How Hard is Inference for Structured Prediction? (Supplementary Material)

A. Missing Proofs

Proof of Lemma 5.4: We first claim that $\widehat{Y}$ agrees with the data on at least half the edges of $\delta(F(B_i))$. The reason is that flipping the label of every vertex of $F(B_i)$ increases the agreement with the data by the number of disagreeing edges of $\delta(F(B_i))$ minus the number of agreeing edges of $\delta(F(B_i))$, and this difference is non-positive by the optimality of $\widehat{Y}$.

On the other hand, since B_i is maximal, every neighbor of B_i is correctly labeled in $\widehat{Y}$. Since the neighborhood of $F(B_i)$ is a subset of B_i, this also holds for $F(B_i)$. Thus, $\widehat{Y}$ disagrees with Y on every edge of $\delta(F(B_i))$.

We conclude that at least half the edges of $\delta(F(B_i))$ are bad. It is easy to see that the proof works for $-\widehat{Y}$, since it also maximizes Eq. 6 ■

Proof of Lemma 5.5: By the definition of a bad set we have that $\mathbf{Pr}[S \text{ is bad}]$ is equal to the probability that at least half of $\delta(S)$ are bad edges. Since $|\delta(S)| = i$ this is the probability that at least $\frac{i}{2}$ edges are bad. Since these events are IID, we can bound it via:

$$\mathbf{Pr}\left[\sum_j Z_j \geq \frac{i}{2}\right] < \binom{i}{\frac{i}{2}} p^{\frac{i}{2}} \leq (2e)^{\frac{i}{2}} p^{\frac{i}{2}} \leq (3\sqrt{p})^i \quad (1)$$

where Z_j is the indicator event of the j-th edge being bad. The first inequality is a union bound on all events where a specific set of size $\frac{i}{2}$ is bad, and the other edges can take any value. ■

Proof of Lemma 5.6: If F is a type 4 or 5 set, then $|\delta(F)| \geq \sqrt{N}$ and the bound is trivial. If F is a type 1 set, let U be the smallest rectangle in the dual graph (Diestel, 1997) which contains F. Let k, m denote the side lengths of U. Then: $|F| \leq km \leq \frac{1}{16}(2k+2m)^2 \leq \frac{1}{16}|\delta(F)|^2$. (To be clear, $km \leq \frac{1}{16}(2k+2m)^2$ because $4k^2+8km+4m^2-16km = (2k-2m)^2 \geq 0$.) Similarly for type 2 sets we have $|F| \leq km \leq \min\left\{(2k+m)^2, (k+2m)^2\right\} \leq |\delta(F)|^2$. Finally for type 3 sets: $|F| \leq km \leq (k+m)^2 \leq |\delta(F)|^2$. ■

Proof of Lemma 5.7: Recall that, by construction, a filled-in set $F \in \mathcal{F}$ is such that both $G[F]$ and $G[V \setminus F]$ are connected. In a planar graph such as G, this translates to an elegant characterization via the dual graph G^d. Recall that the dual graph has a vertex per face in G edges crossing the edges in G. Then it it easy to see that a set $\delta(F)$ is a boundary of a filled in set if and only if the dual edges corresponding to the edges $\delta(F)$ form a simple cycle in G^d (e.g., see Section 4.6 of Diestel, 1997). Note that the dual graph G^d is just an $(n-1) \times (n-1)$ grid, with one vertex per "grid cell" (i.e., face) of G, plus an extra vertex z of degree $4(\sqrt{N}-1)$ that corresponds to the outer face of G. The type-1 sets of $\mathcal{F}$ are in dual correspondence with the simple cycles of G^d that do not include z, the other sets of $\mathcal{F}$ are in dual correspondence with the simple cycles of G^d that do include z. The cardinality of the boundary $|\delta(F)|$ equals the length of the corresponding dual cycle.

Part (a) follows from the fact that $G^d \setminus \{z\}$ is a bipartite graph, with only even cycles, and with no 2-cycles.

For part (b), we count simple cycles of G^d of length i that do not include z. There are at most N choices for a starting point. There are at most 4 choices for the first edge, at most 3 choices for the next $(i-2)$ edges, and at most one choice at the final step to return to the starting point. Each simple cycle of $G^d \setminus \{z\}$ is counted $2i$ times in this way, once for each choice of the starting point and the orientation.

For part (c), we count simple cycles of G^d of length i that include z. We start the cycle at z, and there are at most $4\sqrt{N}$ choices for the first node. There are at most 3 choices for the next $i-2$ edges, and at most one choice for the final edge. This counts each cycle twice, once in each orientation. ■

Additional details for Theorem 5.1 Here we prove Equation (8) in the main text.

Let $\mathcal{F}_1 \subseteq \mathcal{F}$ denote the type-1 sets of $\mathcal{F}$. Recall that the random variable T is defined as:

$$T = \sum_{F \in \mathcal{F}} |F| \cdot 1_{F \text{ is bad}} \quad (2)$$

Then from linearity of expectation:

$$\mathbf{E}[T] = \sum_{F \in \mathcal{F}} |F| \cdot \mathbf{Pr}[F \text{ is bad}] \quad (3)$$

Next, we sum by size of $|\delta(F)|$, separating into $\mathcal{F}_1$ and the rest of $\mathcal{F}$.

$$\mathbf{E}[T] = \sum_{i=2}^{\infty} \sum_{F\in\mathcal{F}_1\,:\,|\delta(F)|=2i} |F| \cdot \mathbf{Pr}[F \text{ is bad}] + \tag{4}$$

$$\sum_{j=2}^{\infty} \sum_{F\in\mathcal{F}\setminus\mathcal{F}_1\,:\,|\delta(F)|=j} |F| \cdot \mathbf{Pr}[F \text{ is bad}] \tag{5}$$

Now use Lemmas 5.5 and 5.6 to bound both the size of $|F|$ and the probability that it is bad:

$$\mathbf{E}[T] \le \sum_{i=2}^{\infty} \sum_{F\in\mathcal{F}_1\,:\,|\delta(F)|=2i} \frac{i^2}{4} \cdot (3\sqrt{p})^{2i} + \tag{6}$$

$$\sum_{j=2}^{\infty} \sum_{F\in\mathcal{F}\setminus\mathcal{F}_1\,:\,|\delta(F)|=j} j^2 \cdot (3\sqrt{p})^{j} \tag{7}$$

Finally, we use Lemma 5.7 to bound the number of sets in $\mathcal{F}$ with a given size, yielding:

$$\mathbf{E}[T] \le \sum_{i=2}^{\infty} N \cdot \frac{2 \cdot 3^{2i-2}}{i} \frac{i^2}{4} \cdot (3\sqrt{p})^{2i} + \tag{8}$$

$$\sum_{j=2}^{\infty} 2\sqrt{N} \cdot 3^{j-2} \cdot j^2 \cdot (3\sqrt{p})^{j} \tag{9}$$

$$= N \sum_{i=2}^{\infty} \frac{i}{16} (81p)^i + \sqrt{N} \sum_{j=2}^{\infty} \frac{2j^2}{9} (9\sqrt{p})^j$$

$$= N(cp^2) + O(p\sqrt{N}), \tag{10}$$

for a constant $c > 0$ that is independent of p and N, and assuming $p < 1/81$.

The factor c can be improved as follows. First, we use the tighter upper bound of $(2ep)^{i/2}$ for the probability that a region of boundary size i is bad (see Lemma 5.5). We then replace the upper bound on the number of regions of each type in Lemma 5.7 with tighter results from statistical physics. In particular, the number of type-1 sets with boundary size i can be upper bounded by $N\mu^i$ (Eq. 3.2.5 of Madras & Slade, 1993), where μ is the so-called connective constant of square lattices and is upper bounded by 2.65 (Clisby & Jensen, 2012). The number of type 2–5 sets with boundary length i can similarly be upper bounded by $4\sqrt{N}\mu^i e^{\kappa\sqrt{i}}$ for the same value of μ and for some fixed constant $\kappa > 0$ (Hammersley & Welsh, 1962).

Next, we recognize that the term in (10) which is linear in N can be attributed to the type-1 regions. We expand the sum in (4) over type-1 regions into two terms: one term that explicitly enumerates over type-1 regions whose corresponding simple cycle in G^d is of length $i = 2$ to 100, and a remainder term. The sum in the first term can be computed exactly as follows. For each value of i, the probability that the region is bad is simply $\sum_{k=i/2}^{i} \binom{i}{k} p^k (1-p)^{i-k}$. We can then use the bound $\sum_{F\in\mathcal{F}_1:|\delta(F)|=i} |F| \le N \sum_{a=1}^{i^2/16} a c_{a,i}$, where $c_{a,i}$ is the number of distinct cycles in an infinite grid of length i and area a (up to translation). These cycles also go by the name of *self-avoiding polygons* in statistical physics, and the numbers $c_{a,i}$ have been exhaustively computed up to $i = 100$ (Jensen, 2000). Finally, the infinite sum in the remainder can be shown to be upper bounded by $51^2 b^{51}/(1-b)^3$ for $b = 2ep(2.65)^2$. The resulting function can then be shown to be upper bounded by $8Np^2$ for $p \le 0.017$, yielding a constant $c = 8$ as mentioned in the main text.

Formal Analysis of Second Stage Our starting point is $\mathbf{E}[H_0] \le N \cdot cp^2$, where H_0 is the Hamming error of the better of $\widehat{Y}$ and $-\widehat{Y}$. To calculate the error of the second stage, we need to consider the probability that it chooses the better of the two.

First, Markov's inequality implies that $\mathbf{Pr}\left[H_0 \ge \frac{1}{kp^2} Ncp^2\right] \le kp^2$, where k is a free parameter.

For the second stage, let B' be the set of wrong node observations. Chernoff bounds imply that, for sufficiently large N, $\mathbf{Pr}[|B'| \ge (1+\delta)Nq] \le \frac{1}{N^2}$. Observe that if the sum of the number of bad node observations and number of misclassified nodes for the better of $\widehat{Y}$ and $-\widehat{Y}$ is less than $N/2$, the two phase algorithm would choose the better of $\widehat{Y}$ and $-\widehat{Y}$. Hence with probability $1 - kp^2 - \frac{1}{N^2}$ the algorithm would choose the better of $\widehat{Y}$ and $-\widehat{Y}$, provided $\frac{1}{kp^2} Ncp^2 + (1+\delta)Nq < \frac{N}{2}$, or equivalently,

$$\frac{c}{k} + (1+\delta)q < \frac{1}{2}$$

For small δ and $k > \frac{c}{1/2-(1+\delta)q}$, this inequality would be satisfied and the better of $\widehat{Y}$ and $-\widehat{Y}$ would be chosen. Thus,

$$\mathbf{E}[H] \le 1 \cdot Ncp^2 + (kp^2 + \frac{1}{N^2}) \cdot N \tag{11}$$

$$\le N \cdot ((c+1)p^2 + kp^2) \le N \cdot Cp^2 \tag{12}$$

for $N > N_0(p,q)$ where H is the error of the 2-step algorithm. (in the second inequality we use $N > \frac{1}{p}$.)

Full proof of lower bound In the main paper, we give a proof sketch of the lower bound, Theorem 5.8. Here, we include a full proof of the fact that every binary classification algorithm suffers worst-case (over the ground truth) expected error $\Omega(p^2 N)$.

Let $G = (V, E)$ denote an $n \times n$ grid with $N = n^2$ vertices. Let $Y : V \to \{-1, +1\}$ denote the ground truth. We consider the case where Y is chosen at random from the following distribution. Color the nodes of G with black and white like a chess board. White nodes are assigned binary values uniformly and independently. Black nodes are assigned the label $+1$.

Given Y, input is generated using the random process described in Sec. 2.

Consider an arbitrary function from inputs to labelings of V. We claim that the expected error of the output of this function, where the expectation is over the choice of ground truth Y and the subsequent random input, is $\Omega(p^2N)$. This implies that, for every function, there exists a choice of ground truth Y such that the expected error of the function (over the random input) is $\Omega(p^2N)$.

Given Y, call a white node *ambiguous* if exactly two of the edges incident to itself are labeled "$+1$" in the input. A white node is ambiguous with probability $6p^2(1-p)^2 \geq 5.1p^2$ for $p \leq 0.078$. Since there are $N/2$ white nodes, and the events corresponding to ambiguous white nodes are independent, Chernoff bounds imply that there are at least $\frac{5p^2}{2}N$ ambiguous white nodes with very high probability.

Let L denote the error contributed by ambiguous white nodes. Since the true labels of different white nodes are conditionally independent (given that all black nodes are known to have value $+1$), the function that minimizes $\mathbf{E}[L]$ just predicts each white node separately. The algorithm that minimizes the expected value of L simply predicts that each ambiguous white node has true label equal to its input label. This prediction is wrong with constant probability, so $\mathbf{E}[L] = \Omega(p^2N)$ for every algorithm. Since L is a lower bound on the Hamming error, the result follows.

B. Illustration of Filled In Sets

Recall that for every subset S we defined a corresponding filled in set $F(S)$. The figures on the next page illustrate the transformation from a subset S to the filled in set $F(S)$.

References

Clisby, Nathan and Jensen, Iwan. A new transfer-matrix algorithm for exact enumerations: self-avoiding polygons on the square lattice. *J. Phys. A*, 45(11):115202, 15, 2012. ISSN 1751-8113.

Diestel, R. *Graph Theory*. Springer-Verlag, 1997.

Hammersley, J. M. and Welsh, D. J. A. Further results on the rate of convergence to the connective constant of the hypercubical lattice. *Quart. J. Math. Oxford Ser. (2)*, 13: 108–110, 1962. ISSN 0033-5606.

Jensen, Iwan. Size and area of square lattice polygons. *J. Phys. A*, 33(18):3533–3543, 2000. ISSN 0305-4470.

Madras, Neal and Slade, Gordon. *The self-avoiding walk*. Probability and its Applications. Birkhäuser Boston Inc., Boston, MA, 1993. ISBN 0-8176-3589-0.

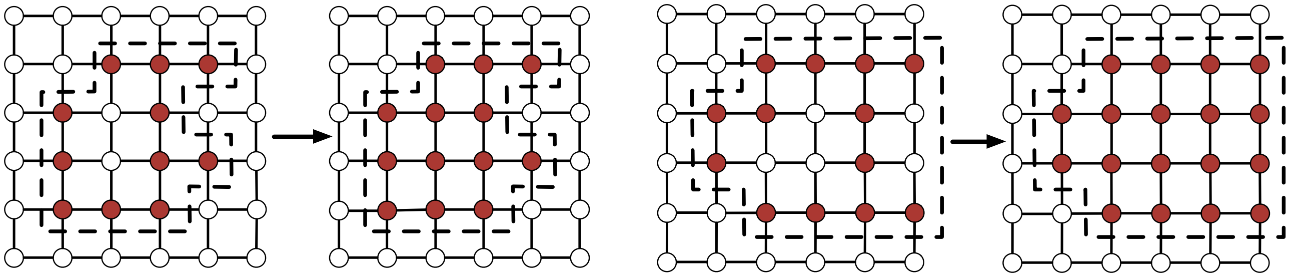

Figure 1. An example of type 1 set and corresponding filled-in set (left) and an example of type 2 set and corresponding filled-in set (right).

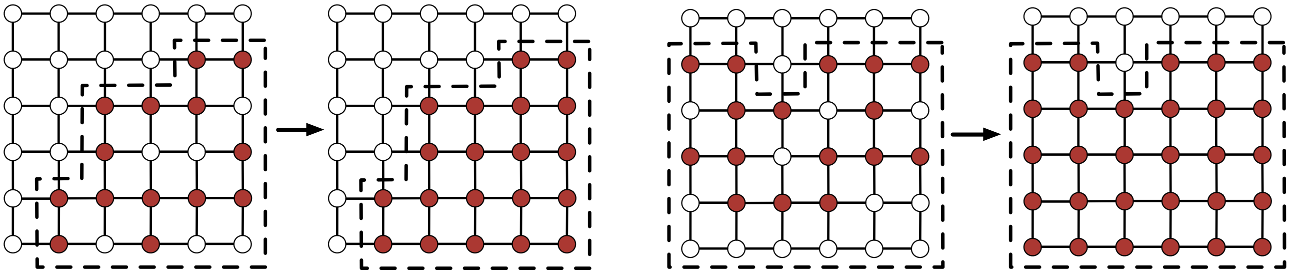

Figure 2. An example of type 3 set and corresponding filled-in set (left) and an example of type 4 set and corresponding filled-in set (right).

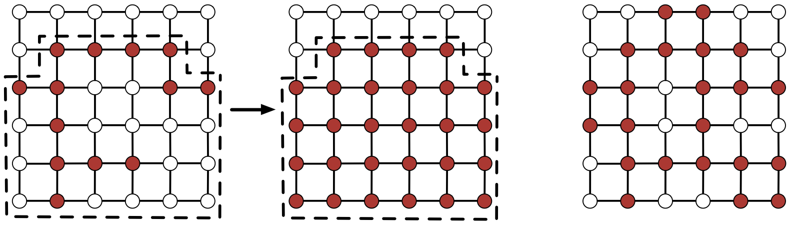

Figure 3. An example of type 5 set and corresponding filled-in set (left) and an example of type 6 set.