Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically (Supplementary Material)

Yuan Fang[† ‡], Kevin Chen-Chuan Chang[† ‡], and Hady W. Lauw[⋆]

[†]University of Illinois at Urbana-Champaign, USA {fang2,kcchang}@illinois.edu
[‡]Advanced Digital Sciences Center, Singapore
[⋆]Singapore Management University, Singapore hadywlauw@smu.edu.sg

This document contains three sections: 1) *Technical Proofs:* the detailed proofs for the propositions in the main paper; 2) *Additional Discussion:* a preliminary discussion on discriminative and generative formulations; 3) *Additional Experiments:* empirical evidence to support the theoretical analysis in the main paper.

1 Technical Proofs

The full proofs for all propositions can be found in this section. The propositions themselves and necessary equations are reproduced here for convenience. *All equations are re-numbered in this document*, which may have different numbering from the main paper. Unless explicitly stated, all equation references are self-contained in this document. In our notation, all unquantified indices such as i, j, k range from 1 to $|\mathcal{X}|$, unless stated explicitly.

1.1 Probability of Events

PROPOSITION 1 (PROBABILITY OF EVENTS): Suppose a random walk is visiting a sequence of points $\{V_t : t = 0, 1, \ldots\}$ on the graph defined by the following adjacency matrix:

$$W_{ij} = \begin{cases} \exp\left(-\|x_i - x_j\|^2/2\sigma^2\right) & i \neq j \\ 0 & i = j. \end{cases} \tag{1}$$

Define $Z_i \triangleq \sum_j W_{ij}$. The following conclusions can be made.

(a) The limit of $p\left(V_t, V_{t+1}\right)$ as $t \to \infty$ exists uniquely.

(b) Given that $(V_t, V_{t+1}) \xrightarrow{d} (X, X')$,

$$\forall ij, \quad p\left(X = x_i, X' = x_j\right) \propto W_{ij}, \tag{2}$$

$$\forall i, \quad p\left(X = x_i\right) = p\left(X, X' = x_i\right) \propto Z_i, \tag{3}$$

■

PROOF:

(a) First, $p\left(V_t, V_{t+1}\right) = p\left(V_{t+1}|V_t\right) p\left(V_t\right)$. On the one hand, since a random walk is a time-homogeneous Markov chain, $p\left(V_{t+1}|V_t\right)$ is constant as t varies. On the other hand, the limit of $p\left(V_t\right)$ as $t \to \infty$ is the (first-order) stationary distribution of the random walk, which exists uniquely for an arbitrary initial state V_0, subject to the irreducibility and aperodicity of the graph (Motwani & Raghavan, 2010). Thus, $\lim_{t\to\infty} p\left(V_t, V_{t+1}\right) = p\left(V_{t+1}|V_t\right) \lim_{t\to\infty} p\left(V_t\right)$, which also exists uniquely regardless of the initial state V_0.

Now we discuss the irreducibility and aperodicity of the graph. In particular, our graph (Eq. 1) satisfies such conditions. As a minor caveat, it is a popular practice to construct a kNN graph instead, where two points x_i and x_j are connected only if x_i is among the k nearest neighbors of x_j or x_j is among the k nearest neighbors of x_i. The two conditions are generally satisfied in a kNN graph as well; in the rare case where the conditions are not met, a simple and common tweak is to add some dummy edges of small weights (Haveliwala, 2003).

(b) Since $\left(V_t, V_{t+1}\right) \xrightarrow{d} \left(X, X'\right)$, we have

$$\begin{aligned} p\left(X = x_i, X' = x_j\right) &\overset{1}{=} \lim_{t\to\infty} p\left(V_t = x_i, V_{t+1} = x_j\right) \\ &\overset{2}{=} \lim_{t\to\infty} p\left(V_{t+1} = x_j | V_t = x_i\right) p\left(V_t = x_i\right) \\ &\overset{3}{=} \frac{W_{ij}}{Z_i} \lim_{t\to\infty} p\left(V_t = x_i\right) \\ &\overset{4}{=} \frac{W_{ij}}{Z_i} \frac{Z_i}{\sum_{uv} W_{uv}} \\ &\overset{5}{=} \frac{W_{ij}}{\sum_{uv} W_{uv}} \quad \overset{6}{\propto} W_{ij} \end{aligned} \tag{4}$$

In step 3, $p\left(V_{t+1} = x_j | V_t = x_i\right) = W_{ij}/Z_i$ is given by the transition probability. In step 4, $\lim_{t\to\infty} p\left(V_t = x_i\right) = Z_i / \sum_{uv} W_{uv}$ is the (first-order) stationary distribution of the random walk, which is established elsewhere (Motwani & Raghavan, 2010).

Next, we find $p\left(X = x_i\right)$ by marginalizing the joint distribution $p\left(X, X'\right)$.

$$p\left(X = x_i\right) = \sum_j p\left(X = x_i, X' = x_j\right) = \sum_j \frac{W_{ij}}{\sum_{uv} W_{uv}} = \frac{Z_i}{\sum_{uv} W_{uv}} \propto Z_i \tag{5}$$

Finally, $p\left(X,\, X' = x_i\right)$ can be found similarly. ■

1.2 Statistical Indistinguishability

DEFINITION 1 (INDISTINGUISHABILITY): Two distributions D_1 and D_2 are ϵ-*statistically indistinguishable* if and only if $\frac{1}{2} \left\| D_1 - D_2 \right\|_1 \leq \epsilon$. ■

PROPOSITION 2 (LABEL COUPLING): Suppose the label distribution of x_i, $p\left(Y|X = x_i\right)$, and the label distribution of some point close to x_i, $p\left(Y|X,\, X' = x_i\right)$, are related:

$$p\left(Y|X = x_i\right) = (1-\alpha) p\left(Y|X,\, X' = x_i\right) + \alpha D, \tag{6}$$

for some $\alpha \in (0, 1)$ and some distribution D. Then, the two distributions $p\left(Y|X = x_i\right)$ and $p\left(Y|X,\, X' = x_i\right)$ are α-*statistically indistinguishable* (Definition 1). ■

PROOF:

$$\begin{aligned} \frac{1}{2} \left\| p\left(Y|X = x_i\right) - p\left(Y|X,\, X' = x_i\right) \right\|_1 &= \frac{1}{2} \left\| (1-\alpha) p\left(Y|X,\, X' = x_i\right) + \alpha D - p\left(Y|X,\, X' = x_i\right) \right\|_1 \\ &= \frac{1}{2} \alpha \left\| D - p\left(Y|X,\, X' = x_i\right) \right\|_1 \\ &\leq \alpha \end{aligned} \tag{7}$$

The inequality follows since the L_1 difference of any distribution is at most 2. Thus, the proof is concluded. ■

1.3 Solution of Probability Constraints

PROPOSITION 3 (SOLUTION): $\forall y \in \mathcal{Y}$, suppose π_y satisfies the constraints in Eq. 8 and 9:

$$\textit{(Constraint on labeled data)} \qquad \begin{aligned} &\pi_{yi} = K \cdot \theta_{yi}, \quad \forall i : x_i \in \mathcal{L}, \\ &\text{where} \quad K = \sum_{k:x_k \in \mathcal{L}} \pi_{yk}, \\ &\text{and} \qquad \theta_{yi} = \frac{p\left(y|x_i\right) Z_i}{\sum_{k:x_k \in \mathcal{L}} p\left(y|x_k\right) Z_k}, \end{aligned} \tag{8}$$

$$\textit{(Constraint on unlabeled data)} \qquad \pi_{yi} = \sum_j \frac{(1-\alpha) W_{ji}}{Z_j} \pi_{yj}, \quad \forall i : x_i \notin \mathcal{L}. \tag{9}$$

Then, we can establish two conclusions below. (a) π_y is the stationary distribution of a Markov chain $\mathcal{C}$ with states $\mathcal{X}$ and transition matrix P, where

$$P_{ji} = \begin{cases} \dfrac{\sum_{k:x_k \in \mathcal{L}} W_{jk} + \alpha \sum_{k:x_k \notin \mathcal{L}} W_{jk}}{Z_j} \cdot \theta_{yi} & i : x_i \in \mathcal{L} \\ \dfrac{(1-\alpha) W_{ji}}{Z_j} & i : x_i \notin \mathcal{L}, \end{cases} \tag{10}$$

(b) The stationary distribution of $\mathcal{C}$ exists uniquely. ■

PROOF: (a) The objective is to derive the corresponding transition matrix P. Intuitively, the unlabeled constraint (Eq. 9) already tells us the transition from each state x_j to $x_i \notin \mathcal{L}$. Thus, our main task is to find out how to transit from each x_j to $x_i \in \mathcal{L}$, *i.e.*, to express K (Eq. 8) as a function of π_{yj}.

$$\begin{aligned} K &\overset{1}{=} \sum_{k:x_k \in \mathcal{L}} \pi_{yk} \\ &\overset{2}{=} 1 - \sum_{k:x_k \notin \mathcal{L}} \pi_{yk} \\ &\overset{3}{=} 1 - \sum_{k:x_k \notin \mathcal{L}} \left((1-\alpha) \sum_j \frac{W_{jk}}{Z_j} \pi_{yj} \right) \\ &\overset{4}{=} 1 - (1-\alpha) \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} \\ &\overset{5}{=} 1 - \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} + \alpha \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} \\ &\overset{6}{=} \sum_j \sum_k \frac{W_{jk}}{Z_j} \pi_{yj} - \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} + \alpha \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} \\ &\overset{7}{=} \sum_j \sum_{k:x_k \in \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} + \alpha \sum_j \sum_{k:x_k \notin \mathcal{L}} \frac{W_{jk}}{Z_j} \pi_{yj} \\ &\overset{8}{=} \sum_j \frac{\sum_{k:x_k \in \mathcal{L}} W_{jk} + \alpha \sum_{k:x_k \notin \mathcal{L}} W_{jk}}{Z_j} \pi_{yj} \end{aligned} \tag{11}$$

Step 2 follows since π_y is a distribution and its entries sum up to 1. Step 3 applies the constraint on unlabeled data

(Eq. 9). In step 6, $\sum_j \sum_k \frac{W_{jk}}{Z_j}\pi_{yj} = \sum_j \left(\pi_{yj} \sum_k \frac{W_{jk}}{Z_j}\right) = \sum_j \pi_{yj} = 1$. Thus, $\forall i : x_i \in \mathcal{L}$, we have

$$\pi_{yi} = K \cdot \theta_{yi} = \sum_j \frac{\sum_{k:x_k \in \mathcal{L}} W_{jk} + \alpha \sum_{k:x_k \notin \mathcal{L}} W_{jk}}{Z_j} \pi_{yj} \cdot \theta_{yi} \tag{12}$$

It is easy to verify that $\pi_y P = \pi_y$ and $\sum_i P_{ji} = 1$, concluding the first part of the proof.

(b) An irreducible and aperiodic Markov chain has a unique stationary distribution. P is irreducible and aperiodic except in one case below. When $\theta_{yi} = 0$ for some $i : x_i \in \mathcal{L}$, P is not irreducible since $P_{ji} = 0$. When this happens, it means $p\left(y|x_i\right) = 0$, which also implies $\pi_{yi} = 0$. Thus, we can do a minor tweak by excluding x_i from the state space $\mathcal{X}$, and derive a new transition matrix P' over $\mathcal{X} \backslash \{x_i\}$, which is always irreducible. The solution would then be π'_y corresponding to P', coupled with $\pi_{yi} = 0$. Alternatively, we can assume $p\left(y|x_i\right) \geq \epsilon, \forall x_i \in \mathcal{X}$, where ϵ is a very small positive constant. ■

1.4 Error Analysis

PROPOSITION 4 (ERROR): Given the two constraints in Eq. 8 and 9, for any constant $\epsilon \in (0, 1)$,

$$\mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1\right] \leq O\left(\left(1 - \lambda_1\right)^{|\mathcal{U}|}\right) + O\left(\exp\left(-2\epsilon^2 \lambda_2 \min_{x_i \in \mathcal{L}, p(y|x_i) > 0} |\mathcal{L}_{x_i}|\right)\right), \tag{13}$$

where $\lambda_1 = \min_{x_i \in \mathcal{X}, p(x_i) > 0} p\left(x_i\right)$, and $\lambda_2 = \min_{x_i \in \mathcal{L}, p(y|x_i) > 0} p\left(y|x_i\right)^2$ are constants in $(0, 1]$. ■

PROOF: Our solution estimator $\hat{\pi}_y$ can differ from the true π_y due to insufficient samples, which produce two types of error as follows. (a) *Data sampling error.* We only observe a potentially partial $\hat{\mathcal{X}}$ for $\mathcal{X}$, through samples in $\mathcal{L}$ and $\mathcal{U}$. Hence, the affinity matrix W would not be correctly constructed, resulting in a different transition matrix P. (Here we suppose that the graph construction function itself is perfect.) (b) *Label sampling error.* We estimate θ_y as $\hat{\theta}_y$ based on $\mathcal{L}$, which is potentially erroneous, also resulting in a different P.

Corresponding to the two types of error, we consider two scenarios $\hat{\mathcal{X}} \subset \mathcal{X}$ and $\hat{\mathcal{X}} = \mathcal{X}$. On the one hand, when $\hat{\mathcal{X}} \subset \mathcal{X}$, we do not need to consider the second type of error caused by $\hat{\theta}_y$, since regardless of the error in $\hat{\theta}_y$, the error in $\hat{\pi}_y$ can be as large as 2 (the maximal L_1 difference between two distributions). On the other hand, when $\hat{\mathcal{X}} = \mathcal{X}$, the first type of error does not exist, and we only need to investigate the error caused by $\hat{\theta}_y$. Formally,

$$\begin{aligned}\mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1\right] &= \mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1 \middle| \hat{\mathcal{X}} \subset \mathcal{X}\right] p\left(\hat{\mathcal{X}} \subset \mathcal{X}\right) + \mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1 \middle| \hat{\mathcal{X}} = \mathcal{X}\right] p\left(\hat{\mathcal{X}} = \mathcal{X}\right) \\ &\leq 2p\left(\hat{\mathcal{X}} \subset \mathcal{X}\right) + \mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1 \middle| \hat{\mathcal{X}} = \mathcal{X}\right] \end{aligned} \tag{14}$$

Now, we only need to bound $p\left(\hat{\mathcal{X}} \subset \mathcal{X}\right)$ and $\mathbb{E}\left[\|\hat{\pi}_y - \pi_y\|_1 \middle| \hat{\mathcal{X}} = \mathcal{X}\right]$ for the two scenarios, respectively. Let $\mathcal{L}_{x_i}$ (or $\mathcal{U}_{x_i}$) denote the set that contains all the samples with x_i in $\mathcal{L}$ (or $\mathcal{U}$). Similarly, let $\mathcal{L}_y$ denote the set that contains all the samples with y in $\mathcal{L}$. Note that since the samples are i.i.d, there can be different samples with x_i or y.

The first scenario $\hat{\mathcal{X}} \subset \mathcal{X}$ happens only when there exists some $x_i \in \mathcal{X}$, such that $p\left(x_i\right) > 0$ but it is not sampled in $\mathcal{L}$ or $\mathcal{U}$. Hence,

$$\begin{aligned} p\left(\hat{\mathcal{X}} \subset \mathcal{X}\right) &\overset{1}{=} p\left(\bigcup_{x_i \in \mathcal{X}, p(x_i)>0} |\mathcal{L}_{x_i}| + |\mathcal{U}_{x_i}| = 0\right) \\ &\overset{2}{\leq} \sum_{x_i \in \mathcal{X}, p(x_i)>0} p\left(|\mathcal{L}_{x_i}| + |\mathcal{U}_{x_i}| = 0\right) \\ &\overset{3}{=} \sum_{x_i \in \mathcal{X}, p(x_i)>0} \left(1 - p\left(x_i\right)\right)^{|\mathcal{L}|+|\mathcal{U}|} \\ &\overset{4}{=} O\left(\left(1 - \min_{x_i \in \mathcal{X}, p(x_i)>0} p\left(x_i\right)\right)^{|\mathcal{L}|+|\mathcal{U}|}\right). \end{aligned} \tag{15}$$

Step 2 is an application of Bonferroni inequalities. Step 3 follows since the samples in $\mathcal{L}$ and $\mathcal{U}$ are i.i.d. In step 4, $\min_{x_i \in \mathcal{X}, p(x_i)>0} p\left(x_i\right)$ is a constant in $(0, 1]$, which we denote as λ_1. Furthermore, if we consider $|\mathcal{U}| \gg |\mathcal{L}|$, we can rewrite Eq. 15 as follows:

$$p\left(\hat{\mathcal{X}} \subset \mathcal{X}\right) \leq O\left(\left(1 - \lambda_1\right)^{|\mathcal{U}|}\right). \tag{16}$$

In the second scenario $\hat{\mathcal{X}} = \mathcal{X}$, we investigate the second type of error due to $\hat{\theta}_y$. Note that θ_{yi} itself is defined in terms of $p\left(y|x_i\right)$ (Eq. 8). Thus, we need to quantify the error in $p\left(y|x_i\right)$, further translate it to the error in $\hat{\theta}_y$, and finally derive the error in $\hat{\pi}_y$.

We first bound the error $|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)|$. Note that we only need to consider x_i such that $p\left(y|x_i\right) > 0$, since when $p\left(y|x_i\right) = 0$, we have $|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)| = 0$ almost surely. Recall that $\hat{p}\left(y|x_i\right)$ is estimated as the sample mean $|\mathcal{L}_y \cap \mathcal{L}_{x_i}|/|\mathcal{L}_{x_i}|$. By Hoeffding's inequality, the error in the sample mean can be bounded. Specifically, for any constant $\epsilon \in (0, 1)$,

$$p\left(|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)| > p\left(y|x_i\right)\epsilon\right) \leq 2\exp\left(-2p\left(y|x_i\right)^2 \epsilon^2 |\mathcal{L}_{x_i}|\right). \tag{17}$$

To obtain a bound on the error in $\hat{\theta}_y$, we need $|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)|$ to be bounded for all $x_i \in \mathcal{L}$ in conjunction, whose probability can be computed as follows.

$$\begin{aligned} & p\left(\bigcap_{x_i \in \mathcal{L}, p(y|x_i)>0} |\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)| \leq p\left(y|x_i\right)\epsilon\right) \\ \overset{1}{=}\ & 1 - p\left(\bigcup_{x_i \in \mathcal{L}, p(y|x_i)>0} |\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)| > p\left(y|x_i\right)\epsilon\right) \\ \overset{2}{\geq}\ & 1 - \sum_{x_i \in \mathcal{L}, p(y|x_i)>0} p\left(|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)| > p\left(y|x_i\right)\epsilon\right) \\ \overset{3}{\geq}\ & 1 - 2\sum_{x_i \in \mathcal{L}, p(y|x_i)>0} \exp\left(-2p\left(y|x_i\right)^2 \epsilon^2 |\mathcal{L}_{x_i}|\right) \\ \overset{4}{=}\ & 1 - \rho \end{aligned} \tag{18}$$

Step 1 follows from De Morgan's laws. Step 2 is an application of Bonferroni inequalities. Step 3 follows from Eq. 17. In step 4 we denote $\rho \triangleq 2 \sum_{x_i \in \mathcal{L}, p(y|x_i)>0} \exp\left(-2p\left(y|x_i\right)^2 \epsilon^2 |\mathcal{L}_{x_i}|\right)$.

Next, we can translate the error $\left|\hat{p}\left(y|x_i\right) - p\left(y|x_i\right)\right|$ to $\left|\hat{\theta}_{yi} - \hat{\theta}_{yi}\right|$. Eq. 18 means that there is a probability of at least $1-\rho$ such that $\forall x_i \in \mathcal{L}, p\left(y|x_i\right) > 0$,

$$1-\epsilon \leq \frac{\hat{p}\left(y|x_i\right)}{p\left(y|x_i\right)} \leq 1+\epsilon. \tag{19}$$

Hence, with probability at least $1-\rho$, $\forall x_i \in \mathcal{L}, p\left(y|x_i\right) > 0$,

$$\frac{1-\epsilon}{1+\epsilon} \leq \frac{\hat{\theta}_{yi}}{\theta_{yi}} \leq \frac{1+\epsilon}{1-\epsilon} \quad \Rightarrow \quad \left|\hat{\theta}_{yi} - \theta_{yi}\right| \leq \frac{2\epsilon}{1-\epsilon}\theta_{yi} \tag{20}$$

Finally, we can translate the error $\left|\hat{\theta}_{yi} - \hat{\theta}_{yi}\right|$ to $\left\|\hat{\pi}_y - \pi_y\right\|_1$. Based on the perturbation theory of Markov chains (Cho & Meyer, 2001; Seneta, 1993), we can bound $\left\|\hat{\pi}_y - \pi_y\right\|_1$ in terms of the ∞-norm of the error matrix $\hat{P} - P$. Specifically, when $\hat{\mathcal{X}} = \mathcal{X}$, with probability at least $1-\rho$,

$$\begin{aligned}
\left\|\hat{\pi}_y - \pi_y\right\|_1 &\overset{1}{\leq} C\left|\hat{P} - P\right|_\infty \\
&\overset{2}{=} C \max_j \sum_i \left|\hat{P}_{ji} - P_{ji}\right| \\
&\overset{3}{=} C \max_j \sum_{i:x_i\in\mathcal{L}} \frac{\sum_k \alpha^{\mathbf{1}\{k:x_k\notin\mathcal{L}\}} W_{jk}}{Z_j} \left|\hat{\theta}_{yi} - \theta_{yi}\right| \\
&\overset{4}{\leq} C \max_j \sum_{i:x_i\in\mathcal{L}} \left|\hat{\theta}_{yi} - \theta_{yi}\right| \\
&\overset{5}{\leq} C \max_j \sum_{i:x_i\in\mathcal{L}} \frac{2\epsilon}{1-\epsilon}\theta_{yi} \\
&\overset{6}{=} \frac{2C\epsilon}{1-\epsilon}.
\end{aligned} \tag{21}$$

Step 1 follows from the perturbation theory of Markov chains, where C is called a condition number and is a constant for a given P. In step 3, we only sum over i where $x_i \in \mathcal{L}$, since $\hat{P}_{ji} = P_{ji}$ when $x_i \notin \mathcal{L}$. Step 4 follows since $Z_j = \sum_k W_{jk} \geq \sum_k \alpha^{\mathbf{1}\{k:x_k\notin\mathcal{L}\}} W_{jk}$ for $0 < \alpha < 1$. Step 5 applies Eq. 20.

As $\left\|\hat{\pi}_y - \pi_y\right\|_1$ is also bounded by 2 (the maximal L_1 difference between two distributions), when $\hat{\mathcal{X}} = \mathcal{X}$, we can determine with probability at least $1-\rho$,

$$\left\|\hat{\pi}_y - \pi_y\right\|_1 \leq \min\left\{2, \frac{2C\epsilon}{1-\epsilon}\right\}. \tag{22}$$

Let $\delta \triangleq \min\left\{2, \frac{2C\epsilon}{1-\epsilon}\right\}$, which is a constant. It is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\pi}_y - \pi_y\right\|_1 \middle| \hat{\mathcal{X}} = \mathcal{X}\right] &\leq \delta(1-\rho) + 2\rho \\
&= \delta + (2-\delta)\rho \\
&= O(\rho) \\
&= O\left(\exp\left(-2\epsilon^2 \min_{x_i\in\mathcal{L}, p(y|x_i)>0} p\left(y|x_i\right)^2 \min_{x_i\in\mathcal{L}} \left|\mathcal{L}_{x_i}\right|\right)\right).
\end{aligned} \tag{23}$$

Here $\min\limits_{x_i\in\mathcal{L}, p(y|x_i)>0} p\left(y|x_i\right)^2$ is a constant in $(0, 1]$, which we denote as λ_2.

By plugging Eq. 16 and 23 into Eq. 14, we conclude the proof. ■

1.5 Robustness Analysis

PROPOSITION 5 (ROBUSTNESS): Suppose a matrix $\tilde{W}$ is perturbed from W, such that for some $s > 1$, $W_{ij}/s \leq \tilde{W}_{ij} \leq W_{ij} \cdot s, \forall ij$. Let $\tilde{\pi}_y$ be the the solution vector based on $\tilde{W}$. It holds that $\|\tilde{\pi}_y - \pi_y\|_1 \leq O(s^2 - 1)$. ■

PROOF: Let $\tilde{P}$ be an estimator of the transition matrix from $\tilde{W}$. As $W_{ij}/s \leq \tilde{W}_{ij} \leq W_{ij} \cdot s$, we have

$$P_{ji}/s^2 \leq \tilde{P}_{ji} \leq P_{ji} \cdot s^2 \quad \Rightarrow \quad \left|\tilde{P}_{ji} - P_{ji}\right| \leq (s^2 - 1)P_{ji} \tag{24}$$

From the sensitivity theory of Markov chains (Cho & Meyer, 2001; Seneta, 1993), for a condition number C which is a constant for a given P,

$$\begin{aligned} \|\tilde{\pi}_y - \pi_y\|_1 &\leq C\left|\tilde{P} - P\right|_\infty & (25) \\ &= C \max_j \sum_i \left|\tilde{P}_{ji} - P_{ji}\right| \\ &\leq C \max_j \sum_i (s^2 - 1)P_{ji} \\ &= C(s^2 - 1). & (26) \end{aligned}$$

Hence, the proof is concluded. ■

2 Additional Discussion

In the main paper, we focus on the generative formulation of smoothness in the label coupling model, as follows. $\forall y \in \mathcal{Y}, \forall x_i \in \mathcal{X}$,

$$p\left(X = x_i | Y = y\right) = (1 - \alpha)p\left(X, X' = x_i | Y = y\right). \tag{27}$$

This result further enables us to derive a set of probability constraints in terms of $p\left(X|Y\right)$. For a given $y \in \mathcal{Y}$,

$$p\left(X = x_i | Y = y\right) = (1 - \alpha) \sum_j W_{ji}/Z_j \cdot p\left(X = x_j | Y = y\right). \tag{28}$$

Their derivations are laid out in the main paper. However, our smoothness framework can accommodate both generative and discriminative formulations. In particular, we are able to derive a discriminative counterpart, in terms of $p\left(Y|X\right)$. Starting from the indistinguishability-based label coupling model, $\forall y \in \mathcal{Y}, \forall x_i \in \mathcal{X}$,

$$\begin{aligned} p\left(Y | X = x_i\right) &= (1 - \alpha)p\left(Y | X, X' = x_i\right) + \alpha D, \\ &= (1 - \alpha)p\left(Y | X, X' = x_i\right), \qquad (\because D(y) = 0, \forall y \in \mathcal{Y}) & (29) \end{aligned}$$

which is already in the discriminative form, we can derive the corresponding set of probability constraints in terms of $p\,(Y|X)$. For a given $y \in \mathcal{Y}$,

$$\begin{aligned}
p\,(Y=y|X=x_i) &\overset{1}{=} (1-\alpha)p\,(Y=y|X,\,X'=x_i)\\
&\overset{2}{=} (1-\alpha)\sum_j p\,(Y=y,X=x_j|X,\,X'=x_i)\\
&\overset{3}{=} (1-\alpha)\sum_j p\,(Y=y|X=x_j,X'=x_i)\,p\,(X=x_j|X'=x_i)\\
&\overset{4}{=} (1-\alpha)\sum_j p\,(Y=y|X=x_j)\,p\,(X=x_j|X'=x_i)\\
&\overset{5}{=} (1-\alpha)\sum_j p\,(Y=y|X=x_j)\,W_{ji}/Z_i
\end{aligned} \tag{30}$$

Step 1 is given by the indistinguishability model. Step 2 expands into the joint distribution. Step 3 is an application of the Bayes' rule. In step 4, the label Y of X, given $X=x_j$, is conditionally independent of X'. In step 5, we have $p\,(X=x_j|X'=x_i) = p\,(X=x_j,X'=x_i)\,/p\,(X'=x_i) = W_{ji}/Z_i$. Thus, $p\,(Y=y|X=x_i)$ is rewritten, relating the label distribution of x_i to that of its close points.

The generative and discriminative constraints in Eq. 28 and 30 appear symmetric. They differ in their normalizations —the generative constraints (Eq. 28) normalize each summand by Z_j, while the discriminative constraints (Eq. 30) normalize each summand by Z_i. Interestingly, such different normalizations correspond to two different but symmetric forms of random walk, namely the forward and backward random walks (Agarwal et al., 2010; Fang & Chang, 2011; Fang et al., 2013). Loosely speaking, on the one hand, the generative constraints correspond to the forward walk, which starts from a labeled point and we are interested in the probability of reaching each x_i. On the other hand, the discriminative constraints correspond to the backward walk, which starts from x_i and we are interested in the probability of reaching a labeled point. These two forms of random walk not only travel in "opposite" directions, but also convey symmetric and complementary semantics: probabilistic recall and precision (Agarwal et al., 2010; Fang & Chang, 2011) respectively, or importance and specificity (Fang et al., 2013) respectively. We refer readers to the given literature for more details.

The discriminative constraints are also similar to the harmonic functions in the GRF method (Zhu et al., 2003) as shown below (which only lacks the $1-\alpha$ factor), potentially giving a new interpretation to GRF.

$$F_i = \textstyle\sum_j W_{ij}/Z_i \cdot F_j, \tag{31}$$

While the given labels naturally give a set of constraints on the labeled data points, the discriminative constraints can be applied on the unlabeled data. By solving these constraints on labeled and unlabeled data, we can ultimately find $p\,(Y|X)$. We leave the development of a concrete solution based on the discriminative constraints to future work.

3 Additional Experiments

We conduct additional experiments to validate the theoretical analysis in the main paper.

3.1 Effect of unlabeled data

While we have seen the effect of increasing labeled points in the main paper, we investigate unlabeled points here to validate the error analysis in Proposition 4. For each dataset, we sample $|\mathcal{L}| + |\mathcal{U}|$ data points, where we fix $|\mathcal{L}| = 10$

while $|\mathcal{U}|$ is varied from Δ to 5Δ. Δ is different for different datasets in order to visualize the performance trend clearer. Among these points, we draw 10 points as labeled data $\mathcal{L}$, and treat the remaining points as unlabeled data $\mathcal{U}$.

We present the mean performance over 10 such samples in Fig. 1. The empirical results are consistent with our analysis. As expected, the performance improves when we use more unlabeled data even though the labeled data remain the same. The result also shows that the rate of improvement slows down as we introduce more unlabeled data, eventually hitting a ceiling. This means growing unlabeled data alone cannot reduce the error indefinitely, since it is also limited by labeled data as explained by the second error term in Proposition 4. In more intuitive words, unlabeled data can only help so much.

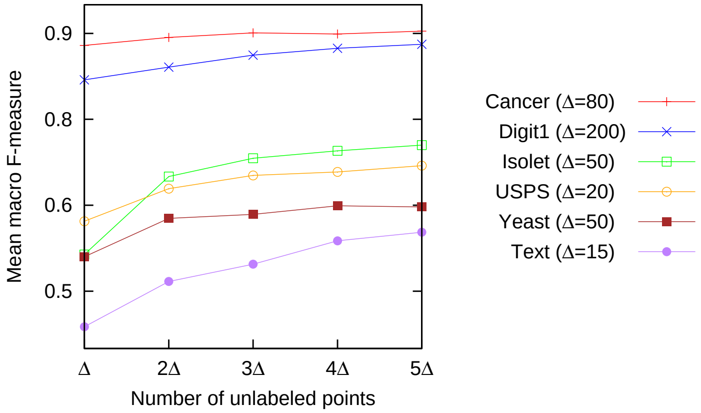

Figure 1: Effect of unlabeled data.

3.2 Effect of graph perturbation

Lastly, we study the robustness of our solution against graph perturbation. As Proposition 5 has established, the change in the solution π_y can be bounded by $O(s^2 - 1)$, where $s \geq 1$ is the degree of perturbation. Given s, we independently generate a random $\tilde{W}_{ij} \in [W_{ij}/s, W_{ij} \cdot s]$, $\forall ij$, and use $\tilde{W}$ as the perturbed affinity matrix.

We illustrate the effect of perturbation on USPS with $|\mathcal{L}| = 10$. In Fig. 2, when the graph is perturbed to various degrees ($s \in \{1.01, 1.02, \ldots, 1.1\}$), the change in π_y (averaged over the testing runs with standard deviation bars) for both $y = 1$ and $y = 2$ is linear in s. This result is better than (yet still consistent with) the theoretical bound $O(s^2 - 1)$. The reason is that we consider the worst case scenario in deriving the bound, but in our experiments here the perturbations are generated randomly within the degree s. Furthermore, we observe that the change in the actual predictive power in terms of macro F-measure is highly correlated with the change in π_y. The same trend is observed on all datasets with different $|\mathcal{L}|$'s. Thus, our solution is robust—small degrees of perturbation cause small changes in π_y and the predictive power.

References

Agarwal, Ganesh, Kabra, Govind, and Chang, Kevin Chen-Chuan. Towards rich query interpretation: walking back and forth for mining query templates. In *WWW*, pp. 1–10, 2010.

Cho, Grace E and Meyer, Carl D. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1):137–150, 2001.

Fang, Yuan and Chang, Kevin Chen-Chuan. Searching patterns for relation extraction over the web: rediscovering the pattern-relation duality. In *WSDM*, pp. 825–834, 2011.

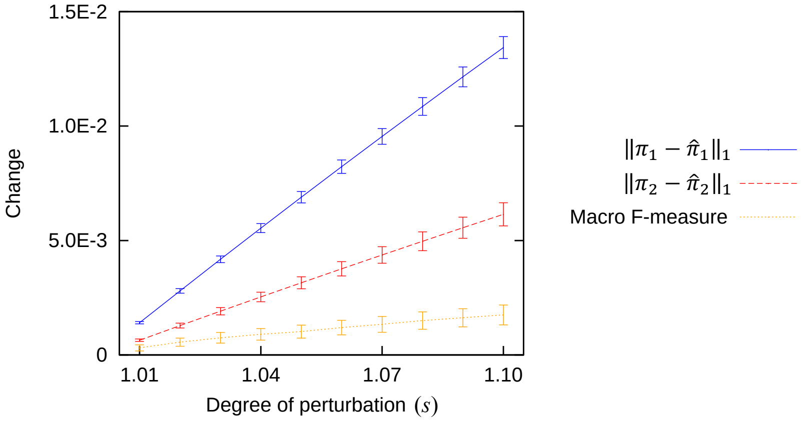

Figure 2: Effect of graph perturbation.

Fang, Yuan, Chang, Kevin Chen-Chuan, and Lauw, Hady Wirawan. Roundtriprank: Graph-based proximity with importance and specificity? In *ICDE*, pp. 613–624, 2013.

Haveliwala, Taher H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

Motwani, R. and Raghavan, P. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

Seneta, E. Sensitivity of finite markov chains under perturbation. *Statistics & probability letters*, 17(2):163–168, 1993.

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, pp. 912–919, 2003.