Randomized Algorithms 2019A – Lecture 9
Edge-Sparsification of Hypergraphs and Some Occupancy Problems*

Robert Krauthgamer

1 Edge-Sparsification of Hypergraphs (via Importance Sampling)

Cuts in Hypergraphs: Let $H = (V, E, w)$ be a hypergraph with edge weights $w : E \mapsto \mathbb{R}_+$. Every (nontrivial) $S \subset V$ defines a cut

$$\delta_H(S) := \{e \in E : \text{both } e \cap S \neq \emptyset, e \cap \bar{S} \neq \emptyset\}.$$

For every subset of edges $E' \subset E$ define $w(E') = \sum_{e \in E'} w(e)$, which in particular defines the weight of a cut S as $w(\delta_H(S))$.

Cut sparsifier: Let $H = (V, E, w)$ be a hypergraph, and let $\varepsilon \in (0, 1)$. A hypergraph $H' = (V, E', w')$ (on same vertex set) is a $(1 + \varepsilon)$-cut-sparsifier if

$$\forall S \subset V, \qquad w'(\delta'_H(S)) \in (1 \pm \varepsilon) w(\delta_H(S)).$$

Theorem 1 [Kogan and Krauthgamer, 2015]: For every n-vertex hypergraph $H = (V, E, w)$ and every $\varepsilon \in (0, 1/2)$, there is a $(1 + \varepsilon)$-cut-sparsifier H' with $m = O(n^2/\varepsilon^2)$ hyperedges.

The original proof was an extension of an earlier result, by [Benczur and Karger, 1996], which introduced the concept for graphs (all hyperedges are of size 2) and proved a (better) bound of $O(\varepsilon^{-2} n \log n)$ edges and also gave an algorithm with near-linear running time. We will see a different proof based on Importance Sampling.

Idea: We will construct H' by sampling m edges, where each edge is drawn according to probabilities $\{p(e)\}_{e \in E}$, and a sampled edge e is given new weight $w'(e) = \frac{w(e)}{mp(e)}$. It follows immediately that the expected weight of a (every) cut S in H' equals the weight of the same cut S in H. Viewing this as importance sampling, it will be easy to reduce the variance. But this holds only for any one cut S, and a sparsifier H' requires a guarantee for *all cuts*, and thus we will prove a concentration bound and then apply a union bound.

Construction of sparsifier: For each edge $e \in E$ define its *sensitivity*

$$s(e) := \max_{S \subset V:\ e \in \delta_H(S)} \frac{w(e)}{w(\delta_H(S))},$$

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

and define the total sensitivity to be $s(E) = \sum_{e \in E} s(e)$.

Construct H' at random by picking m edges from H, each chosen independently according to the distribution on edges given by $p(e) = \frac{s(e)}{S(E)}$, and every edge $e \in E$ that is chosen is given a new weight $w'(e) = \frac{w(e)}{m\ p(e)}$.

Remark: This construction may create parallel edges, because the same edge may be picked multiple times (up to m). We can always merge parallel edges at the end, which does not change the weight of any cut, but it will be easier for us to analyze the hypergraph before such merges.

Expectation:

$$\forall S \subset V, \qquad \mathbb{E}[w'(\delta_{H'}(S))] = m \sum_{e \in \delta_H(S)} p(e) \frac{w(e)}{m\ p(e)} = \sum_{e \in \delta_H(S)} w(e) = w(\delta_H(S)).$$

Lemma 2: $s(E) \leq n^2$.

Proof: Was seen in class by "charging" the sensitivity of each hyperedge to the minimum cut between a pair of vertices.

Lemma 2': $s(E) \leq n - 1$.

Exer: Prove this bound by repeatedly removing from G a minimum cut whose removal increases the number of connected components by 1 (this is a global minimum cut in one of the components).

Hint: It then suffices to show that in each of the $n - 1$ iterations, the sensitivity of the removed edges sums up to at most 1.

Lemma 3 (Importance Sampling): Let $S \subset V$ and $\lambda = s(E)$. Then

$$\forall e \in \delta_H(S), \qquad p(e) \geq \frac{1}{\lambda} \frac{w(e)}{w(\delta_H(S))}.$$

Proof: Given our S and e,

$$p(e) = \frac{s(e)}{s(E)} \geq \frac{1}{s(E)} \frac{w(e)}{w(\delta_H(S))}.$$

QED

Corollary 4: If $m \geq c \cdot s(E)$ for a suitable constant $c > 0$, then

$$\forall S \subset V, \qquad \Pr[w'(\delta_{H'}(S)) \in (1 \pm \varepsilon) w(\delta_H(S))] \geq 3/4.$$

Proof: Using the importance sampling theorem we saw in previous classes, and the bound in Lemma 3, $\mathrm{Var}(w'(\delta_{H'}(S))) \leq m \cdot \frac{1}{m^2} \cdot \lambda \cdot (w(\delta_H(S)))^2 = (\frac{1}{2}\varepsilon w(\delta_H(S)))^2$. The corollary now follows by Chebyshev's inequality.

QED

Proof of Theorem 1: Was seen in class, by proving a concentration bound for each cut, and then applying a union bound over all 2^n cuts.

Exer: Show that with high probability the total weight of all edges of H' is approximately equal to that in H, i.e., $w'(E') = \Theta(w(E))$.

Exer: Analyze a variant of this algorithm, where the sampling is different: Independently for each edge $e \in E$, with probability $q(e) = \min\{1, O(\varepsilon^{-2}n) \cdot s(e)\}$ add this edge to H' with new weight $w'(e) = \frac{w(e)}{q(e)}$, and otherwise do not add to H'. Note that now the number of edges is random (and has to be analyzed).

2 Balls and Bins (Occupancy Problems)

Problem definition: Suppose we throw balls independently and uniformly into n bins. We stop after $m = m(n)$ balls and examine the most/least loaded bin.

One motivation: load balancing or an ideal model for a hash function

Expected behavior: Let X_i be the load of bin $i \in [n]$. The expected load of a fixed bin i is $\mathbb{E}[X_i] = m/n$.

But don't we expect deviations?

$m = n$ **balls, empty bins:** We expect one ball per bin. But how many bins will be empty?

$$\Pr[\text{bin } i \text{ is empty}] = \Pr[X_i = 0] = (1 - 1/n)^n \approx 1/e.$$

Therefore,

$$\mathbb{E}[\#\text{ of empty bins}] = \mathbb{E}[\sum_i I_{\{X_i=0\}}] = \sum_i \Pr[X_i = 0] \approx n/e.$$

Exer: Give a high probability bound by bounding the variance and using Chebyshev's inequality. Compare to Markov's inequality.

Exer: How many bins are expected to have load 1?

$m = n$ **balls, maximum load:**

$$\Pr[X_i \geq 2\log n] \leq 2^{-2\log n} = 1/n^2$$

(by one of the Chernoff bounds) and therefore

$$\Pr[\max_i X_i \geq 2\log n] \leq \sum_i \Pr[X_i \geq 2\log n] \leq 1/n.$$

Exer: Show that if $m \geq 10n\log n$, then with high probability the maximum load and the minimum load are within factor 2 of the expected load. For what value of m the ratio will be $1 + \varepsilon$?

Review of key points:

1. Use a union bound to analyze the maximum
2. High probability often implies expectation

Hitting all bins (Coupon Collector):

Let Y_i be the number balls thrown until i distinct bins are hit. We are interested in Y_n, and by definition $Y_1 = 1$. Observe that $Z_i = Y_i - Y_{i-1}$ has geometric distribution $G(p = \frac{n-(i-1)}{n})$. Thus,

$$\mathbb{E}[Z_i] = \tfrac{1}{p} = \tfrac{n}{n-i+1}, \quad \mathrm{Var}(Z_i) = (1-p)/p^2 = \tfrac{i-1}{n} \cdot \tfrac{n^2}{(n-i+1)^2} = \tfrac{(i-1)n}{(n-i+1)^2}.$$

Since we can write $Y_n = \sum_{i=1}^{n} Z_i$ (by convention $Z_1 = 1$),

$$\mathbb{E}[Y_n] = \mathbb{E}[\sum_{i=1}^{n} Y_i - Y_{i-1}] + \mathbb{E}[Y_1] = \sum_{i=1}^{n} \tfrac{n}{n-i+1} = nH_n \approx n \ln n.$$

$$\mathrm{Var}(Y_n) = \mathrm{Var}(\sum_{i=1}^{n} Z_n) = \sum_{i=1}^{n} \mathrm{Var}(Z_n) \leq \sum_{j=1}^{n} \tfrac{(n-j)n}{j^2} \leq O(n^2).$$

Using Chebyshev's inequality,

$$\Pr[Y_n \geq 3n \ln n] \leq \Pr[Y_n - \mathbb{E}\, Y_n \geq 2n \ln n] \leq O(1/\ln^2 n).$$

We can get a stronger bound using a direct calculation:

$$\Pr[X_1 = 0] \leq (1 - 1/n)^m \leq e^{-m/n} = 1/n^3,$$

hence

$$\Pr[\exists i, X_i = 0] \leq n \Pr[X_1 = 0] \leq 1/n^2.$$