Supplementary Material

A Derivation of G and H

In this section, we derive the gradient and second order gradient given in Eq (5). The result is the same as standard CRF, we include the derivation here for completeness of the paper. We can write the negative log-likelihood

$$l(\mathbf{y}, \mathbf{x}, \phi) = -\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}) + \ln Z(\mathbf{x}) \tag{19}$$

Here $Z(\mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}'))$. The following equality holds for $Z(\mathbf{x})$

$$\begin{aligned}\partial_{\phi_k} Z(\mathbf{x}) &= \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}')) \\ &= Z(\mathbf{x}) \sum_{\mathbf{y}' \in \mathcal{Y}} \frac{\exp(\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}'))\mu_k(\mathbf{y}')}{Z(\mathbf{x})} \\ &= Z(\mathbf{x})E[\mu_k]\end{aligned} \tag{20}$$

In the calculation, ϕ is viewed as a vector, and partial derivative is defined by the derivative at $\phi(\mathbf{y}, \mathbf{x})$. Using this property, we can calculate the gradient as

$$\begin{aligned}\mathbf{G}_i(\mathbf{x}) &\triangleq \partial_{\phi_i} l(\mathbf{y}, \mathbf{x}, \phi) \\ &= -\mu_i(\mathbf{y}) + \frac{\partial_{\phi_i} Z(\mathbf{x})}{Z(\mathbf{x})} \\ &= -\mu_i(\mathbf{y}) + E[\mu_i] = p_i - \mu_i(\mathbf{y})\end{aligned} \tag{21}$$

Here the last equality holds because $\mu_i(\mathbf{y}) \in \{0, 1\}$. We can further calculate the second order gradient as

$$\begin{aligned}\mathbf{H}_{ij}(\mathbf{x}) &\triangleq \partial_{\phi_i \phi_j} l(\mathbf{y}, \mathbf{x}, \phi) \\ &= \partial_{\phi_j} \mathbf{G}_i(\mathbf{x}) \\ &= \sum_{\mathbf{y}' \in \mathcal{Y}} \frac{\exp(\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}'))\mu_i(\mathbf{y}')\mu_j(\mathbf{y}')}{Z(\mathbf{x})} \\ &\quad - \sum_{\mathbf{y}' \in \mathcal{Y}} \frac{\exp(\sum_{i=1}^{m} \phi_i(\mathbf{x})\mu_i(\mathbf{y}'))\mu_i(\mathbf{y}')}{Z^2(\mathbf{x})} \partial_{\phi_j} Z(\mathbf{x}) \\ &= E[\mu_i\mu_j] - E[\mu_i]E[\mu_j] = p_{ij} - p_i p_j\end{aligned} \tag{22}$$

The hessian $\mathbf{H}$ is also known as Fisher information matrix.

B Proof for Lemma 2.1

Proof. The following inequality holds for γ that satisfies the condition

$$\begin{aligned}\sum_{i \in \mathcal{U}} \gamma_i \mathbf{H}_{ii} \delta_i^2 &\geq \textstyle\sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} |\mathbf{H}_{ij}| \delta_i^2 \\ &= \tfrac{1}{2} \textstyle\sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} |\mathbf{H}_{ij}| (\delta_i^2 + \delta_j^2) \\ &\geq \textstyle\sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} |\mathbf{H}_{ij}| \delta_i \delta_j\end{aligned}$$

Applying it to Talyor expansion in Eq (4), we have

$$\begin{aligned}l(\mathbf{y}, \mathbf{x}, \phi + \delta) &= l(\mathbf{y}, \mathbf{x}, \phi) + \sum_{i \in \mathcal{U}} \delta_i \mathbf{G}_i(\mathbf{y}, \mathbf{x}) \\ &\quad + \frac{1}{2} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} |\mathbf{H}_{ij}| \delta_i \delta_j + o(\delta^2) \\ &\leq l(\mathbf{y}, \mathbf{x}, \phi) + \sum_{i \in \mathcal{U}} \delta_i \mathbf{G}_i(\mathbf{y}, \mathbf{x}) \\ &\quad + \frac{1}{2} \sum_{i \in \mathcal{U}} \gamma_i \mathbf{H}_{ii} \delta_i^2 + o(\delta^2).\end{aligned}$$

□

C Proofs for Lemma 3.1 and 3.2

Proof. The proof is exactly the same for both node and potential case, we present the proof for $\mathcal{U}$ to be all node potentials here. Recall the definition of $\mathbf{H}$: $\mathbf{H}_{ij} = p_{ij}$. Note that p_i and p_{ij} are short hand notations for $p_i \triangleq P(\mu_i = 1|\mathbf{x})$, $p_{ij} \triangleq P(\mu_i\mu_j = 1|\mathbf{x})$, we have

$$\begin{aligned}\frac{1}{2p_i} \sum_{j \in \mathcal{U}} |\mathbf{H}_{ij}| &= \sum_j |p_{ij}/p_i - p_j| \\ &= \sum_{j \in \mathcal{U}} |P(\mu_j = 1|\mu_i = 1, \mathbf{x}) - P(\mu_j = 1|\mathbf{x})| \\ &= \sum_{s,k'} |P(y_s = k'|y_t = k, \mathbf{x}) - P(y_s = k'|\mathbf{x})| \\ &= \sum_s \|P(y_s|\mathbf{x}, y_t = k) - P(y_s|\mathbf{x})\|_{tv}\end{aligned}$$

□

D Proof for Lemma 3.3

Proof. Taking the fact that μ_i and μ_j are mutually exclusive for $j \neq i$, we have

$$\begin{aligned}&\sum_{j \in M} |P(\mu_j = 1|\mu_i = 1, \mathbf{x}) - P(\mu_j = 1|\mathbf{x})| \\ =&|P(\mu_i = 1|\mu_i = 1, \mathbf{x}) - P(\mu_i = 1|\mathbf{x})| \\ &+ \sum_{j \neq i} |P(\mu_j = 1|\mu_i = 1, \mathbf{x}) - P(\mu_j = 1|\mathbf{x})| \\ =&|1 - P(\mu_i = 1|\mathbf{x})| + \sum_{j \neq i} |0 - P(\mu_j = 1|\mathbf{x})| \\ =&(1 - P(\mu_i = 1|\mathbf{x})) + \sum_{j \neq i} P(\mu_j = 1|\mathbf{x}) \\ =&2(1 - P(\mu_i = 1|\mathbf{x}))\end{aligned}$$

□

E Proof for Theorem 4.2

Proof. In this proof, we will reduce the total variation distance between joint distribution of edge states into total variation distance of marginal distribution over nodes, as in Theorem 4.1. Assume the edge pairs are (y_t, y_{t+1}),

(y_s, y_{s+1}), and y_s is closer to y_{t+1} (without loss of generality), then

$$P(y_s, y_{s+1}|y_t, y_{t+1}, \mathbf{x}) = P(y_{s+1}|y_s, \mathbf{x})P(y_s|y_{t+1}, \mathbf{x})$$

We can convert total variation by

$$\begin{aligned}
&\|P(y_s, y_{s+1}|y_t, y_{t+1}, \mathbf{x}) - P(y_s, y_{s+1}|\mathbf{x})\|_{tv} \\
=&\sum_{y_s, y_{s+1}} |P(y_s, y_{s+1}|y_t, y_{t+1}, \mathbf{x}) - P(y_s, y_{s+1}|\mathbf{x})| \\
=&\sum_{y_s, y_{s+1}} P(y_{s+1}|y_s, \mathbf{x})|P(y_s|y_{t+1}, \mathbf{x}) - P(y_s|\mathbf{x})| \\
=&\sum_{y_s} |P(y_s|y_{t+1}, \mathbf{x}) - P(y_s|\mathbf{x})| \\
=&\|P(y_s|y_{t+1}, \mathbf{x}) - P(y_s|\mathbf{x})\|_{tv}
\end{aligned}$$

Now the case become same as node potential, we can make use of Corollary 3.1 bound the total variation.

$$\begin{aligned}
&\|P(y_s, y_{s+1}|y_t = k_t, y_{t+1} = k_{t+1}, \mathbf{x}) - P(y_s, y_{s+1}|\mathbf{x})\|_{tv} \\
=&\|P(y_s|y_{t+1} = k_{t+1}, \mathbf{x}) - P(y_s|\mathbf{x})\|_{tv} \\
\leq&\|P(y_{t+1}|y_{t+1} = k_{t+1}, \mathbf{x}) - P(y_{t+1}|\mathbf{x})\|_{tv} \prod_{(a,b)\in\mathcal{Q}(s,t+1)} \alpha_{b,a} \\
=&[1 - P(y_{t+1} = k_{t+1}|\mathbf{x})] \prod_{(a,b)\in\mathcal{Q}(s,t+1)} \alpha_{b,a} \\
\leq&[1 - P(y_t = k_t, y_{t+1} = k_{t+1}|\mathbf{x})] \prod_{(a,b)\in\mathcal{Q}(s,t+1)} \alpha_{b,a}
\end{aligned}$$

Here the first inequality is due to Corollary 3.1. Intuitively, this means that the total variational distance of between two edge states, can be bounded by recursively applying the mixing rate bound along the path between two edges. Summing the results of $(s, s+1)$ over all edges will give us Eq. (15). □