EE 736: Introduction to Stochastic Optimization Project

Compression of Graphical Data

Adway Girish (180070002)

Submitted on: April 11, 2022

Abstract

In this report, I present a survey of some recent work on the compression of graphical data. The aim is to understand the techniques used in the compression schemes. I do not provide complete proofs here, since they may be directly referred to in the original publications. Instead, I look to provide intuitive explanations and highlight the key ideas that make the schemes successful.

1 Introduction

Data in the form of graphs is now ubiquitous. This is primarily due to their simplicity of representation which offers an incredible level of versatility and abstraction. In applications such as social networks and molecular biology, the graphs used could be very large. Naturally, finding efficient ways to compress graphs becomes essential to make them useful for real-world applications. This forms the central theme of this report.

In Section 2, I introduce the relevant preliminaries and present the notation that will be used throughout, identical to [DA20a]. Also from the same paper is a *universal lossless compression* scheme that forms the basis of this study, described in Section 3. When the graphs are *sparse*, [DA20b] obtains similar compression with a computationally optimal scheme, covered in Section 4. Section 5 also deals with sparse graphs, but in a different regime of sparsity – *heavy-tailed sparse graphs*, as done in [DA21]. I provide concluding remarks in Section 6.

2 Preliminaries and Notation

Before moving to studying different settings and their associated compression schemes, there is an extensive set of notation that must be described.

General symbols and terms. $[n] = \{1, 2, \ldots, n\}$, log is the natural logarithm, $\{0,1\}^*$ is the set of finite nonempty binary sequences, and $\mathsf{nats}(x)$ is the length of x in nats, i.e. $\log 2$ times the length of the binary representation of x. To define $d_{\mathrm{LP}}(\mu, \nu)$, the Lévy-Prokhorov distance between Borel probability measures on a complete separable (Polish) space Ω, we require B^ϵ, the union of all open balls of radius ϵ centered at points in the Borel set $B \subset \Omega$. Then $d_{\mathrm{LP}}(\mu, \nu)$ is the infimum of all $\epsilon > 0$ such that $\mu(B) \le \nu(B^\epsilon) + \epsilon$ and $\nu(B) \le \mu(B^\epsilon) + \epsilon$ for all Borel sets B. Further, on the set of Borel probability measures on a Polish space Ω, a sequence of measures μ_n converges weakly to μ ($\mu_n \Rightarrow \mu$) if for any continuous bounded function on Ω, $\int f\, d\mu_n \to \int f\, d\mu$. For $x \in \Omega$, δ_x is the Dirac measure at x

Basic graph notation. Note that we consider *marked graphs*, i.e. graphs where the edges and vertices carry marks from some fixed, finite edge mark set Ξ and vertex mark set Θ respectively. Let G be a graph, then $V(G)$ represents its vertex set. Given a vertex $v \in V(G)$, $\tau_G(v) \in \Theta$ is the mark of vertex v, and similarly, given directed edge from vertex v to w, denoted by (v, w), we have $\xi_G(v, w) \in \Xi$, the mark of edge (v, w). If there is an edge connecting vertex v to w, we write $v \sim_G w$. The number of edges connected to a vertex v is given by $\deg_G(v)$. Another notion of the degree is given by $\deg_G^{x,x'}(v)$, which is the number of neighbours w such that $w \sim_G v$, $\xi_G(w, v) = x$, and $\xi_G(v, w) = x'$. For some distinguished vertex o, (G, o) is a rooted marked graph, rooted at o. An equivalence relation can be defined through a root preserving bijection, then the equivalence class is given by $[G, o]$. Given an integer $h \ge 1$, $(G, o)_h$ is the restriction of (G, o) to a depth of h, i.e. containing vertices that are at most h distance away from o. Naturally, we have $[G, o]_h = [(G, o)_h]$. Finally, G^Δ, for some positive integer Δ, is the graph with the same vertex, but containing only those edges that are whose endpoints are of degree at most Δ.

Convergence of graphs. Let $\bar{\mathcal{G}}_*$ be the space of equivalence classes $[G, o]$. To define a metric on this space, first define $\hat{h}$ as the supremum over integers $h \ge 0$ such that $[G, o]_h = [G', o']_h$, then set $\bar{d}_*([G, o], [G', o']) = 1/(1 + \hat{h})$. With this metric, we have that $\bar{\mathcal{G}}_*$ forms a Polish space. We define $U(G)$ to be the law of a graph $[G, o]$, where o is drawn uniformly at random in G, i.e. $U(G) = \frac{1}{|V(G)|} \sum_{o \in V(G)} \delta_{[G,o]}$. These can also be restricted to the space of trees $\bar{\mathcal{T}}_*$, instead of $\bar{\mathcal{G}}_*$, as the set of equivalence classes $[G, o]$ where G is a tree. We can further define $\bar{\mathcal{G}}_*^h$ and $\bar{\mathcal{T}}_*^h$ as the restrictions of $\bar{\mathcal{G}}_*$ and $\bar{\mathcal{T}}_*$ to a depth $h \ge 1$.

Compression basics. To encode a vector with k elements, we require a sequence of length given by $\log k$ nats $= \log_2 k$ bits.

3 Universal Lossless Compression

The aim is to have a compression scheme that is *universal* (does not make any assumptions about the statistical properties of the data to be compressed) and *lossless* (a decompression function should be prescribable to recover the original data perfectly). The main result from [DA20a] is that not only is this possible, but also in an optimal sense. Before formally describing these results, we require some more definitions, which are given below.

Fix some finite marked graph G, with Ξ and Θ as its fixed and finite edge and vertex mark sets respectively. Then define $\vec{m}_G = (m_G(x, x') : x, x' \in \Xi)$, the *edge mark count vector* of G, where $m_G(x, x')$ is the number of edges (v, w) in G with $(\Xi_G(v, w), \Xi_G(w, v)) = (x, x')$ or (x', x). Similarly, the *vertex mark count vector* of G, $\vec{u}_G = (u_G(\theta) : \theta \in \Theta)$, where $u_G(\theta)$ is the number of vertices v in G with $\tau_G(v) = \theta$.

Now let $\vec{m}$ and $\vec{u}$ be some edge and vertex mark count vectors respectively. We define $\|\vec{m}\|_1 = \sum_{x \leq x' \in \Xi} m(x, x')$ and $\|\vec{u}\|_1 = \sum_{\theta \in \Theta} u(\theta)$. Given an $n \in \mathbb{N}$, we also define $\mathcal{G}^{(n)}_{\vec{m},\vec{u}}$ to be the set of marked graphs on the vertex set [n] with $\vec{m}_G = \vec{m}$ and $\vec{u}_G = \vec{u}$. Clearly, we require $\|\vec{u}\|_1 = n$ and $\|\vec{m}\|_1 \leq \binom{n}{2}$ for $\mathcal{G}^{(n)}_{\vec{m},\vec{u}}$ to be nonempty.

3.1 Main results

Let $\bar{\mathcal{G}}_n$ be the set of marked graphs with vertex set [n] and edge and vertex mark set given by Ξ and Θ respectively. Let $\{f_n, g_n\}_{n=1}^{\infty}$ be a compression-decompression scheme such that f_n compresses $\bar{\mathcal{G}}_n$ to $\{0,1\}^*$ and g_n decompresses $\{0,1\}^*$ to $\bar{\mathcal{G}}_n$, and $g_n \circ f_n(G)$ for all $G \in \bar{\mathcal{G}}_n$. Then we have the following results, which are rephrased from [DA20a].

Theorem 1. *For all 'good' mark count vectors $\vec{m}^{(n)}$ and $\vec{u}^{(n)}$, and $\mu \in \mathcal{P}(\bar{\mathcal{T}}_*)$, there exists a sequence of positive $\epsilon_n \downarrow 0$, and $G^{(n)}$ drawn uniformly from $\mathcal{G}^{(n)}_{\vec{m},\vec{u}}(\mu, \epsilon_n) = \left\{G \in \mathcal{G}^{(n)}_{\vec{m},\vec{u}} : d_{\mathrm{LP}}(U(G), \mu) < \epsilon_n\right\}$, such that*

$$\liminf_{n\to\infty} \frac{\mathsf{nats}\left(f_n\left(G^{(n)}\right)\right) - \left\|\vec{m}^{(n)}\right\|_1 \log n}{n} \geq \Sigma(\mu) \ a.s.,$$

where $\Sigma(\mu)$ is the marked BC entropy *(generalized from [BC14]).*

The definition of the BC entropy is itself the result of several assumptions, which are subsumed in our restriction to 'good' count vectors and distributions (see Theorems 1 and 2 from [DA20a] for a more complete characterization – we need the count vectors to be 'adapted to' average degree vectors, and distributions to be 'unimodular'; additionally, μ defined on the space of graphs not restricted to trees are also 'bad' – $\Sigma(\mu) = -\infty$ in all these cases).

Theorem 2. *There exists a lossless compression scheme $\{f_n\}_{n=1}^{\infty}$ such that for any sequence of marked graphs $G^{(n)}$ with weak local limit $\mu \in \mathcal{P}(\mathcal{T}_*)$, we obtain*

$$\limsup_{n\to\infty} \frac{\mathsf{nats}\left(f_n\left(G^{(n)}\right)\right) - \|\vec{m}_{G^{(n)}}\|_1 \log n}{n} \leq \Sigma(\mu) \ a.s. \tag{1}$$

Theorem 1 says that for any lossless compression scheme, we have a lower bound on the compressed length per unit vertex of the graph, that is linear with a notion of 'entropy' associated with μ. Conversely, Theorem 2 says that we have a scheme (given below) that attains the same bound, thus making the scheme asymptotically optimal.

3.2 Compression scheme

We start with a simpler version of the scheme that makes an assumption on the graph – that its maximum degree is at most Δ_n.

We need some more definitions. Let k and Δ be integers, then $\mathcal{A}_{k,\Delta}$ is defined as the (finite) set of equivalence classes of rooted marked graphs $[G,o] \in \bar{\mathcal{G}}_*$ with depth at most k and maximum degree at most Δ. Let $[G,o] \in \mathcal{A}_{k,\Delta}$, and let $G^{(n)}$ be a marked graph on [n] with maximum degree at most Δ_n. Then $\psi_{G^{(n)}}^{(n)}([G,o])$ is the set of vertices of $G^{(n)}$ that is locally isomorphic to $[G,o]$ up to depth k_n, i.e. $= \{1 \leq i \leq n : [G^{(n)}, i]_{k_n} = [G,o]\}$. These sets $\psi_{G^{(n)}}^{(n)}([G,o])$ form a partition of $[n]$ over the range of $[G,o]$ if $G^{(n)}$ has maximum degree at most Δ_n.

3.2.1 First-step scheme

If the graph $G^{(n)}$ has maximum degree at most Δ_n, we may encode it as follows:

1. Encode the vector $\left(\left|\psi_{G^{(n)}}^{(n)}([G,o])\right|, [G,o] \in \mathcal{A}_{k_n,\Delta_n}\right)$, which lists the number of times each element of $\mathcal{A}_{k_n,\Delta_n}$ appears in $G^{(n)}$.
2. Add elements to the encoded vector to specify $G^{(n)}$ from W^n, which is the set of marked graphs G on $[n]$ such that $\left|\psi_{G^{(n)}}^{(n)}([G,o])\right| = \left|\psi_{G^{(n)}}^{(n)}([G',o'])\right|$ for all $[G',o'] \in \mathcal{A}_{k_n,\Delta_n}$.

Step 1 requires at most $|\mathcal{A}_{k_n,\Delta_n}|(1 + \lfloor \log_2 n \rfloor)$ nats, since $\left|\psi_{G^{(n)}}^{(n)}([G,o])\right| \leq n$ for all $[G,o] \in \mathcal{A}_{k_n,\Delta_n}$ and step 2 requires at most $1 + \lfloor \log_2 |W_n| \rfloor$ nats. If k_n and Δ_n are chosen such that $|\mathcal{A}_{k_n,\Delta_n}| = o\left(\frac{n}{\log n}\right)$, and $k_n \to \infty$ with n, it can be shown by simply writing out the expressions and calculating the required terms, that (1) is satisfied.

We can now use this first-step scheme to develop a general scheme.

3.2.2 General scheme

We first reduce our graph to satisfy the requirements for it to be compressed using the first-step scheme above, then encode the leftover information separately, as follows:

1. Remove all edges connected to any vertex with degree more than $\Delta_n = \log \log n$, and let the trimmed graph (with the same vertex set) be $\tilde{G^n} = (G^{(n)})^{\Delta_n}$.
2. Encode $\tilde{G^n}$ using the first-step scheme above.
3. Define R_n to be the set of endpoints of removed edges, $= \{1 \le i \le n : \deg_{G(n)}(i) > \delta_n \text{ or } \deg_{G(n)}(i) > \delta_n \text{ for some } j \sim_{G^{(n)}} (i)\}$, and encode $|R_n|$, then R_n.
4. Encode the vector $\vec{m}_{G^{(n)}} - \vec{m}_{\tilde{G^{(n)}}}$, which is the number of removed edges, then encode the removed edges themselves.

Just as for the first-step scheme, a lengthy calculation verifies that (1) holds. The complete proof is given in [DA20a].

4 Efficient Universal Compression of Sparse Graphs

A graph with n vertices can be called sparse if the number of edges are much smaller than n^2. An interesting way to maintain sparsity with structure is to have the number of edges grow linearly in n. This is the regime that has been assumed here. Similar to the compression scheme developed for arbitrary marked graphs in the previous section, [DA20b] develops a universal lossless compression scheme that exploits sparsity to be of a lower complexity. The key difference is that here, we further partition and encode the edges separately based on their types, which reduces the complexity.

4.1 Main result

Similar to Theorem 2, we have a theorem guaranteeing the existence of an optimal compression scheme, given by Theorem 3 below.

Theorem 3. *There exists a lossless compression scheme* $\{f^{(n)}_{h,\delta}\}$ *with positive integer parameters* h *and* δ*, satisfying:*

1. optimality. *For a sequence* $G^{(n)}$ *of marked graphs such that* $U(G^{(n)}) \Rightarrow \mu$*, again a 'good' distribution on* $\bar{\mathcal{T}}_*$*, let* $m^{(n)}$ *be the number of edges in* $G^{(n)}$*. Then we have*

$$\limsup_{h\to\infty} \limsup_{\delta\to\infty} \limsup_{n\to\infty} \frac{\mathsf{nats}\left(f^{(n)}_{h,\delta}\left(G^{(n)}\right)\right) - m^{(n)} \log n}{n} \le \Sigma(\mu) \text{ a.s.},$$

where $\Sigma(\mu)$ *is the marked BC entropy, as described in Section 3.*

2. low computational complexity. *The complexity of the compression and decompression algorithm is* $O(n \,\mathrm{polylog}(n)) = O\left(n(\log n)^k\right)$ *for some positive integer* k*. We also have a lower bound of* $\Omega(n \log n)$*, which makes the scheme asymptotically computationally optimal up to factors of* $\log n$*.*

To justify the optimality claim, there exists a converse statement similar to Theorem 2, which says that we need $\Sigma(\mu)$ nats to encode each vertex of the graph.

4.2 Compression algorithm

Recall that the improvement that makes this algorithm efficient is that the edges are further divided into groups and each group (which is now an unmarked graph) is encoded separately. Here I provide a summary of the scheme, describing essentially what happens at each step. For more details refer to [DA20b].

1. Define $\mathcal{F}^{\delta,h}$ to be the set of all $(x,[T,o]) \in \Xi \times \bar{\mathcal{T}}_*^{h-1}$ such that $\deg_T(o) < \delta$ and $\deg_T(v) \le \delta$ for $v \neq o$. Further, we also define a fictitious symbol $\star_x$ for each $x \in \Xi$. Then the set of edge types that we will partition our graph into is given by $\mathcal{F}^{\delta,h} \cup \{\star_x : x \in \Xi\}$. The exact allotment of types is a little more complicated.
2. Calculate the edge types across all vertices using a message passing algorithm.
3. Encode the 'star vertices' and 'star edges', which are the vertices and edges corresponding to the edges marked with the $\star_x$ type for some $x \in \Xi$. The star edges account for the largest length portion in the compressed output. The encoding is done by comparing with possible vertices that it could be connected to, sharing an edge with the same type, and adding that information to the output.
4. Next, we encode the 'vertex types', which store the marks and number of vertices with those marks for each edge type.
5. Finally, we encode the edges that are not star edges. This is done by partitioning the leftover edges from the graph to form several unmarked graphs (with vertices having small degrees). Each of them can be easily encoded separately.

While the scheme may appear complicated with details and unclear without, the takeaway is quite straightforward – we first separate the edges with high degrees at the vertices and encode them, then encode the unmarked graphs that partition the remaining edges.

5 Heavy-Tailed Sparse Graphs

In Section 4, the sparsity regime considered had the number of edges growing linearly with the number of vertices n. However, they can remain sparse even if their edges grow superlinearly with n – they give rise to sparse graphs whose degree distributions have heavy tails. To differentiate between the two regimes, we call the former *sparse graphs* and the latter *heavy-tailed sparse graphs*. The heavy-tailed regime is addressed by introducing the notion of *sparse graphons*, which then define a new metric.

5.1 Sparse graphon framework

On a probability space $(\Omega, \mathcal{F}, \pi)$, define the *graphon* $W : \Omega \times \Omega \to \mathbb{R}_+$, the set of nonnegative real numbers, which is symmetric and satisfies $\|W\|_1 = \int W(x,y)\, d\pi(x) d\pi(y) <$

∞. More generally, an L^p graphon satisfies $\|W\|_p^p = \int (W(x,y))^p \, d\pi(x) d\pi(y) < \infty$. If $\|W\|_1 = 1$, the graphon W is normalized.

For L^2 graphons W and W' defined on $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$, we define the metric

$$\delta_2(W, W') = \inf_{\nu} \sqrt{\int |W(x,y) - W'(x',y')|^2 d\nu(x,y) d\nu(x',y')},$$

where ν takes all possible joint distributions on $\Omega \times \Omega'$ such that the marginals are π and π'.

Given a normalized graphon W and the target densities ρ_n, we can generate a sequence of graphs $G^{(n)}$ that are said to be W*-random* by independently placing edges between vertices v and w with probability $1 \wedge \rho_n W(X_v, X_w)$, where X_i are generated i.i.d. according to π. The law of $G^{(n)}$ is given by $\mathcal{G}(n; \rho_n W)$. We further define a notion of *sparse graphon entropy*,

$$\mathrm{Ent}(W) = \mathbb{E}[W \log W] - \mathbb{E}[W] \log \mathbb{E}[W].$$

When W is normalized, we simply have $\mathrm{Ent}(W) = \mathbb{E}[W \log W]$. Recall that the definition of information-theoretic entropy for a random variable X with mass function p on $\{1, \ldots, n\}$ is given by $H(X) = -\sum_{i=1}^n p(i) \log p(i)$. This suggests that the sparse graphon notion of entropy in fact measures the uncertainty in an inverse sense. Thus we expect that the bound on the length of the code needed should vary opposite to $\mathrm{Ent}(W)$ (which, as we will see in (2), is indeed the case).

5.2 Main result

Just as in the previous settings, we have a result that gives the existence of an optimal compression scheme.

Theorem 4. *Let* $n\rho_n \geq a_n$ *where* $(a_n : n \geq 1)$ *is known to both the encoder and the decoder, with* $a_n \to \infty$ *as* $n \to \infty$. *Then, we can choose parameters* $(\Delta_n : n \geq 1)$ *such that the compression scheme* $\left(\left(f_n^{\Delta_n}, g_n^{\Delta_n}\right) : n \geq 1\right)$ *achieves optimal universal compression, i.e. if* $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ *is a sequence of* W*-random graphs with target densities* ρ_n*, where* W *is a normalized* L^2 *graphon, with* $\binom{\bar{m}_n = n}{2\rho_n}$ *assuming that* $\rho_n \to 0$ *as* $n \to \infty$*, with probability 1 we have*

$$\limsup_{n \to \infty} \frac{\mathrm{nats}\left(f_n^{\Delta_n}\left(G^{(n)}\right)\right) - \bar{m}_n \log \frac{1}{\rho_n}}{\bar{m}_n} \leq 1 - \mathrm{Ent}(W). \quad (2)$$

The scheme that achieves the bound is given below.

5.3 Compression scheme

The scheme makes use of a splitting method, with *splitting parameter* Δ_n, to split $G^{(n)}$ into $G^{(n)}_{\Delta_n}$, which dominates when $G^{(n)}$ comes from the weak convergence regime from the previous sections, and $G_*^{(n)}$ dominates when $G^{(n)}$ comes from the sparse graphon framework. The compression of $G_{\Delta_n}^{(n)}$ is done similarly to that of linear sparse graphs as in Section 4. For $G_*^{(n)}$, we require *sparse graphon estimation*, which estimates W given a sequence of W-random graphs. Details are provided in [DA21]. The key idea that makes this extension possible is that it is possible to split the graph into a part which is taken care of by the existing sparse regime, and the remainder can be dealt using the new sparse graphon framework.

6 Conclusion

I have summarized three papers that deal with universal lossless compression of marked graphs. However, there were several other papers I had hoped to cover to make this a more complete representation of the developments in graphical compression. Time-constraints restrict them to merely concluding remarks. All of the above schemes assume that the entire data is available with a single entity, which may not always be the case – hence [DA18] discusses *distributed compression*. Further, the schemes detailed in this report have all been asymptotic; [BS22] achieves fixed-length, universal, but *lossy* compression, which allows for some inaccuracy locally around each vertex. Almost posing as a converse problem to the compression of sparse graphs, [TMS18] looks at the compression of graphs that are generated by *duplication models*, where one picks random vertices from existing graphs and clone them – here there is redundant data hiding in these duplications – this is exploited in compressing unmarked graphs with vertices that may be labelled or unlabelled. Since graphs have innumerable applications, there is no dearth of possible directions to probe.

References

[DA20a] Payam Delgosha and Venkat Anantharam. "Universal Lossless Compression of Graphical Data". In: *IEEE Transactions on Information Theory* 66.11 (2020), pp. 6962–6976. DOI: 10.1109/TIT.2020.2991384.

[DA20b] Payam Delgosha and Venkat Anantharam. "A Universal Low Complexity Compression Algorithm for Sparse Marked Graphs". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. 2020, pp. 2349–2354. DOI: 10.1109/ISIT44484.2020.9174300.

[DA21] Payam Delgosha and Venkat Anantharam. "A Universal Lossless Compression Method applicable to Sparse Graphs and heavy-tailed Sparse Graphs". In: *2021 IEEE International Symposium on Information Theory (ISIT)*. 2021, pp. 2792–2797. DOI: 10.1109/ISIT45174.2021.9517897.

[BC14] Charles Bordenave and Pietro Caputo. "Large deviations of empirical neighborhood distribution in sparse random graphs". In: *Probability Theory and Related Fields* 163.1-2 (Nov. 2014), pp. 149–222. DOI: 10.1007/s00440-014-0590-8. URL: https://doi.org/10.1007%2Fs00440-014-0590-8.

[DA18] Payam Delgosha and Venkat Anantharam. "Distributed Compression of Graphical Data". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. 2018, pp. 2216–2220. DOI: 10.1109/ISIT.2018.8437614.

[BS22] Ronit Bustin and Ofer Shayevitz. "On Lossy Compression of Directed Graphs". In: *IEEE Transactions on Information Theory* 68.4 (2022), pp. 2101–2122. DOI: 10.1109/TIT.2021.3132290.

[TMS18] Krzysztof Turowski, Abram Magner, and Wojciech Szpankowski. "Compression of Dynamic Graphs Generated by a Duplication Model". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2018, pp. 1089–1096. DOI: 10.1109/ALLERTON.2018.8636034.