Appendix A: Perfect Recovery Guarantee for the Problem (5)

The following theorem shows the perfect recovery guarantee for the problem (5). Appendix C provides the proof for completeness.

Theorem 7.1. *Let $\boldsymbol{X}^* \in \mathbb{R}^{n\times n}$ be a rank k matrix with a singular value decomposition $\boldsymbol{X}^* = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$, where $\boldsymbol{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbb{R}^{n\times k}$ and $\boldsymbol{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathbb{R}^{n\times k}$ are the left and right singular vectors of $\boldsymbol{X}^*$, respectively. Similar to many related works of matrix completion, we assume that the following two assumptions are satisfied:*

1. *The row and column spaces of $\boldsymbol{X}$ have coherence bounded above by a positive number μ_0.*
2. *Max absolute value in matrix $\boldsymbol{U}\boldsymbol{V}^\top$ is bounded above by $\mu_1\sqrt{r}/n$ for a positive number μ_1.*

Suppose that m_1 entries of $\boldsymbol{X}^*$ are observed with their locations sampled uniformly at random, and among the m_1 observed entries, m_2 randomly sampled entries are corrupted. Using the resulting partially observed matrix as the input to the problem (5), then with a probability at least $1 - n^{-3}$, the underlying matrix $\boldsymbol{X}^*$ can be perfectly recovered, given

1. $\mu(\boldsymbol{E})\xi(\boldsymbol{X}) \leq \frac{1}{4k+5}$,
2. $\frac{\xi(\boldsymbol{X})-(2k-1)\mu(\boldsymbol{E})\xi^2(\boldsymbol{X})}{1-2(k+1)\mu(\boldsymbol{E})\xi(\boldsymbol{X})} < \lambda < \frac{1-(4k+5)\mu(\boldsymbol{E})\xi(\boldsymbol{X})}{(k+2)\mu(\boldsymbol{E})}$,
3. $m_1 - m_2 \geq C[\max(\mu_0, \mu_1)]^4 n \log^2 n$,

where C is a positive constant; $\xi(\circ)$ and $\mu(\circ)$ denotes the low-rank and sparsity incoherence (Chandrasekaran et al., 2011).

Theorem 7.1 implies that even if some of the observed entries computed by (4) are incorrect, problem (5) can still perfectly recover the underlying similarity matrix $\mathbf{X}^*$ if the number of observed correct entries is at least $O(n\log^2 n)$. For MATL with large n, this implies that only a tiny fraction of all task pairs is needed to reliably infer similarities over all task pairs. Moreover, the completed similarity matrix $\mathbf{X}$ is symmetric, due to symmetry of the input matrix $\mathbf{Y}$. This enables analysis by similarity-based clustering algorithms, such as spectral clustering.

Appendix B: Proof of Low-rankness of Matrix X

We first prove that the full similarity matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ is of low-rank. To see this, let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$ be the underlying perfect clustering result, where k is the number of clusters and $\mathbf{a}_i \in \{0, 1\}^n$ is the membership vector for the i-th cluster. Given $\mathbf{A}$, the similarity matrix $\mathbf{X}$ is computed as

$$\mathbf{X} = \sum_{i=1}^{k} \mathbf{a}_i\mathbf{a}_i^\top = \sum_{i=1}^{k} \mathbf{B}_i$$

where $\mathbf{B}_i = \mathbf{a}_i\mathbf{a}_i^\top$ is a rank one matrix. Using the fact that $\text{rank}(\mathbf{X}) \leq \sum_{i=1}^{k} \text{rank}(\mathbf{B}_i)$ and $\text{rank}(\mathbf{B}_i) = 1$, we have $\text{rank}(\mathbf{X}) \leq k$, i.e., the rank of the similarity matrix $\mathbf{X}$ is upper bounded by the number of clusters. Since the number of clusters is usually small, the similarity matrix $\mathbf{X}$ should be of low rank.

Appendix C: Proof of Theorem 7.1

We then prove our main theorem. First, we define several notations that are used throughout the proof. Let $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the singular value decomposition of matrix $\mathbf{X}$, where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbb{R}^{n\times k}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathbb{R}^{n\times k}$ are the left and right singular vectors of matrix $\mathbf{X}$, respectively. Similar to many related works of matrix completion, we assume that the following two assumptions are satisfied:

1. **A1**: the row and column spaces of $\mathbf{X}$ have coherence bounded above by a positive number μ_0, i.e., $\sqrt{n/r}\max_i \|\mathbf{P}_\mathbf{U}(\mathbf{e}_i)\| \leq \mu_0$ and $\sqrt{n/r}\max_i \|\mathbf{P}_\mathbf{V}(\mathbf{e}_i)\| \leq \mu_0$, where $\mathbf{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{P}_\mathbf{V} = \mathbf{V}\mathbf{V}^\top$, and $\mathbf{e}_i$ is the standard basis vector, and

2. **A2**: the matrix $\mathbf{UV}^\top$ has a maximum entry bounded by $\mu_1\sqrt{r}/n$ in absolute value for a positive number μ_1.

Let T be the space spanned by the elements of the form $\mathbf{u}_i\mathbf{y}^\top$ and $\mathbf{x}\mathbf{v}_i^\top$, for $1 \le i \le k$, where $\mathbf{x}$ and $\mathbf{y}$ are arbitrary n-dimensional vectors. Let $T^\perp$ be the orthogonal complement to the space T, and let $\mathbf{P}_T$ be the orthogonal projection onto the subspace T given by

$$\mathbf{P}_T(\mathbf{Z}) = \mathbf{P}_\mathbf{U}\mathbf{Z} + \mathbf{Z}\mathbf{P}_\mathbf{V} - \mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V}.$$

The following proposition shows that for any matrix $\mathbf{Z} \in T$, it is a zero matrix if enough amount of its entries are zero.

Proposition 1. *Let Ω be a set of m entries sampled uniformly at random from $[1,\dots,n]\times[1,\dots,n]$, and $\boldsymbol{P}_\Omega(\mathbf{Z})$ projects matrix $\mathbf{Z}$ onto the subset Ω. If $m > m_0$, where $m_0 = C_R^2\mu_0 rn\beta\log n$ with $\beta > 1$ and C_R being a positive constant, then for any $\mathbf{Z} \in T$ with $\boldsymbol{P}_\Omega(\mathbf{Z}) = 0$, we have $\mathbf{Z} = 0$ with probability $1 - 3n^{-\beta}$.*

Proof. According to the Theorem 3.2 in (Candès and Tao, 2010), for any $\mathbf{Z} \in T$, with a probability at least $1 - 2n^{2-2\beta}$, we have

$$\|\mathbf{P}_T(\mathbf{Z})\|_F - \delta\|\mathbf{Z}\|_F \le \frac{n^2}{m}\|\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T(\mathbf{Z})\|_F^2 = 0 \tag{8}$$

where $\delta = m_0/m < 1$. Since $\mathbf{Z} \in T$, we have $P_T(\mathbf{Z}) = \mathbf{Z}$. Then from (8), we have $\|\mathbf{Z}\|_F \le 0$ and thus $\mathbf{Z} = 0$. □

In the following, we will develop a theorem for the dual certificate that guarantees the unique optimal solution to the following optimization problem

$$\begin{aligned} \min_{\mathbf{X},\,\mathbf{E}} \quad & \|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{P}_\Omega(\mathbf{X}+\mathbf{E}) = \mathbf{P}_\Omega(\mathbf{Y}). \end{aligned} \tag{9}$$

Theorem 1. *Suppose we observe m_1 entries of $\boldsymbol{X}$ with locations sampled uniformly at random, denoted by Ω. We further assume that m_2 entries randomly sampled from m_1 observed entries are corrupted, denoted by Δ. Suppose that $\boldsymbol{P}_\Omega(\boldsymbol{Y}) = \boldsymbol{P}_\Omega(\boldsymbol{X}+\boldsymbol{E})$ and the number of observed correct entries $m_1 - m_2 > m_0 = C_R^2\mu_0 rn\beta\log n$. Then, for any $\beta > 1$, with a probability at least $1 - 3n^{-\beta}$, the underlying true matrices $(\boldsymbol{X},\boldsymbol{E})$ is the unique optimizer of (9) if both assumptions* **A1** *and* **A2** *are satisfied and there exists a dual $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ such that (a) $\boldsymbol{Q} = \boldsymbol{P}_\Omega(\boldsymbol{Q})$, (b) $\boldsymbol{P}_T(\boldsymbol{Q}) = \boldsymbol{U}\boldsymbol{V}^\top$, (c) $\|\boldsymbol{P}_{T^\top}(\boldsymbol{Q})\| < 1$, (d) $\boldsymbol{P}_\Delta(\boldsymbol{Q}) = \lambda\, sgn(\boldsymbol{E})$, and (e) $\|\boldsymbol{P}_{\Delta^c}(\boldsymbol{Q})\|_\infty < \lambda$.*

Proof. First, the existence of $\mathbf{Q}$ satisfying the conditions (a) to (e) ensures that $(\mathbf{X},\mathbf{E})$ is an optimal solution. We only need to show its uniqueness and we prove it by contradiction. Assume there exists another optimal solution $(\mathbf{X}+\mathbf{N}_\mathbf{X}, \mathbf{E}+\mathbf{N}_\mathbf{E})$, where $\mathbf{P}_\Omega(\mathbf{N}_\mathbf{X}+\mathbf{N}_\mathbf{E}) = 0$. Then we have

$$\|\mathbf{X}+\mathbf{N}_\mathbf{X}\|_* + \lambda\|\mathbf{E}+\mathbf{N}_\mathbf{E}\|_1 \;\;\ge\;\; \|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_1 + \langle\mathbf{Q}_\mathbf{E},\mathbf{N}_\mathbf{E}\rangle + \langle\mathbf{Q}_\mathbf{X},\mathbf{N}_\mathbf{X}\rangle$$

where $\mathbf{Q}_\mathbf{E}$ and $\mathbf{Q}_\mathbf{X}$ satisfying $\mathbf{P}_\Delta(\mathbf{Q}_\mathbf{E}) = \lambda\ \mathrm{sgn}(\mathbf{E})$, $\|\mathbf{P}_{\Delta^c}(\mathbf{Q}_\mathbf{E})\|_\infty \le \lambda$, $\mathbf{P}_T(\mathbf{Q}_\mathbf{X}) = \mathbf{U}\mathbf{V}^\top$ and $\|\mathbf{P}_{T^\perp}(\mathbf{Q}_\mathbf{X})\| \le 1$. As a result, we have

$$\begin{aligned} & \lambda\|\mathbf{E}+\mathbf{N}_\mathbf{E}\|_1 + \|\mathbf{X}+\mathbf{N}_\mathbf{X}\|_* \\ \ge\;& \lambda\|\mathbf{E}\|_1 + \|\mathbf{X}\|_* + \langle\mathbf{Q}+\mathbf{P}_{\Delta^c}(\mathbf{Q}_\mathbf{E}) - \mathbf{P}_{\Delta^c}(\mathbf{Q}),\mathbf{N}_\mathbf{E}\rangle + \langle\mathbf{Q}+\mathbf{P}_{T^\perp}(\mathbf{Q}_\mathbf{X}) - \mathbf{P}_{T^\perp}(\mathbf{Q}),\mathbf{N}_\mathbf{X}\rangle \\ =\;& \lambda\|\mathbf{E}\|_1 + \|\mathbf{X}\|_* + \langle\mathbf{Q},\mathbf{N}_\mathbf{E}+\mathbf{N}_\mathbf{X}\rangle + \langle\mathbf{P}_{\Delta^c}(\mathbf{Q}_\mathbf{E}) - \mathbf{P}_{\Delta^c}(\mathbf{Q}),\mathbf{N}_\mathbf{E}\rangle + \langle\mathbf{P}_{T^\perp}(\mathbf{Q}_\mathbf{X}) - \mathbf{P}_{T^\perp}(\mathbf{Q}),\mathbf{N}_\mathbf{X}\rangle \\ =\;& \lambda\|\mathbf{E}\|_1 + \|\mathbf{X}\|_* + \langle\mathbf{P}_{\Delta^c}(\mathbf{Q}_\mathbf{E}) - \mathbf{P}_{\Delta^c}(\mathbf{Q}),\mathbf{P}_{\Delta^c}(\mathbf{N}_\mathbf{E})\rangle + \langle\mathbf{P}_{T^\perp}(\mathbf{Q}_\mathbf{X}) - \mathbf{P}_{T^\perp}(\mathbf{Q}),\mathbf{P}_{T^\perp}(\mathbf{N}_\mathbf{X})\rangle \end{aligned}$$

We then choose $\mathbf{P}_{\Delta^c}(\mathbf{Q_E})$ and $\mathbf{P}_{T^\perp}(\mathbf{Q_X})$ to be such that $\langle \mathbf{P}_{\Delta^c}(\mathbf{Q_E}), \mathbf{P}_{\Delta^c}(\mathbf{N_E})\rangle = \lambda\|\mathbf{P}_{\Delta^c}(\mathbf{N_E})\|_1$ and $\langle \mathbf{P}_{T^\perp}(\mathbf{Q_X}), \mathbf{P}_{T^\perp}(\mathbf{N_X})\rangle = \|\mathbf{P}_{T^\perp}(\mathbf{N_X})\|_*$. We thus have

$$\begin{aligned} & \lambda\|\mathbf{E}+\mathbf{N_E}\|_1 + \|\mathbf{X}+\mathbf{N_X}\|_* \\ \geq\ & \lambda\|\mathbf{E}\|_1 + \|\mathbf{X}\|_* + (\lambda - \|\mathbf{P}_{\Delta^c}(\mathbf{Q})\|_\infty)\|\mathbf{P}_{\Delta^c}(\mathbf{N_E})\|_1 + (1 - \|\mathbf{P}_{T^\perp}(\mathbf{Q})\|)\|\mathbf{P}_{T^\perp}(\mathbf{N_X})\|_* \end{aligned}$$

Since $(\mathbf{X}+\mathbf{N_X}, \mathbf{E}+\mathbf{N_E})$ is also an optimal solution, we have $\|\mathbf{P}_{\Omega^c}(\mathbf{N}_E)\|_1 = \|\mathbf{P}_{T^\perp}(\mathbf{N_X})\|_*$, leading to $\mathbf{P}_{\Omega^c}(\mathbf{N_E}) = \mathbf{P}_{T^\perp}(\mathbf{N_X}) = 0$, or $\mathbf{N_X} \in T$. Since $\mathbf{P}_\Omega(\mathbf{N_X}+\mathbf{N_E}) = 0$, we have $\mathbf{N_X} = \mathbf{N_E} + \mathbf{Z}$, where $P_\Omega(\mathbf{Z}) = 0$ and $\mathbf{P}_{\Omega^c}(\mathbf{N_E}) = 0$. Hence, $\mathbf{P}_{\Omega^c\cap\Omega}(\mathbf{N_X}) = 0$, where $|\Omega^c \cap \Omega| = m_1 - m_2$. Since $m_1 - m_2 > m_0$, according to Proposition 1, we have, with a probability $1 - 3n^{-\beta}$, $\mathbf{N_X} = 0$. Besides, since $\mathbf{P}_\Omega(\mathbf{N_X}+\mathbf{N_E}) = \mathbf{P}_\Omega(\mathbf{N_E}) = 0$ and $\Delta \subset \Omega$, we have $\mathbf{P}_\Delta(\mathbf{N_E}) = 0$. Since $\mathbf{N_E} = \mathbf{P}_\Delta(\mathbf{N_E}) + \mathbf{P}_{\Delta^c}(\mathbf{N_E})$, we have $\mathbf{N_E} = 0$, which leads to the contradiction. □

Given Theorem 1, we are now ready to prove Theorem 3.1.

Proof. The key to the proof is to construct the matrix $\mathbf{Q}$ that satisfies the conditions (a)-(e) specified in Theorem 1. First, according to Theorem 1, when $m_1 - m_2 > m_0 = C_R^2\mu_0 rn\beta\log n$, with a probability at least $1 - 3n^{-\beta}$, mapping $\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T(\mathbf{Z}) : T \mapsto T$ is an one to one mapping and therefore its inverse mapping, denoted by $(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}$ is well defined. Similar to the proof of Theorem 2 in (Chandrasekaran et al., 2011), we construct the dual certificate $\mathbf{Q}$ as follows

$$\mathbf{Q} = \lambda\,\mathrm{sgn}(\mathbf{E}) + \epsilon_\Delta + \mathbf{P}_\Delta\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\mathbf{U}\mathbf{V}^\top + \epsilon_T)$$

where $\epsilon_T \in T$ and $\epsilon_\Delta = \mathbf{P}_\Delta(\epsilon_\Delta)$. We further define

$$\begin{aligned} \mathbf{H} &= \mathbf{P}_\Omega\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\mathbf{U}\mathbf{V}^\top) \\ \mathbf{F} &= \mathbf{P}_\Omega\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\epsilon_T) \end{aligned}$$

Evidently, we have $\mathbf{P}_\Omega(\mathbf{Q}) = \mathbf{Q}$ since $\Delta \subset \Omega$, and therefore the condition (a) is satisfied. To satisfy the conditions (b)-(e), we need

$$\begin{aligned} \mathbf{P}_T(\mathbf{Q}) = \mathbf{U}\mathbf{V}^\top &\rightarrow \epsilon_T = -\mathbf{P}_T(\lambda\,\mathrm{sgn}(\mathbf{E}) + \epsilon_\Delta) && (10)\\ \|\mathbf{P}_{T^\perp}(\mathbf{Q})\| < 1 &\rightarrow \mu(\mathbf{E})\,(\lambda + \|\epsilon_\Delta\|_\infty) + \|\mathbf{P}_{T^\perp}(\mathbf{H})\| + \|\mathbf{P}_{T^\perp}(\mathbf{F})\| < 1 && (11)\\ \mathbf{P}_\Delta(\mathbf{Q}) = \lambda\,\mathrm{sgn}(\mathbf{E}) &\rightarrow \epsilon_\Delta = -\mathbf{P}_\Delta(\mathbf{H}+\mathbf{F}) && (12)\\ |\mathbf{P}_{\Delta^c}(\mathbf{Q})|_\infty < \lambda &\rightarrow \xi(\mathbf{X})(1 + \|\epsilon_T\|) < \lambda && (13) \end{aligned}$$

Below, we will first show that there exist solutions $\epsilon_T \in T$ and ϵ_Δ that satisfy conditions (10) and (12). We will then bound $\|\epsilon_\Omega\|_\infty$, $\|\epsilon_T\|$, $\|\mathbf{P}_{T^\perp}(\mathbf{H})\|$, and $\|\mathbf{P}_{T^\perp}(\mathbf{F})\|$ to show that with sufficiently small $\mu(\mathbf{E})$ and $\xi(\mathbf{X})$, and appropriately chosen λ, conditions (11) and (13) can be satisfied as well.

First, we show the existence of ϵ_Δ and ϵ_T that obey the relationships in (10) and (12). It is equivalent to show that there exists ϵ_T that satisfies the following relation

$$\epsilon_T = -\mathbf{P}_T(\lambda\,\mathrm{sgn}(\mathbf{E})) + \mathbf{P}_T\mathbf{P}_\Delta(\mathbf{H}) + \mathbf{P}_T\mathbf{P}_\Delta\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\epsilon_T)$$

or

$$\mathbf{P}_T\mathbf{P}_{\Omega\setminus\Delta}\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\epsilon_T) = -\mathbf{P}_T(\lambda\,\mathrm{sgn}(\mathbf{E})) + \mathbf{P}_T\mathbf{P}_\Delta(\mathbf{H}),$$

where $\Omega \setminus \Delta$ indicates the complement set of set Δ in Ω and $|\Omega \setminus \Delta|$ denotes its cardinality. Similar to the previous argument, when $|\Omega \setminus \Delta| = m_1 - m_2 > m_0$, with a probability $1 - 3n^{-\beta}$, $\mathbf{P}_T\mathbf{P}_{\Omega\setminus\Delta}\mathbf{P}_T(\mathbf{Z}) : T \mapsto T$ is an one to one mapping, and therefore $(\mathbf{P}_T\mathbf{P}_{\Omega\setminus\Delta}\mathbf{P}_T(\mathbf{Z}))^{-1}$ is well defined. Using this result, we have the following solution to the above equation

$$\epsilon_T = \mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_{\Omega\setminus\Delta}\mathbf{P}_T)^{-1}\left(-\mathbf{P}_T(\lambda\,\mathrm{sgn}(\mathbf{E})) + \mathbf{P}_T\mathbf{P}_\Delta(\mathbf{H})\right)$$

We now bound $\|\epsilon_T\|$ and $\|\epsilon_\Delta\|_\infty$. Since $\|\epsilon_T\| \leq \|\epsilon_T\|_F$, we bound $\|\epsilon_T\|_F$ instead. First, according to Corollary 3.5 in (Candès and Tao, 2010), when $\beta = 4$, with a probability $1 - n^{-3}$, for any $\mathbf{Z} \in T$, we have

$$\left\|\mathbf{P}_{T^\perp}\mathbf{P}_\Omega\mathbf{P}_T(\mathbf{P}_T\mathbf{P}_\Omega\mathbf{P}_T)^{-1}(\mathbf{Z})\right\|_F \leq \|\mathbf{Z}\|_F.$$

Using this result, we have

$$\begin{aligned}
\|\epsilon_\Delta\|_\infty &\leq \xi(\mathbf{X})\left(\|\mathbf{H}\| + \|\mathbf{F}\|\right) \\
&\leq \xi(\mathbf{X})\left(1 + \|\mathbf{P}_{T^\perp}(\mathbf{H})\|_F + \|\epsilon_T\| + \|\mathbf{P}_{T^\perp}(\mathbf{F})\|_F\right) \\
&\leq \xi(\mathbf{X})\left(2 + \|\epsilon_T\| + \|\epsilon_T\|_F\right) \\
&\leq \xi(\mathbf{X})\left[2 + (2k+1)\|\epsilon_T\|\right]
\end{aligned}$$

In the last step, we use the fact that $\text{rank}(\epsilon_T) \leq 2k$ if $\epsilon_T \in T$. We then proceed to bound $\|\epsilon_T\|$ as follows

$$\|\epsilon_T\| \leq \mu(\mathbf{E})\left(\lambda + \|\epsilon_\Delta\|_\infty\right)$$

Combining the above two inequalities together, we have

$$\begin{aligned}
\|\epsilon_T\| &\leq \xi(\mathbf{X})\mu(\mathbf{E})(2k+1)\|\epsilon_T\| + 2\xi(\mathbf{X})\mu(\mathbf{E}) + \lambda\mu(\mathbf{E}) \\
\|\epsilon_\Delta\|_\infty &\leq \xi(\mathbf{X})\left[2 + (2k+1)\mu(\mathbf{E})(\lambda + \|\epsilon_\Delta\|_\infty)\right],
\end{aligned}$$

which lead to

$$\begin{aligned}
\|\epsilon_T\| &\leq \frac{\lambda\mu(\mathbf{E}) + 2\xi(\mathbf{X})\mu(\mathbf{E})}{1 - (2k+1)\xi(\mathbf{X})\mu(\mathbf{E})} \\
\|\epsilon_\Delta\|_\infty &\leq \frac{2\xi(\mathbf{X}) + (2k+1)\lambda\xi(\mathbf{X})\mu(\mathbf{E})}{1 - (2k+1)\xi(\mathbf{X})\mu(\mathbf{E})}
\end{aligned}$$

Using the bound for $\|\epsilon_\Delta\|_\infty$ and $\|\epsilon_T\|$, we now check the condition (11)

$$1 > \mu(\mathbf{E})\left(\lambda + |\epsilon_\Delta|_\infty\right) + \frac{1}{2} + \frac{k}{2}\|\epsilon_T\|$$

or

$$\lambda < \frac{1 - \xi(\mathbf{X})\mu(\mathbf{E})(4k+5)}{\mu(\mathbf{E})(k+2)}$$

For the condition (13), we have

$$\lambda > \xi(\mathbf{X}) + \xi(\mathbf{X})\|\epsilon_T\|$$

or

$$\lambda > \frac{\xi(\mathbf{X}) - (2k-1)\xi^2(\mathbf{X})\mu(\mathbf{E})}{1 - 2(k+1)\xi(\mathbf{X})\mu(\mathbf{E})}$$

To ensure that there exists $\lambda \geq 0$ satisfies the above two conditions, we have

$$1 - 5(k+1)\xi(\mathbf{X})\mu(\mathbf{E}) + (10k^2 + 21k + 8)[\xi(X)\mu(\mathbf{E})]^2 > 0$$

and

$$1 - \xi(\mathbf{X})\mu(\mathbf{E})(4k+5) \geq 0$$

Since the first condition is guaranteed to be satisfied for $k \geq 1$, we have

$$\xi(\mathbf{X})\mu(\mathbf{E}) \leq \frac{1}{4k+5}.$$

Thus we finish the proof. □

Appendix D: Data Statistics

We listed the detailed domains of the sentiment analysis tasks in Table 3. We removed the *musical_instruments* and *tools_hardware* domains from the original data because they have too few labeled examples. The statistics for the 10 target tasks of intent classification in Table 4.

Domains	#train	#validation	#test
apparel	7398	926	928
automotive	601	69	66
baby	3405	437	414
beauty	2305	280	299
books	19913	2436	2489
camera_photo	5915	744	749
cell_phones_service	816	109	98
computer_video_games	2201	274	296
dvd	19961	2624	2412
electronics	18431	2304	2274
gourmet_food	1227	182	166
grocery	2101	268	263
health_personal_care	5826	687	712
jewelry_watches	1597	188	196
kitchen_housewares	15888	1978	1990
magazines	3341	427	421
music	20103	2463	2510
office_products	337	54	40
outdoor_living	1321	143	135
software	1934	254	202
sports_outdoors	4582	566	580
toys_games	10634	1267	1246
video	19941	2519	2539

Table 3: Statistics of the Multi-Domain Sentiment Classification Data.

Dataset ID	#labeled instances	#labels
1	497	11
2	3071	14
3	305	21
4	122	7
5	110	11
6	126	12
7	218	45
8	297	10
9	424	4
10	110	17

Table 4: Statistics of the User Intent Classification Data.