A Note on Minimax Learning of Tree Models

Frederic Koehler

June 24, 2020

In recent work, Devroye et al [1] studied minimax rates for learning various kinds of graphical models in Total Variation (TV) distance. They proved that for a tree Ising model on n nodes, the minimax rate for reconstructing the tree model in TV from m samples is upper bounded by $O\left(\sqrt{\frac{n\log(n)}{m}}\right)$ and posed the tightness of this result as an open question. In this note, we resolve this open question by proving a matching information-theoretic lower bound, showing that the minimax rate is $\Theta\left(\sqrt{\frac{n\log n}{m}}\right)$ up to a universal constant.

1 Proof of the Lower Bound

Recall from Stirling's formula that $\log(n!) \sim n\log(n)$. This motivates the following simple construction for the lower bound:

1. $\mathcal{S}$ is a family of permutations on $[n]$ to be specified later.
2. Pick a permutation π from family of permutations $\mathcal{S}$.
3. Build an Ising model on the matching graph with covariance $\alpha/\sqrt{n}$ between vertices i and $n+\pi(i)$ for every $i \in [n]$. (This corresponds to edge weight essentially $\alpha/\sqrt{n}$ since the correlation for edge weight β is $\tanh(\beta) \approx \beta$)

This yields a family of distributions $\{P_\pi\}_{\pi\in\mathcal{S}}$ on $2n$ nodes.

For the set of permutations, we choose a set which satisfies $\log|\mathcal{S}| = \Omega(n\log(n))$ and such that every set of permutations has Hamming distance at least $n/4$, where the Hamming distance is $|\{x : \pi(x) \neq \pi'(x)\}|$. In the survey of Quistorff [3], such a result is given as (6) and attributed to Deza.

The proof reduces to the following claims:

1. Every two elements have large total variation distance. More specifically, to distinguish models from π_1, π_2 look at statistic $\sum_{i=1}^n X_i X_{\pi_1(i)}$. Under P_{π_1} the expectation is $\alpha\sqrt{n}$ and variance is $\Theta(n)$, whereas under P_{π_2} the expectation is less than $(3/4)\alpha\sqrt{n}$ and variance is $\Theta(n)$. Applying the Central Limit Theorem, this shows the total variation distance is $\Omega(\alpha)$.

2. Every two elements have reasonably small KL. Recall by the Gibbs variational formula [2] that

$$KL(P_{\pi_1}, P_{\pi_2}) = (\frac{1}{2}\,\mathrm{E}_{\pi_2}[X^T J_{\pi_2} X] + H_{\pi_2}(X)) - (\frac{1}{2}\,\mathrm{E}_{\pi_1}[X^T J_{\pi_2} X] + H_{\pi_2}(X)).$$

   Since the entropies are the same, this is just a difference of expectations. By similar reasoning to above, it is of order $\Theta(\alpha^2)$ (there are $n/4$ missing edges, and they each contribute $(\alpha/\sqrt{n}) \cdot (c\alpha/\sqrt{n})$).

3. Any algorithm given $m \leq C_2 n \log(n)/\alpha^2$ samples fails to reconstruct with probability at least $1/2$. Since between any two models the KL for m samples is of order $m\alpha^2$ by tensorization and (2), and $\log|\mathcal{S}| \sim n\log(n)$, this follows directly from Fano's inequality (see e.g. [4]).

Combining these claims shows that $\sqrt{n\log(n)/m}$ is the tight rate for learning tree Ising models in TV distance.

References

[1] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and ising undirected graphical models. *arXiv preprint arXiv:1806.06887*, 2018.

[2] Richard S Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.

[3] Jörn Quistorff. A survey on packing and covering problems in the hamming permutation space. *the electronic journal of combinatorics*, 13(1):A1, 2006.

[4] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.