Planted Coloring and Clustering of solutions

Belief propagation for planted graph coloring

In HW3 you implemented and tested BP for random graph coloring. In this homework the same code will come handy as you can modify it in order to verify the properties of planted coloring we defined and discussed in class.

In particular, consider N nodes, each randomly colored with one of the the q colors. Call this color assignment the planted configuration $\{s_i^*\}_{i=1}^N$. Create a graph G on those nodes with M edges in such a way that edges only connect different planted colors. Run the following tests:

(a) Initialize BP close to the uniform fixed point, i.e. $1/q+\epsilon_s^{j\to i}$ and iterate the equations until convergence. Define converge as the time when the $\sum_{s,(ij)} |\chi_s^{i\to j}(t+1) - \chi_s^{i\to j}(t)| /(2qM) < \tilde{\epsilon}$ with suitably chosen small ϵ. Define overlap as

$$Q = \max_{\pi} \left[\frac{1}{N} \sum_{j=1}^{N} \chi^j_{\pi(s_j^*)} - \frac{1}{q} \right] \Bigg/ \left(1 - \frac{1}{q}\right) \tag{1}$$

Plot the overlap for $q = 3$ and $q = 5$ as a function of the average degree $c = 2M/N$. Try several sizes of graphs. Describe the threshold $c_{\rm alg}$ you observe.

(b) Initialize BP in the planted configuration, i.e. $\chi_s^{j\to i} = \delta_{s,s_j^*}$ and iterate the BP equations until convergence. Repeat the task from the previous point and discuss the results. Describe the threshold c_d you observe.

(c) In the region where the two above initializations do not lead to the same fixed point, compare the Bethe free entropies of the two fixed points and compute and discuss the threshold $c_{\rm IT}$.

(d) Monitor presence of frozen variables in the BP fixed points. At what average degree do frozen variables appear and how many of them? Find and discuss this threshold $c_{\rm r}$.

Solution.

(a)

(b)

(c)

(d)

□

Large q expansion

(a) In class we defined the annealed (1st moment) upper bound on the colorability threshold, $c_{\rm 1st}$. Write the large q expansion of the corresponding expression.

(b) In class we also discussed that when the number of colors is large the rigidity threshold $c_{\rm r}$ at which frozen variables appear in equilibrium scales like $q\log(q)$. Persuade yourself of this fact for the random regular graphs for which we derived that the fraction of frozen variables is given as

$$\eta_\ell = 1 + \sum_{r=1}^{q-1} (-1)^r \binom{q-1}{r} \left(1 - \frac{r\eta_{\ell+1}}{q-1}\right)^{d-1}$$

We remind the the rigidity threshold $c_{\rm r}$ was defined as the largest average degree for which the above update initialized at $\eta_\infty = 1$ converges to $\eta = 0$. Hint: When q is large, almost all the variables are frozen already close to the threshold $c_{\rm r}$.

Solution.

(a) From previous lecture,

$$S_{\rm annealed} = S_{\rm Bethe}|_{\chi\equiv\frac{1}{q}} = \log(q) + \frac{c}{2}\log\left(1-\frac{1}{q}\right)$$

Since $c_{\rm 1st}$ is the specific c at where $S_{\rm annealed} = 0$, we have for large q

$$\begin{aligned} c_{\rm 1st} &= -\frac{2\log(q)}{\log\left(1-\frac{1}{q}\right)} = -\frac{2\log(q)}{-\sum_{k=1}^{\infty} q^{-k}/k} \\ &= 2\log(q)\left[q - \frac{1}{2} - \frac{1}{12q} + \mathrm{O}\left(q^{-2}\right)\right] = (2q-1)\log(q) + \mathrm{o}(1) \end{aligned}$$

(b) When q is large, almost all the variables are frozen, so instead looking at the fraction of frozen variables, we look at the fraction of free variables $\theta_\ell \triangleq 1-\eta_\ell$ and assume the fixed point of θ is $\mathrm{o}(1)$. Then we have

$$\theta_\ell = -\sum_{r=1}^{q-1} (-1)^r \binom{q-1}{r} \left(1 - \frac{r(1-\theta_{\ell+1})}{q-1}\right)^{d-1}$$

We show that $c_{\rm r}$ scales like $(q-1)\left[\log(q-1) + \log(\log(q-1)) + \alpha\right]$. To the leading exponential order, we have

$$\begin{aligned} \binom{q-1}{r} &\simeq \exp\left(q\mathcal{H}_{\mathbb{e}}\left(\frac{r}{q}\right)\right) \\ \left(1 - \frac{r(1-\theta_{\ell+1})}{q-1}\right)^{d-1} &\simeq \exp\left(-\left[\log(q) + \log(\log(q)) + \alpha\right] r(1-\theta_{\ell+1})\right) \end{aligned}$$

So each term in the summation scales like

$$\exp\left(q\left[\mathcal{H}_{\mathbb{e}}\left(\frac{r}{q}\right) - \frac{r}{q}\left[\log(q) + \log(\log(q)+\alpha)\right](1-\theta_{\ell+1})\right]\right)$$

In the large-q limit, it is sufficient to only look at the term with $r=1$, the self-consistency equation reads:

$$\begin{aligned} \theta &= (q-1)\exp\left(-(1-\theta)\left[\log(q-1) + \log(\log(q-1)) + \alpha\right]\right) \\ &= \frac{1}{\log(q-1)}\mathbb{e}^{-\alpha}\left[(q-1)\log(q-1)\log(\alpha)\right]^\theta \simeq \frac{1}{\log(1-q)}\mathbb{e}^{-\alpha}(q-1)^\theta \end{aligned}$$

which is solved by $\theta = \gamma(\alpha)/\log(q-1)$ where $\gamma(\alpha)\mathbb{e}^{-\gamma(\alpha)} = \mathbb{e}^{-\alpha}$.

□

Condensed phase in the random sub-cubes model

The random-subcube model is defined by its solution space $S \subset \{0,1\}^N$ (not by a graphical model). We define S as the union of $\lfloor 2^{(1-\alpha)N} \rfloor$ random clusters (where $\lfloor x \rfloor$ denotes the integer value of x). A random cluster A being defined as:

$$A = \left\{ \vec{\sigma} \,\middle|\, \sigma_i \in \pi_i^A, \quad \forall\, i \in \{1,\dots,N\} \right\} \tag{2}$$

where π^A is a random mapping:

$$\begin{aligned} \pi^A \colon \{1,\dots,N\} &\to \{\{0\},\{1\},\{0,1\}\} \\ i &\mapsto \pi_i^A \end{aligned}$$

such that for each variable i, $\pi_i^A = \{0\}$ with probability $p/2$, $\{1\}$ with probability $p/2$, and $\{0,1\}$ with probability $1-p$. A cluster is thus a random subcube of $\{0,1\}$. If $\pi_i^A = \{0\}$ or $\{1\}$, variable i is said "frozen" in A; otherwise it is said "free" in A. One given configuration $\vec{\sigma}$ might belong to zero, one or several clusters. A "solution" belongs to at least one cluster.

We will analyze the properties of this model in the limit $N \to \infty$, the two parameters α and p being fixed and independent of N. The internal entropy s of a cluster A is defined as $\frac{1}{N}\log_2(|A|)$, i.e. the fraction of free variables in A. We also define complexity $\Sigma(s)$ as the (base 2) logarithm of the number of clusters of internal entropy s per variable (i.e. divide by N).

(a) What is the analog of the satisfiability threshold α_s in this model?

(b) Compute the $\alpha_{\rm d}$ threshold below which most configurations belong to at least one cluster.

(c) For $\alpha > \alpha_{\rm d}$ write the expression for the complexity $\Sigma(s)$ as a function of the parameters p and α. Compute the total entropy defined as $s_{\rm tot} = \max_s [\Sigma(s) + s \,|\, \Sigma(s) \geq 0]$. Observe that there are two regimes in the interval $\alpha \in (\alpha_{\rm d}, 1)$, discuss their properties and write the value of the "condensation" threshold $\alpha_{\rm c}$.

Solution.

(a) The $\alpha_{\rm s}$ is the threshold beyond which there is no "solution". A configuration is a "solution" if it belongs to at least one cluster, and the size of each cluster equals to $2^{\#\text{free variables}} \geq 1$.

Hence, a RSM has no "solution" if and only if there is zero cluster, i.e. $\alpha > 1$, which implies that the satisfiability threshold is $\alpha_{\rm s} = 1$.

(b) The probability that a given configuration belongs to a random cluster A is

$$\mathbb{P}\left(\left\{\sigma_i \in \pi_i^A \,\middle|\, i \in \{1,\dots,N\}\right\}\right) = \prod_{i=1}^{N} \mathbb{P}\left(\sigma_i \in \pi_i^A\right) = \prod_{i=1}^{N} \mathbb{P}\left(\pi_i^A \in \{\{\sigma_i\},\{0,1\}\}\right) = \left(1-\frac{p}{2}\right)^N$$

Therefore, the for a given configuration, the average number it belongs to is

$$2^{(1-\alpha)N}\left(1-\frac{p}{2}\right)^N = 2^{N\left[(1-\alpha)+\log_2\left(1-\frac{p}{2}\right)\right]} = 2^{N[\log_2(2-p)-\alpha]}$$

This means in the large-N limit, if $\alpha > \log_2(2-p)$, w.h.p. a given configuration does not belong to any random subcube, implying the clustering threshold $\alpha_{\rm d} = \log_2(2-p)$.

Note when $\alpha < \alpha_{\rm d}$, almost every configuration belongs to at least one cluster, so almost all configurations are "solutions", $s_{\rm tot} = \frac{1}{N}\log_2(|S|) = 1$.

(c) The probability that a cluster has internal entropy s is

$$\mathcal{P}(s) = \binom{N}{sN} (1-p)^{sN} p^{(1-s)N}$$

Let $\mathcal{N}(s)$ be the number of clusters with internal entropy s, then it is easy to see $\mathcal{N}(s) \sim \mathsf{Binomial}\left(2^{(1-\alpha)N}, \mathcal{P}(s)\right)$, i.e.

$$\mathbb{E}\left[\mathcal{N}(s)\right] = 2^{(1-\alpha)N}\mathcal{P}(s), \qquad \mathsf{Var}\left(\mathcal{N}(s)\right) = 2^{(1-\alpha)N}\mathcal{P}(s)\left(1-\mathcal{P}(s)\right)$$

By Markov and Chebyshev's inequality, we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{N}(s) \geq 1\right) &\leq \mathbb{E}\left[\mathcal{N}(s)\right] \\ \mathbb{P}\left(\left|\frac{\mathcal{N}(s)}{\mathbb{E}\left[\mathcal{N}(s)\right]} - 1\right| > \varepsilon\right) &\leq \frac{\mathsf{Var}\left(\mathcal{N}(s)\right)}{\left\{\mathbb{E}\left[\mathcal{N}(s)\right]\right\}^2 \varepsilon^2} \leq \frac{1}{2^{N(1-\alpha)}}\varepsilon^2\mathcal{P}(s), \quad \forall\, \varepsilon > 0 \end{aligned}$$

These equation prove the the random quantity $\frac{1}{N}\log\left(\mathcal{N}(s)\right)$ concentrates around its expectation in the large-N limit, that is,

$$\lim_{N\to\infty} \frac{1}{N}\log\left(\mathcal{N}(s)\right) = \begin{cases} \Sigma(s) := 1 - \alpha - D\left(s \,||\, 1-p\right), & \text{if } \Sigma(s) \geq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

where the binary KL divergence is defined as

$$D\left(x \,||\, y\right) = x\log_2\left(\frac{x}{y}\right) + (1-x)\log_2\left(\frac{1-x}{1-y}\right)$$

Now we consider the regime $\alpha > \alpha_{\mathrm{d}}$, The total entropy is given by a saddle-point estimation, to the leading exponential order we have

$$\sum_A 2^{Ns(A)} \simeq N\int_s \mathrm{d}s\; 2^{N[\Sigma(s)+s]}\mathbb{I}\left(\Sigma(s) \geq 0\right)$$

Although it seems to over count the solutions since a configuration may belong to several clusters and this sum just adds the cluster size, but in the regime $\alpha > \alpha_{\mathrm{d}}$ it is valid because since in every cluster the fraction of solutions belonging to more than one cluster is exponentially small.

Taking derivative of $\Sigma(s) + s$ and setting it to zero solves

$$\frac{\partial}{\partial s}\left[\Sigma(s)+s\right] = \log_2\left(\frac{1-p}{p}\frac{1-s}{s}\right)+1 \quad \Rightarrow \quad \tilde{s} = \frac{2(1-p)}{2-p}, \quad \Sigma(\tilde{s}) = \frac{p}{2-p} - \alpha + \log_2(2-p)$$

- When $\Sigma(\tilde{s}) > 0$, i.e. $\alpha < \alpha_{\mathrm{c}} := \frac{p}{2-p} + \log_2(2-p)$, the correct maximizer is $s^* = \tilde{s}$

$$s_{\mathrm{tot}} = \Sigma(\tilde{s}) + \tilde{s} = 1 - \alpha + \log_2(2-p)$$

- When $\Sigma(\tilde{s}) < 0$, the correct maximizer $s^* = s_M = \max\left\{s \,|\, \Sigma(s) >= 0\right\}$, which is the cross point of $\Sigma(s)$ and the x-axis

$$s_{\mathrm{tot}} = \Sigma(s_M) + s_M = 0 + s_M = s_M$$

Therefore, there are four different phases

(1) Liquid phase: $\alpha < \alpha_{\mathrm{d}}$, almost all configurations are solutions, this can be also seen from $s_{\mathrm{tot}} = 1$

(2) Clustered phase: $\alpha_{\mathrm{d}} < \alpha < \alpha_{\mathrm{c}}$, the solutions set S is partitioned into exponentially many non-overlapping clusters. Most solutions are in the $\mathbb{e}^{N\Sigma(\tilde{s})}$ clusters with internal entropy $\tilde{s}$.

(3) Condensed clustered phase: $\alpha_{\mathrm{c}} < \alpha < \alpha_{\mathrm{s}}$, the solutions set S is partitioned into exponentially many non-overlapping clusters. However, most solutions are in the clusters with internal entropy s_M. The number of such clusters is not exponentially large, as $\Sigma(s_M) = 0$.

(4) Unsatisfiable phase: $\alpha > \alpha_{\mathrm{s}}$, there is no cluster, and thus no solutions.

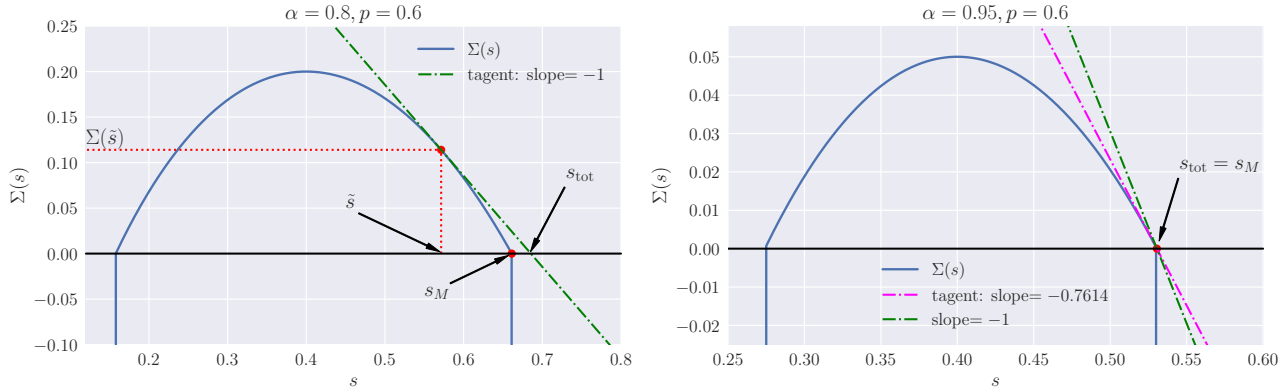

Figure 1: Complexity function $\Sigma(s)$ under clustered phase (left, $\alpha = 0.8, p = 0.6$) and condensed clustered phase (right, $\alpha = 0.95, p = 0.6$)

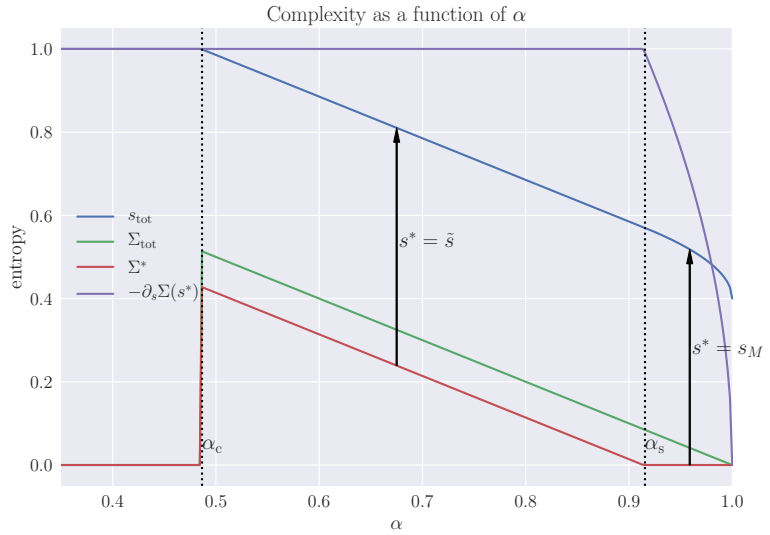

Figure 2: Complexity as a function of α for $p = 0.6$.

□