Supplementary material for "More data means less inference: A pseudo-max approach to structured learning"

David Sontag Microsoft Research
Ofer Meshi Hebrew University
Tommi Jaakkola CSAIL, MIT
Amir Globerson Hebrew University

1 Canonical Form

Every function of the form $f(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{ij\in E}\theta_{ij}(y_i,y_j) + \sum_i \theta_i(y_i)$, where the variables $y_i \in \{1,2,\ldots,k\}$ has an equivalent function in *canonical form*. We have already made use of the canonical form for binary pairwise MRFs (i.e., when $k=2$), and in this section we describe the generalization of this to non-binary MRFs. This notion is defined with respect to a *canonical assignment*.

Definition 1.1. *Parameters $\boldsymbol{\theta}^{can}$ are in* canonical form *with respect to the assignment $\boldsymbol{y}^{can}$ if:*

- *For all $ij \in E$ it holds that: $\theta_{ij}^{can}(y_i^{can}, y_j) = 0$ for all y_j and $\theta_{ij}^{can}(y_i, y_j^{can}) = 0$ for all y_i.*
- *For all $i \in V$ it holds that: $\theta_i^{can}(y_i^{can}) = 0$.*

Given an assignment $\boldsymbol{y}^{\text{can}}$ and parameters $\boldsymbol{\theta}$, the corresponding canonical parameters $\boldsymbol{\theta}^{\text{can}}$ can be easily obtained via a sequence of reparameterizations as follows:

1. Initialize $\boldsymbol{\theta}^{\text{can}} \leftarrow \boldsymbol{\theta}$.
2. For all $ij \in E, \forall y_i, y_j$, do
$$\theta_{ij}^{\text{can}}(y_i,y_j) \leftarrow \theta_{ij}^{\text{can}}(y_i,y_j) - \theta_{ij}^{\text{can}}(y_i,y_j^{\text{can}}) \quad (1)$$
$$\theta_i^{\text{can}}(y_i) \leftarrow \theta_i^{\text{can}}(y_i) + \theta_{ij}^{\text{can}}(y_i,y_j^{\text{can}}). \quad (2)$$
3. For all $ij \in E, \forall y_i, y_j$, do
$$\theta_{ij}^{\text{can}}(y_i,y_j) \leftarrow \theta_{ij}^{\text{can}}(y_i,y_j) - \theta_{ij}^{\text{can}}(y_i^{\text{can}},y_j) \quad (3)$$
$$\theta_j^{\text{can}}(y_j) \leftarrow \theta_j^{\text{can}}(y_j) + \theta_{ij}^{\text{can}}(y_i^{\text{can}},y_j). \quad (4)$$
4. For all $i, \forall y_i$, do
$$\theta_i^{\text{can}}(y_i) \leftarrow \theta_i^{\text{can}}(y_i) - \theta_i^{\text{can}}(y_i^{\text{can}}). \quad (5)$$

An example of this transformation into canonical form is shown in Fig. 1. The first three steps of the transformation correspond to reparameterizations of the model, while the last step only adds a constant to each single node potential. We thus have that $f(\boldsymbol{y};\boldsymbol{\theta}) = f(\boldsymbol{y};\boldsymbol{\theta}^{\text{can}}) + C$ which implies that $\text{argmax}_{\boldsymbol{y}} f(\boldsymbol{y};\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{y}} f(\boldsymbol{y};\boldsymbol{\theta}^{\text{can}})$. We conclude that every function $f(\boldsymbol{y};\boldsymbol{\theta})$ has an equivalent canonical form, and focus on this case in the following proposition. We denote by Θ^{can} the set of all parameters in canonical form.

2 Identifiability of True Parameters

Proposition 2.1. *For any $\boldsymbol{\theta}^* \in \Theta^{can}$, there is a set of $2|V|(k-1) + 2|E|(k-1)^2$ examples, $\{\boldsymbol{x}^m, \boldsymbol{y}(\boldsymbol{x}^m;\boldsymbol{\theta}^*)\}$, such that any pseudo-max consistent $\boldsymbol{\theta} \in \Theta_{ps}(\{\boldsymbol{y}^m, \boldsymbol{x}^m\}) \cap \Theta^{can}$ is arbitrarily close to $\boldsymbol{\theta}^*$.*

$\theta_{ij}(y_i, y_j)$

y_i \ y_j	1	2	3
1	0	1	$-n^2$
2	1	0	$-n^2$
3	$-n^2$	$-n^2$	0

$\theta^{\text{can}}_{ij}(y_i, y_j)$

y_i \ y_j	1	2	3
1	$2n^2$	$1+2n^2$	0
2	$1+2n^2$	$2n^2$	0
3	0	0	0

$\theta^{\text{can}}_{i}(y_i)$

$-\lvert N(i)\rvert n^2$
$-\lvert N(i)\rvert n^2$
0

Figure 1: Illustration of the transformation from a set of parameters to the equivalent canonical form, using the canonical assignment $\boldsymbol{y}^{\text{can}} = \mathbf{3}$. Shown on the left are the original edge potential functions (zero fields), and shown on the right is its transformation into canonical form. These edge potentials are also used in the NP-hardness proof.

Proof. The output of the classifier is $\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{y}} f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta})$ where the function f is given by (see Section 2.1 in the paper):

$$f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) = \textstyle\sum_{ij \in E} \theta_{ij}(y_i, y_j) + \sum_i \theta_i(y_i) + \sum_i x_i(y_i), \tag{6}$$

We assume that the data is generated from some function $f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}^*)$, with corresponding parameters $\boldsymbol{\theta}^*$ in canonical form. Without loss of generality we assume that $\boldsymbol{\theta}^*$ is canonical with respect to the assignment $\boldsymbol{y}^{\text{can}} = \mathbf{1}$ (i.e., $y_1^{\text{can}} = 1, y_2^{\text{can}} = 1, \ldots, y_n^{\text{can}} = 1$).

For this special case we can omit the loss $e(y_i^m, y_i)$, because the input $x_i(y_i)$ will suffice to rule out solutions such as $\boldsymbol{\theta} = 0$. The pseudo constraints then simplify to

$$\Theta_{ps} = \Bigg\{ \boldsymbol{\theta} \mid \forall m, i, y_i \neq y_i^m, \quad \sum_{j \in N(i)} \theta_{ij}(y_i^m, y_j^m) + \theta_i(y_i^m) + x_i^m(y_i^m) \geq \sum_{j \in N(i)} \theta_{ij}(y_i, y_j^m) + \theta_i(y_i) + x_i^m(y_i) \Bigg\} \tag{7}$$

We now show how to construct a set of $2n(k-1) + 2|E|(k-1)^2$ labeled examples $\{(\boldsymbol{x}^m, \boldsymbol{y}^m)\}_{m=1}^M$ such that $\boldsymbol{\theta}^* \in \Theta_{ps} \cap \Theta^{\text{can}}$ (i.e., it is non-empty), and all other $\boldsymbol{\theta} \in \Theta_{ps} \cap \Theta^{\text{can}}$ are close to $\boldsymbol{\theta}^*$. For convenience, we define $\text{Max}_i(\boldsymbol{\theta}^*) = \max_{y_i} \left[|\theta_i^*(y_i)| + \sum_{j \in N(i)} \max_{y_j} |\theta_{ij}^*(y_i, y_j)| \right]$. The key idea is to set coordinates of $\boldsymbol{x}^m$ to large enough values such that either y_i^m or its neighbors are set to their canonical state.

The first set of examples enforces that $\theta_i^*(y_i) - \epsilon/2 \leq \theta_i(y_i) \leq \theta_i^*(y_i) + \epsilon/2$. In particular, for every $i \in V$ and every label $y_i \neq 1$ (the canonical state for i), we have two examples (unless otherwise specified, assume $\boldsymbol{x}^m = \mathbf{0}$). For both examples, and for all $j \in N(i)$, we set $y_j^m = 1$ and $x_j^m(1) = \text{Max}_j(\boldsymbol{\theta}^*) + 1$. This enforces that all nodes that are neighbors of i are in their canonical state (and thus y_i is effectively separated from the rest of the graph). In addition, for both examples and for $\hat{y_i} \notin \{1, y_i\}$, we set $x_i^m(\hat{y_i}) = -2\text{Max}_i(\boldsymbol{\theta}^*) - 1$. This will ensure that the states $\hat{y_i} \notin \{1, y_i\}$ will not be in the maximizing assignment. Then,

1. For the first example, set $y_i^m = y_i$ and $x_i^m(y_i) = -\theta_i^*(y_i) + \epsilon/2$. This gives us the following *pseudo-max* constraint for variable y_i:

$$\sum_{j \in N(i)} \theta_{ij}(y_i, 1) + \theta_i(y_i) + x_i^m(y_i) \quad \geq \quad \sum_{j \in N(i)} \theta_{ij}(1, 1) + \theta_i(1) + x_i^m(1) \tag{8}$$

$$\theta_i(y_i) - \theta_i^*(y_i) + \epsilon/2 \quad \geq \quad 0 \tag{9}$$

$$\theta_i(y_i) \quad \geq \quad \theta_i^*(y_i) - \epsilon/2. \tag{10}$$

2. For the second example, set $y_i^m = 1$ and $x_i^m(y_i) = -\theta_i^*(y_i) - \epsilon/2$. This gives us the following *pseudo-max* constraint for variable y_i:

$$\begin{aligned}
\sum_{j \in N(i)} \theta_{ij}(1,1) + \theta_i(1) + x_i^m(1) &\geq \sum_{j \in N(i)} \theta_{ij}(y_i,1) + \theta_i(y_i) + x_i^m(y_i) && (11)\\
0 &\geq \theta_i(y_i) - \theta_i^*(y_i) - \epsilon/2 && (12)\\
\theta_i^*(y_i) + \epsilon/2 &\geq \theta_i(y_i). && (13)
\end{aligned}$$

We have thus far not specified y_k^m for $k \notin \{i\} \cup N(i)$. These values should be set such that $\boldsymbol{y}^m = \boldsymbol{y}(\boldsymbol{x}^m; \boldsymbol{\theta}^*)$, so that $\boldsymbol{\theta}^* \in \Theta_{ps}$.

The second set of examples enforces that $\theta_{ij}^*(y_i, y_j) - \epsilon \leq \theta_{ij}(y_i, y_j) \leq \theta_{ij}^*(y_i, y_j) + \epsilon$. In particular, for every $ij \in E$ and $y_i \neq 1, y_j \neq 1$, we have two examples. For both examples, and for all $k \in N(i) \backslash \{j\}$, we set $y_k^m = 1$ and $x_k^m(1) = \text{Max}_k(\boldsymbol{\theta}^*) + 1$. This enforces that all nodes that are neighbors of i (except j) are in their canonical state. As before, for both examples and for $\hat{y_i} \notin \{1, y_i\}$, we set $x_i^m(\hat{y_i}) = -2\text{Max}_i(\boldsymbol{\theta}^*) - 1$. We also set $y_j^m = y_j$ and $x_j^m(y_j) = \text{Max}_j(\boldsymbol{\theta}^*) + 1$.

1. For the first example, set $y_i^m = y_i$ and $x_i^m(y_i) = -\theta_{ij}^*(y_i, y_j) - \theta_i^*(y_i) + \epsilon/2$. This gives us the following *pseudo-max* constraint for variable y_i:

$$\begin{aligned}
\theta_{ij}(y_i, y_j) + \sum_{k \in N(i) \backslash \{j\}} \theta_{ik}(y_i, 1) + \theta_i(y_i) + x_i^m(y_i) &\geq \theta_{ij}(1, y_j) + \sum_{k \in N(i) \backslash \{j\}} \theta_{ik}(1,1) + \theta_i(1) + x_i^m(1)\\
\theta_{ij}(y_i, y_j) + \theta_i(y_i) - \theta_{ij}^*(y_i, y_j) - \theta_i^*(y_i) + \frac{\epsilon}{2} &\geq 0\\
\theta_{ij}(y_i, y_j) &\geq \theta_{ij}^*(y_i, y_j) + \Big(\theta_i^*(y_i) - \theta_i(y_i)\Big) - \epsilon/2 \Rightarrow\\
\theta_{ij}(y_i, y_j) &\geq \theta_{ij}^*(y_i, y_j) - \epsilon.
\end{aligned}$$

2. For the second example, set $y_i^m = 1$ and $x_i^m(y_i) = -\theta_{ij}^*(y_i, y_j) - \theta_i^*(y_i) - \epsilon/2$. This gives us the following *psuedo-max* constraint for variable y_i:

$$\begin{aligned}
\theta_{ij}(1, y_j) + \sum_{k \in N(i) \backslash \{j\}} \theta_{ik}(1,1) + \theta_i(1) + x_i^m(1) &\geq \theta_{ij}(y_i, y_j) + \sum_{k \in N(i) \backslash \{j\}} \theta_{ik}(y_i, 1) + \theta_i(y_i) + x_i^m(y_i)\\
0 &\geq \theta_{ij}(y_i, y_j) + \theta_i(y_i) - \theta_{ij}^*(y_i, y_j) - \theta_i^*(y_i) - \frac{\epsilon}{2}\\
\theta_{ij}^*(y_i, y_j) + \Big(\theta_i^*(y_i) - \theta_i(y_i)\Big) + \epsilon/2 &\geq \theta_{ij}(y_i, y_j) \Rightarrow\\
\theta_{ij}^*(y_i, y_j) + \epsilon &\geq \theta_{ij}(y_i, y_j).
\end{aligned}$$

As before, y_k^m for $k \notin \{i\} \cup N(i)$ should be set such that $\boldsymbol{y}^m = \boldsymbol{y}(\boldsymbol{x}^m; \boldsymbol{\theta}^*)$, so that $\boldsymbol{\theta}^* \in \Theta_{ps}$. We now let $\epsilon \to 0$ to get the result. □