\sqrt{\rho_r}\max_{j\in\rho_r}\mathbb{E}_{\mathbb{Q}}|\mathcal{E}(\boldsymbol{X}_{\bar{r}})_i e_j| \\ &\leqslant \max_{i\in\mathcal{S}_r}\sqrt{\rho_r}\max_{j\in\rho_r}\mathbb{E}_{\mathbb{Q}}|e_j|.\end{aligned}$$

Since $|e_j|$ is a bounded random variable according to Assumption 1, we apply Hoeffding's inequality to get

$$\text{Prob}(\mathbb{E}_{\tilde{\mathbb{P}}_m}|e_j| \geqslant \mu + t) \leqslant \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Base on a similar argument as Equation (9), we can derive

$$\mathbb{E}_{\mathbb{Q}}|e_j| - \mathbb{E}_{\tilde{\mathbb{P}}_m}|e_j| \leqslant 2\varepsilon\sigma,$$

which leads to

$$\text{Prob}(\mathbb{E}_{\mathbb{Q}}|e_j| \geqslant 2\varepsilon\sigma + \mu + t) \leqslant \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Taking the union bound over all $i \in \mathcal{S}_r$ and $j \in \rho_r$, we find that

$$\text{Prob}(\|\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^{\mathsf{T}}\|\|_{2,\infty} \geqslant \sqrt{\rho_r}(2\varepsilon\sigma + \mu + t)) \leqslant |\mathcal{S}_r|\rho_r \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Setting $t = \mu$ and $\varepsilon \leqslant \frac{\mu}{\sigma}$ while requiring $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$. With probability at least $1 - |\mathcal{S}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$, we have

$$\|\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^{\mathsf{T}}\|\|_{2,\infty} \leqslant \frac{\lambda_B^*\alpha}{8(1-\alpha/2)}. \tag{15}$$

Then we consider $\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^{\mathsf{T}}\|_{B,2,\infty}$:

$$\begin{aligned}\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^{\mathsf{T}}\|_{B,2,\infty} &\leqslant \max_{i\in\mathbf{Co}_r}\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(X_i)\boldsymbol{e}^{\mathsf{T}}\|_{2,2} \\ &\leqslant \max_{i\in\mathbf{Co}_r}\sqrt{\rho_i\rho_r}\max_{j\in\rho_i,k\in\rho_r}|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(X_i)_j e_k| \\ &\leqslant \max_{i\in\mathbf{Co}_r}\sqrt{\rho_i\rho_r}\max_{k\in\rho_r}\mathbb{E}_{\mathbb{Q}}|e_k|.\end{aligned}$$

Similarly, applying Hoeffding's inequality and the Kantorovich-Rubinstein theorem gives us

$$\text{Prob}(\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^{\mathsf{T}}\|_{B,2,\infty} \geqslant \sqrt{\rho_{\max}\rho_r}(2\varepsilon\sigma + \mu + t)) \leqslant |\mathbf{Co}_r|\rho_r \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Let $t = \mu$, $\varepsilon \leqslant \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_{\max}\rho_r}}{\alpha}$ hold, we have, with probability at least $1 - |\mathbf{Co}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$,

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^{\mathsf{T}}\|_{B,2,\infty} \leqslant \frac{\lambda_B^*\alpha}{8}.$$

□

Theorem 9. *Given a Bayesian network $(\mathcal{G}, \mathbb{P})$ of n categorical random variables and its skeleton $\mathcal{G}_{skel} := (\mathcal{V}, \mathcal{E}_{skel})$. Assume that the condition $\|\boldsymbol{W}^*\|_{B,2,1} \leqslant \bar{B}$ holds for some $\bar{B} > 0$ associated with an optimal Lagrange multiplier $\lambda_B^* > 0$ for $\boldsymbol{W}^*$ defined in Equation* (1)*. Suppose that $\hat{\boldsymbol{W}}$ is a DRO risk minimizer of Equation* (4) *with a Wasserstein distance of order* 1 *and an ambiguity radius*

$\varepsilon = \varepsilon_0/m$ where m is the number of samples drawn i.i.d. from $\mathbb{P}$. Under Assumptions 1, 2, 3, 4, if the number of samples satisfies

$$m = \mathcal{O}(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}),$$

where C only depends on α, Λ, and if the Lagrange multiplier satisfies

$$\frac{32\mu\rho_{max}}{\alpha} < \lambda_B^* < \frac{\beta}{(\alpha/(4-2\alpha)+2)\rho_{max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}},$$

then for any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1 - \delta$, the following properties hold:

(a) The optimal estimator $\hat{\boldsymbol{W}}$ is unique.

(b) All the non-neighbor nodes are excluded: $\boldsymbol{Co}_r \subseteq \hat{\boldsymbol{Co}}_r$.

(c) All the neighbor nodes are identified: $\boldsymbol{Ne}_r \subseteq \hat{\boldsymbol{Ne}}_r$.

(d) The true skeleton is successfully reconstructed: $\mathcal{G}_{skel} = \hat{\mathcal{G}}_{skel}$.

Proof. We prove the statements in this theorem in several steps. In order to prove (a) and (b), we will show that the DRO problem is strictly convex if true non-neighbors are known so that there is an optimal solution. Next we would like to demonstrate that this solution with a non-neighbor constraint is indeed unique for all the solutions without constraints. The proof for uniqueness comes with a conclusion that we do not accidentally include any edge between the current node and its non-neighbors. Next, to prove (c), we present a generalization bound for the DRO estimator in terms of its true risk, which leads to a ℓ_∞ bound of the difference between the estimator $\hat{\boldsymbol{W}}$ and the true weight matrix $\boldsymbol{W}^*$. Combined with the assumption on the minimum weight, it implies that we include all the neighbor nodes successfully. Finally, by taking a union bound for all the nodes, we could conclude that the correct skeleton is recovered with high probability, which proves (d).

(i) Given the true non-neighbors, there is a unique solution.

We start with the Wasserstein DRO problem, which we recapitulate here for convenience:

$$\hat{\boldsymbol{W}} \in \arg\inf_{\boldsymbol{W}} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)} \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal \mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2.$$

The objective is convex because it is a supremum of convex functions.

For now, we assume that the non-neighbor nodes $\mathbf{Co}_r$ are given. We can then explicitly restrict $\boldsymbol{W}_i = \mathbf{0}$ for all $i \in \mathbf{Co}_r$. The Hessian of $\boldsymbol{W}_{\mathcal{S}_r \cdot}$ is a block diagonal matrix reads

$$\nabla^2 R^{\mathbb{Q}}(\boldsymbol{W}_{\mathcal{S}_r \cdot}) = \begin{bmatrix} \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r \mathcal{S}_r} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r \mathcal{S}_r} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r \mathcal{S}_r} \end{bmatrix} \in \mathbb{R}^{\rho_r \rho_{\mathbf{Ne}_r} \times \rho_r \rho_{\mathbf{Ne}_r}},$$

where

$$\boldsymbol{H}^{\mathbb{Q}} := \mathbb{E}_{\mathbb{Q}}[\mathcal{E}(\boldsymbol{X}_{\bar{r}})\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal] \in \mathbb{R}^{\rho_{\bar{r}} \times \rho_{\bar{r}}}$$

is the covariance matrix of encodings of $\boldsymbol{X}_{\bar{r}}$ under some distribution $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$.

Since $\boldsymbol{W}_{\mathcal{S}_r^c \cdot}$ is fixed to be zero and $\nabla^2 R^{\mathbb{Q}}(\boldsymbol{W}_{\mathcal{S}_r \cdot})$ is a block diagonal matrix, we focus on showing that $\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r \mathcal{S}_r} \succ \mathbf{0}$.

We apply Lemma 7 to get the bound

$$\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r \mathcal{S}_r}) \geqslant \Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r \mathcal{S}_r}) - 4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}} - t,$$

with probability at least $1-2|\mathcal{S}_r|^2\exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right)$. $\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})-4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}}-t>0$ will guarantee that the DRO problem in Equation (4) has a unique solution when the $\boldsymbol{W}_i=\mathbf{0}$ is satisfied for non-neighbor nodes.

(ii) Given the true non-neighbors, the solution is optimal.

We would like to show that the solution to Equation (4) with true non-neighbor constraints is optimal. In this way, we do not recover any non-neighbor nodes in the skeleton. We adopt the primal-dual witness (PDW) [Wainwright, 2009] method to show optimality for the constrained unique solution.

Recall that we assume $\|\boldsymbol{W}\|_{B,2,1}\leqslant\bar{B}$. To begin with, we write the dual problem as

$$\hat{\boldsymbol{W}}\in\arg\inf_{\boldsymbol{W}}\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m),\|\boldsymbol{Z}\|_{B,2,\infty}\leqslant 1,\lambda_B\geqslant 0}\frac{1}{2}\mathbb{E}_{\mathbb{Q}}\|\mathcal{E}(X_r)-\boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2+\lambda_B(\langle\boldsymbol{Z},\boldsymbol{W}\rangle-\bar{B}) \tag{16}$$

$$\text{s.t.}\quad \forall i\in\mathbf{Co}_r\quad \boldsymbol{W}_i=\mathbf{0},$$

where λ_B is the Lagrange multiplier for the norm constraint on $\boldsymbol{W}$.

$\hat{\boldsymbol{W}}$ is optimal if and only if there exists $(\mathbb{Q}^*,\boldsymbol{Z}^*,\lambda_B^*)$ that satisfies the KKT condition:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal\hat{\boldsymbol{W}}-\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\mathcal{E}(\boldsymbol{X}_r)^\intercal+\lambda_B^*\boldsymbol{Z}^*=\mathbf{0}\\
&\mathbb{Q}^*\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m),\|\boldsymbol{Z}^*\|_{B,2,\infty}\leqslant 1,\lambda_B^*\geqslant 0,\|\hat{\boldsymbol{W}}\|_{B,2,1}\leqslant\bar{B}\\
&\langle\boldsymbol{Z}^*,\hat{\boldsymbol{W}}\rangle=\|\hat{\boldsymbol{W}}\|_{B,2,1},\lambda_B^*(\|\hat{\boldsymbol{W}}\|_{B,2,1}-\bar{B})=0.
\end{aligned}$$

Note that we assume that the constraint $\|\boldsymbol{W}\|_{B,2,1}\leqslant\bar{B}$ is active such that $\lambda_B^*>0$. This assumption is only for convenience of theoretical analysis and not restrictive. If it is not active, we have $\|\hat{\boldsymbol{W}}\|_{B,2,1}=\check{B}<\bar{B}$ for some $\check{B}$ and $\lambda_B^*=0$, which leads to an unconstrained problem similar to the ordinary least square problem, which is known to suffer from overfitting. Instead, we are usually interested in solutions that have finite norms so we can always find $\bar{B}=\check{B}-\epsilon<\check{B}$ for some small positive constant $\epsilon>0$ to make the constraint active and thus $\lambda_B^*>0$.

Substituting $\mathcal{E}(X_r)=\boldsymbol{W}^{*\intercal}\mathcal{E}(\boldsymbol{X}_{\bar{r}})+\boldsymbol{e}$ into the first-order optimality condition yields

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\mathcal{E}(\boldsymbol{X}_{\bar{r}})^\intercal(\hat{\boldsymbol{W}}-\boldsymbol{W}^*)-\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\boldsymbol{e}^\intercal+\lambda_B^*\boldsymbol{Z}^*=\mathbf{0}\\
\iff&\begin{bmatrix}\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r} & \boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r^c}\\ \boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r} & \boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r^c}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{W}}_{\mathcal{S}_r\cdot}-\boldsymbol{W}^*_{\mathcal{S}_r\cdot}\\ \mathbf{0}\end{bmatrix}-\begin{bmatrix}\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\\ \mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\end{bmatrix}+\lambda_B^*\begin{bmatrix}\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\\ \boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}.
\end{aligned}\tag{17}$$

Solving for $\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}$, we find that

$$\lambda_B^*\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}=\lambda_B^*\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}-\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal+\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal,$$

which can be bounded such that

$$\begin{aligned}
&\lambda_B^*\|\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\|_{B,2,\infty}\\
=&\|\lambda_B^*\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}-\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal+\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\|_{B,2,\infty}\\
\leqslant&\lambda_B^*\|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{B,2,\infty}+\|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{B,2,\infty}+\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\|_{B,2,\infty}\\
\leqslant&\lambda_B^*\|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}\|\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{2,\infty}+\|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty}\\
&+\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\|_{B,2,\infty}.
\end{aligned}$$

Note that

$$\|\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{2,\infty}\leqslant\|\boldsymbol{Z}^*\|_{B,2,\infty}\leqslant 1.$$

Recall that $0 < \alpha \leqslant 1$ in Assumption 4. Based on Lemma 8 and Lemma 12, we may write

$$\begin{aligned}
&\lambda_B^* \|\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\|_{B,2,\infty} \\
\leqslant &\lambda_B^* \|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \|\boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{2,\infty} + \|\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} \\
&+ \|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r^c}\boldsymbol{e}^\intercal\|_{B,2,\infty} \\
\leqslant &\lambda_B^*(1-\frac{\alpha}{2}) + (1-\frac{\alpha}{2})(\frac{\lambda_B^*\alpha}{8(1-\alpha/2)}) + \frac{\lambda_B^*\alpha}{8} \\
\leqslant &\lambda_B^*(1-\frac{\alpha}{4}) \\
< &\lambda_B^*,
\end{aligned}$$

with high probability and certain conditions on λ_B^* and ε.

Henceforth, $\|\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\|_{B,2,\infty} < 1$ satisfies strict dual feasibility and we must have $\|\hat{\boldsymbol{W}}_{\mathcal{S}_r^c\cdot}\|_{B,2,1} = 0$ according to complementary slackness: $\langle \boldsymbol{Z}^*, \hat{\boldsymbol{W}}\rangle = \|\hat{\boldsymbol{W}}\|_{B,2,1}$. In other words, we have

$$\forall i \in \mathbf{Co}_r \quad \hat{\boldsymbol{W}}_i = \mathbf{0},$$

with high probability. This guarantees that we do not recover any node that is not a neighbor of r with high probability.

(iii) Without information about the true skeleton, we have a unique and optimal solution.

We follow the proof of Lemma 11.2 in Hastie et al. [2015].

We have shown that $\hat{\boldsymbol{W}}$ satisfying $\hat{\boldsymbol{W}}_i = \mathbf{0} \quad \forall i \in \mathbf{Co}_r$ is an optimal solution with optimal dual variables $\|\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\|_{B,2,\infty} < 1$.

To avoid clutter of notations, we define

$$L^{\text{DRO}}(\boldsymbol{W}) := \sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)} \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2.$$

Let $(\check{\boldsymbol{W}}, \check{\lambda})$ be any other optimal solution to $\inf_{\boldsymbol{W}}\sup_\lambda L^{\text{DRO}}(\boldsymbol{W}) + \lambda(\|\boldsymbol{W}\|_{B,2,1} - \bar{B})$. By definition,

$$\begin{aligned}
&L^{\text{DRO}}(\check{\boldsymbol{W}}) + \check{\lambda}(\|\check{\boldsymbol{W}}\|_{B,2,1} - \bar{B}) = L^{\text{DRO}}(\hat{\boldsymbol{W}}) + \lambda_B^*(\langle \boldsymbol{Z}^*, \hat{\boldsymbol{W}}\rangle - \bar{B}) \\
\iff &L^{\text{DRO}}(\check{\boldsymbol{W}}) + \check{\lambda}(\|\check{\boldsymbol{W}}\|_{B,2,1} - \bar{B}) - \lambda_B^*\langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle = L^{\text{DRO}}(\hat{\boldsymbol{W}}) + \lambda_B^*(\langle \boldsymbol{Z}^*, \hat{\boldsymbol{W}} - \check{\boldsymbol{W}}\rangle - \bar{B}).
\end{aligned}$$

The first-order optimality condition for $\hat{\boldsymbol{W}}$ says

$$\nabla L^{\text{DRO}}(\hat{\boldsymbol{W}}) + \lambda_B^*\boldsymbol{Z}^* = \mathbf{0},$$

which implies

$$\check{\lambda}(\|\check{\boldsymbol{W}}\|_{B,2,1} - \bar{B}) + \lambda_B^*(\bar{B} - \langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle) = L^{\text{DRO}}(\hat{\boldsymbol{W}}) + \langle \nabla L^{\text{DRO}}(\hat{\boldsymbol{W}}), \check{\boldsymbol{W}} - \hat{\boldsymbol{W}}\rangle - L^{\text{DRO}}(\check{\boldsymbol{W}}).$$

By definition, $\|\check{\boldsymbol{W}}\|_{B,2,1} - \bar{B} = 0$ and $\lambda_B^* > 0$. Since $L^{\text{DRO}}(\cdot)$ is convex, the RHS of the above equation should be non-positive, or equivalently,

$$\|\check{\boldsymbol{W}}\|_{B,2,1} \leqslant \langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle.$$

On the other hand,

$$\langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle \leqslant \|\boldsymbol{Z}^*\|_{B,2,\infty}\|\check{\boldsymbol{W}}\|_{B,2,1} \leqslant \|\check{\boldsymbol{W}}\|_{B,2,1}.$$

Therefore, the equality holds for the above inequalities, which leads to

$$\|\check{\boldsymbol{W}}\|_{B,2,1} = \langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle.$$

Recall that $\|\boldsymbol{Z}^*_{\mathcal{S}_r^c\cdot}\|_{B,2,\infty} < 1$. In order for $\|\check{\boldsymbol{W}}\|_{B,2,1} = \langle \boldsymbol{Z}^*, \check{\boldsymbol{W}}\rangle$ to hold, we must have

$$\check{\boldsymbol{W}}_{\mathcal{S}_r^c\cdot} = \mathbf{0}.$$

In that wise, all the optimal solutions $\check{\boldsymbol{W}}$ have

$$\check{\boldsymbol{W}}_i = \mathbf{0} \quad \forall i \in \mathbf{Co}_r.$$

This implies that we have a unique solution that excludes all the non-neighbor nodes without information about the true skeleton. Until now, we have proven properties (a) and (b).

(iv) The set of correct neighbors is recovered.

Consider again the first-order optimality condition in Equation (17),

$$\begin{aligned}
\hat{\boldsymbol{W}}_{\mathcal{S}_r\cdot} - \boldsymbol{W}^*_{\mathcal{S}_r\cdot} =& (\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}(\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal - \lambda^*_B \boldsymbol{Z}^*_{\mathcal{S}_r\cdot}) \\
\implies \|\hat{\boldsymbol{W}}_{\mathcal{S}_r\cdot} - \boldsymbol{W}^*_{\mathcal{S}_r\cdot}\|_{B,2,\infty} =& \|(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}(\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal - \lambda^*_B \boldsymbol{Z}^*_{\mathcal{S}_r\cdot})\|_{B,2,\infty} \\
\leqslant& \|(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal - \lambda^*_B \boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{2,\infty} \\
\leqslant& \|(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}(\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} + \|\lambda^*_B \boldsymbol{Z}^*_{\mathcal{S}_r\cdot}\|_{2,\infty}) \\
\leqslant& \rho_{\max}\|(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{\infty,\infty}(\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} + \lambda^*_B) \\
\leqslant& \rho_{\max}\sqrt{|\mathcal{S}_r|}\|(\boldsymbol{H}^{\mathbb{Q}^*}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{2,2}(\|\mathbb{E}_{\mathbb{Q}^*}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} + \lambda^*_B).
\end{aligned}$$

According to Equation (13), with probability at least $1 - 2|\mathcal{S}_r|^2 \exp\left(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right)$ and $\varepsilon \leqslant \frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}$,

$$\|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{2,2} \leqslant \sqrt{\frac{4}{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}}.$$

According to Equation (15), with probability at least $1 - |\mathcal{S}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$, $\varepsilon \leqslant \frac{\mu}{\sigma}$ and $\lambda^*_B > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$, we have

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})_{\mathcal{S}_r}\boldsymbol{e}^\intercal\|_{2,\infty} \leqslant \frac{\lambda^*_B\alpha}{8(1-\alpha/2)}.$$

On that account, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp\left(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right) - |\mathcal{S}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$ and $\varepsilon \leqslant \min\left(\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}, \frac{\mu}{\sigma}\right)$ while requiring $\lambda^*_B > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$,

$$\|\hat{\boldsymbol{W}}_{\mathcal{S}_r\cdot} - \boldsymbol{W}^*_{\mathcal{S}_r\cdot}\|_{B,2,\infty} \leqslant \rho_{\max}\sqrt{|\mathcal{S}_r|}\sqrt{\frac{4}{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}}\lambda^*_B\left(\frac{\alpha}{8(1-\alpha/2)} + 1\right).$$

By Assumption 2, if the condition $\lambda^*_B < \frac{\beta}{2\left(\frac{\alpha}{8(1-\alpha/2)}+1\right)\rho_{\max}\sqrt{|\mathcal{S}_r|}}\sqrt{\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{4}}$ is satisfied, the following inequality holds:

$$\|\hat{\boldsymbol{W}}_{\mathcal{S}_r\cdot} - \boldsymbol{W}^*_{\mathcal{S}_r\cdot}\|_{B,2,\infty} < \beta/2.$$

In this way, we are able to recover all the neighbor nodes with a threshold $\beta/2$. This proves (c).

(v) The true skeleton is recovered with high probability.

The above arguments tell us that with high probability and certain conditions for ε and λ^*_B satisfied, for each node r, we do not recover any non-neighbor and we do recover all the neighbor nodes. The correct $\mathbf{Ne}_r$ and $\mathbf{Co}_r$ are thus identified. Now we are ready to prove (d).

Putting everything together and taking the the union bound for all nodes $r \in [n]$, with probability at least $1 - \mathcal{O}(n\exp\left(-\frac{Cm\mu^2}{\sigma^2\rho^4_{\max}\rho^3_{[n]}} + 2\log\rho_{[n]}\right))$, $\varepsilon \leqslant \frac{C\mu}{\sigma\rho_{\max}\rho^{3/2}_{[n]}}$ and $\frac{32\mu\rho_{\max}}{\alpha} < \lambda^*_B < \frac{\beta}{2\left(\frac{\alpha}{8(1-\alpha/2)}+1\right)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}}$, where C only depends on α, Λ, we have

$$\hat{\mathcal{G}}_{\text{skel}} = \mathcal{G}_{\text{skel}}.$$

Setting $\varepsilon = \frac{\varepsilon_0}{m}$ and making the dependence on the sample size more explicit. We draw the conclusion that, if the number of samples satisfies

$$m = \mathcal{O}(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}),$$

where C only depends on α, Λ, and if λ_B^* satisfies

$$\frac{32\mu\rho_{\max}}{\alpha} < \lambda_B^* < \frac{\beta}{(\alpha/(4-2\alpha)+2)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}},$$

then with probability at least $1-\delta$ for $\delta \in (0, 1]$:

$$\hat{\mathcal{G}}_{\text{skel}} = \mathcal{G}_{\text{skel}}.$$

Moreover, if we assume that the target graph has a bounded degree of d, the sample complexity becomes logarithmic in n:

$$m = \mathcal{O}(\frac{C(\varepsilon_0 + \log(n/\delta) + \log n + \log \rho_{\max})\sigma^2 \rho_{\max}^7 d^3}{\min(\mu^2, 1)}).$$

□

Theorem 10. *Suppose that $\hat{\boldsymbol{W}}$ is a DRO risk minimizer of Equation* (4) *with the KL divergence and an ambiguity radius $\varepsilon = \varepsilon_0/m$. Given the same definitions of $(\mathcal{G}, \mathbb{P})$, $\mathcal{G}_{skel}$, $\bar{B}$, λ_B^*, m in Theorem 9. Under Assumptions 1, 2, 3, 4, if the number of samples satisfies*

$$m = \mathcal{O}(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}).$$

where C depends on α, Λ while independent of n, and if the Lagrange multiplier satisfies the same condition as in Theorem 9, then for any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1-\delta$, the properties (a)-(d) in Theorem 9 hold.

Proof. Define

$$\ell_{\boldsymbol{W}}(\boldsymbol{X}) := \frac{1}{2}\|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal \mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2.$$

According to Theorem 7 in Lam [2019], the worst-case risk with a KL divergence ambiguity set can be bounded as follows:

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^D(\hat{\mathbb{P}}_m)} \mathbb{E}_{\mathbb{Q}}\ell_{\boldsymbol{W}}(\boldsymbol{X}) \leqslant & \mathbb{E}_{\hat{\mathbb{P}}_m}\ell_{\boldsymbol{W}}(\boldsymbol{X}) + \sqrt{\varepsilon}\sqrt{\frac{1}{m}\sum_{i\in[m]}(\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)}) - \bar{\ell_{\boldsymbol{W}}})^2} + C\varepsilon\frac{\sum_{i\in[m]}|\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)}) - \bar{\ell_{\boldsymbol{W}}}|^3}{\sum_{i\in[m]}(\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)}) - \bar{\ell_{\boldsymbol{W}}})^2} \\ \leqslant & \mathbb{E}_{\hat{\mathbb{P}}_m}\ell_{\boldsymbol{W}}(\boldsymbol{X}) + \sqrt{\varepsilon}\max_{i\in[m]}|\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)}) - \bar{\ell_{\boldsymbol{W}}}| + C\varepsilon\max_{i\in[m]}|\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)}) - \bar{\ell_{\boldsymbol{W}}}|,\end{aligned}$$

where $\bar{\ell_{\boldsymbol{W}}} = \frac{1}{m}\sum_{i\in[m]}\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})$ and $C > 0$ is constant independent of n.

Consider

$$\begin{aligned}\max_{i\in[m]}|\ell_{\boldsymbol{W}}(\boldsymbol{x}^{(i)})-\bar{\ell_{\boldsymbol{W}}}| \leqslant & \max_{\boldsymbol{W},\boldsymbol{W}',\boldsymbol{x},\boldsymbol{x}'}|\ell_{\boldsymbol{W}}(\boldsymbol{x})-\ell_{\boldsymbol{W}'}(\boldsymbol{x}')| \\ \leqslant & \max_{\boldsymbol{W},\boldsymbol{x}}|\ell_{\boldsymbol{W}}(\boldsymbol{x})| \\ \leqslant & \frac{1}{2}\max_{\boldsymbol{W},\boldsymbol{x}}(\|\mathcal{E}(X_r)\|_2+\|\boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2)^2 \\ \leqslant & \frac{1}{2}\max_{\boldsymbol{W},\boldsymbol{x}}(\sqrt{\rho_{\max}}+\|\!|\boldsymbol{W}^\intercal|\!\|_{\infty,2})^2 \\ \leqslant & \frac{1}{2}\max_{\boldsymbol{W},\boldsymbol{x}}(\sqrt{\rho_{\max}}+\|\boldsymbol{W}\|_{1,2})^2 \\ \leqslant & \frac{1}{2}\max_{\boldsymbol{W},\boldsymbol{x}}(\sqrt{\rho_{\max}}+\sqrt{\rho_{[n]}}\|\boldsymbol{W}\|_F)^2 \\ \leqslant & \frac{1}{2}\max_{\boldsymbol{W},\boldsymbol{x}}(\sqrt{\rho_{\max}}+\sqrt{\rho_{[n]}}\|\boldsymbol{W}\|_{B,2,1})^2 \\ \leqslant & \frac{1}{2}(\sqrt{\rho_{\max}}+\sqrt{\rho_{[n]}}\bar{B})^2 \\ := & B_\rho.\end{aligned}$$

Define $\varepsilon_{\max}:=\max(\sqrt{\varepsilon},\varepsilon)$. Therefore, we find that

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m)}\mathbb{E}_{\mathbb{Q}}\ell_{\boldsymbol{W}}(\boldsymbol{X})\leqslant\mathbb{E}_{\tilde{\mathbb{P}}_m}\ell_{\boldsymbol{W}}(\boldsymbol{X})+C\varepsilon_{\max}B_\rho.$$

Similar to the Wasserstein robust risk, we observe that the following results hold for any $\mathbb{Q}\in\mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m)$.

With probability at least $1-2|\mathcal{S}_r|^2\exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$, we have

$$\Lambda_{\min}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})\geqslant\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})-C\varepsilon_{\max}|\mathcal{S}_r|^{\frac{1}{2}}-t.$$

With probability at least $1-2|\mathcal{S}_r^c||\mathcal{S}_r|\exp(-\frac{mt^2}{2\rho_{\max}^2|\mathcal{S}_r|^2})$,

$$\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}-\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty}\leqslant C\varepsilon_{\max}\rho_{\max}|\mathcal{S}_r|+t.$$

With probability at least $1-2|\mathcal{S}_r|^2\exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$,

$$|\!|\!|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r}-\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}|\!|\!|_{\infty,\infty}\leqslant C\varepsilon_{\max}|\mathcal{S}_r|+t.$$

With probability at least $1-2|\mathcal{S}_r|^2\exp(-\frac{mt^2(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3})-2|\mathcal{S}_r|^2\exp(-\frac{m(\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and $\varepsilon_{\max}\leqslant C\min(\frac{t\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}},\frac{\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}})$,

$$|\!|\!|(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}-(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}|\!|\!|_{\infty,\infty}\leqslant t.$$

With probability at least $1-\mathcal{O}(\exp(-\frac{Cm}{\rho_{\max}^2|\mathcal{S}_r|^3}+\log|\mathcal{S}_r^c|+\log|\mathcal{S}_r|))$ and $\varepsilon_{\max}\leqslant\frac{C}{\rho_{\max}|\mathcal{S}_r|^{3/2}}$,

$$\|\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r^c\mathcal{S}_r}(\boldsymbol{H}^{\mathbb{Q}}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}\leqslant 1-\frac{\alpha}{2},$$

where C only depends on α, $\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r})$.

Thanks to the boundedness of the error term $\boldsymbol{e}$, we have similar conclusions to Lemma 12 if $\varepsilon_{\max}\leqslant\frac{\mu}{\sigma}$ holds.

In such wise, the properties in Theorem 9 hold with the same condition on λ_B^* and the condition on $\varepsilon_{\max}$ that $\varepsilon_{\max}\leqslant\frac{C\mu}{\sigma\rho_{\max}\rho_{[n]}^{3/2}}$. Since we set $\varepsilon=\frac{\varepsilon_0}{m}$ and define $\varepsilon_{\max}:=\max(\sqrt{\varepsilon},\varepsilon)$, the condition on $\varepsilon_{\max}$ implies that

$$m\geqslant\max(\frac{\varepsilon_0C^2\sigma^2\rho_{\max}^2\rho_{[n]}^3}{\mu^2},\frac{\varepsilon_0C\sigma\rho_{\max}\rho_{[n]}^{3/2}}{\mu}).$$

Table 2: Comparisons of F1 scores for benchmark datasets and BIC for real-world datasets (backache, voting). BIC is not applicable to skeletons. The best and runner-up results are marked in bold. Significant differences are marked by † (paired t-test, $p < 0.05$).

Dataset	n	m	Noise	ζ	Wass	KL	Reg	MMPC	GRASP	Wass+HC	KL+HC	Reg+HC	MMPC+HC	GRASP+HC	HC
asia	8	1000	Noisefree	0	0.7800†	0.7285†	0.7897†	**0.9067**	**0.8167**	0.5123	0.6367	0.5743	**0.6667**	**0.6583**	0.6550
asia	8	1000	Huber	0.2	**0.7333**†	0.7124†	**0.7297**†	0.5468	0.6570	**0.3943**	**0.3724**	0.3487	0.2907	0.3664	0.2183
asia	8	1000	Independent	0.2	**0.6933**	0.6797	**0.6868**	0.6359	0.3632†	**0.2676**	**0.2632**	0.2581	0.2469	0.1794	0.2443
cancer	5	1000	Noisefree	0	**1.0000**†	**1.0000**†	**1.0000**†	0.6133	0.6133	**0.2800**	**0.2800**	**0.2800**	**0.2800**	**0.2800**	**0.2800**
cancer	5	1000	Huber	0.5	**0.9156**†	0.8933†	**0.9092**†	0.6133	0.5357	**0.4333**	0.3833	**0.4143**	0.2589	0.2714	0.2589
cancer	5	1000	Independent	0.2	**0.9048**†	**0.9029**†	0.8992†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
earthquake	5	1000	Noisefree	0	0.8447†	0.9333†	**0.9778**	**1.0000**	**0.9778**	0.2000	**0.2500**	**0.2500**	**0.2500**	**0.2500**	0.2278†
earthquake	5	1000	Huber	0.2	**0.7509**†	**0.7509**†	**0.7509**†	0.5978	0.6583†	**0.4618**	**0.4618**	**0.4618**	0.3860	0.4547	0.3860
earthquake	5	1000	Independent	0.2	**0.6786**†	**0.6350**†	**0.6350**†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
sachs	11	1000	Noisefree	0	0.8357†	**0.8402**†	0.8374†	**0.9697**	0.7678†	0.4310†	0.4535†	0.4641†	**0.5935**	0.4112†	**0.5873**
sachs	11	1000	Huber	0.2	0.7765	**0.8064**	**0.7893**	0.7498	0.5663†	**0.5194**	0.4815	0.4520	0.4736	0.2380	**0.5028**
sachs	11	1000	Independent	0.5	**0.5268**†	**0.5208**†	0.5172†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
survey	6	1000	Noisefree	0	**0.6596**	**0.6545**	0.6506	0.6533	0.1714†	**0.1789**	**0.1789**	**0.1789**	**0.1789**	0.0571	**0.1789**
survey	6	1000	Huber	0.2	**0.7303**†	0.6778†	**0.7095**†	0.5396	0.3810	**0.1444**	**0.1444**	**0.1444**	**0.1444**	**0.1516**	**0.1444**
survey	6	1000	Independent	0.2	**0.6311**†	**0.6705**†	0.6220†	0.2032	0.0000†	**0.1071**	**0.1071**	**0.1143**	**0.1071**	0.0000	**0.1071**
alarm	37	1000	Noisefree	0	0.4750†	0.7863†	**0.8042**†	**0.8530**	0.6824†	0.3483†	0.4949†	0.4470†	**0.5635**	**0.4976**	0.4494†
alarm	37	1000	Huber	0.2	0.1432†	0.1619†	**0.6571**†	**0.5486**	0.1945†	0.2192	0.1680†	**0.3148**	**0.2774**	0.2092†	0.2582
alarm	37	1000	Independent	0.2	0.1419†	0.1448†	**0.5458**†	**0.4309**	0.2830†	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
barley	48	1000	Noisefree	0	0.1521†	0.2632†	0.4913†	**0.5847**	**0.5636**	0.1995	0.1970†	0.2503	**0.2510**	0.2245	**0.2526**
barley	48	1000	Huber	0.2	0.1452†	0.1592†	**0.4027**†	**0.4522**	0.4000†	0.1396	0.1151	**0.1658**	0.1463	0.1530	**0.1685**
barley	48	1000	Independent	0.2	0.1463†	0.1501†	0.2767†	**0.4273**	**0.4923**†	0.0598	0.0769	**0.0838**	0.0727	**0.0840**	**0.0838**
voting	17	216	Noisefree	0	N/A	N/A	N/A	N/A	N/A	**−2451.8631**	**−2453.2737**	−2453.4091	−2475.5799	−2482.3835	−2456.1489
voting	17	216	Huber	0.2	N/A	N/A	N/A	N/A	N/A	**−4418.9731**	**−4418.9731**	−4487.4544	−4450.3941	−4445.0175	**−4418.9731**
voting	17	216	Independent	0.2	N/A	N/A	N/A	N/A	N/A	**−4453.8298**	**−4453.8298**	−4522.5521	−4465.1076	−4473.8612	**−4453.8298**
backache	32	90	Noisefree	0	N/A	N/A	N/A	N/A	N/A	−1729.8364	−1726.8465	**−1710.7248**	−1719.5002	**−1713.7583**	−1729.7991
backache	32	90	Huber	0.2	N/A	N/A	N/A	N/A	N/A	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**
backache	32	90	Independent	0.2	N/A	N/A	N/A	N/A	N/A	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**
connect-4_6000	43	6000	Noisefree	0	N/A	N/A	N/A	N/A	N/A	**−38956.4300**	**−38956.4300**	**−38954.9501**	−39004.8512	−39933.6041†	**−38956.4300**
connect-4_6000	43	6000	Huber	0.2	N/A	N/A	N/A	N/A	N/A	**−99616.2848**	**−99616.2848**	−102878.2766	−99673.5320	−100212.9773	**−99616.2848**
connect-4_6000	43	6000	Independent	0.2	N/A	N/A	N/A	N/A	N/A	**−107403.2543**	**−107403.2543**	**−107403.2543**	**−107403.2543**	**−107403.2543**	**−107403.2543**

The final sample complexity becomes

$$m = \mathcal{O}(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}).$$

□

D More Empirical Results

Table 2 lists the complete experimental results.