Supplementary: Semi-supervised clustering for de-duplication

Shrinu Kushagra
University of Waterloo

Shai Ben-David
University of Waterloo

Ihab F. Ilyas
University of Waterloo

1 Definitions

Definition 1 (Informative metric)**.** *Given (X, d), a clustering C^* and a parameter λ. We say that the metric d is (α, β)-informative w.r.t C^* and λ if*

$$\mathop{\mathbf{P}}_{(x,y)\sim U^2} \big[d(x,y) > \lambda \mid C^*(x,y) = 1\big] \;\leq\; \alpha \qquad (1)$$

$$\mathop{\mathbf{P}}_{(x,y)\sim U^2} \big[C^*(x,y) = 1 \mid d(x,y) \leq \lambda\big] \;\geq\; \beta \qquad (2)$$

Here U^2 is the uniform distribution over $X^{[2]}$.

2 Hardness of PCC in the presence of an oracle

Theorem 2. *Given that the Exponential Time Hypothesis (ETH) holds then any algorithm for the Promise Correlation Clustering problem that runs in polynomial time makes $\Omega(|X|)$ same-cluster queries for all $M \geq 3$ and for $\alpha = 0$ and $\beta = \frac{1}{2}$.*

The exponential time hypothesis says that any solver for 3-SAT runs in $2^{o(m)}$ time (where m is the number of clauses in the 3-SAT formula). We use a reduction from 3-SAT to 3DM to X3C to show that the exact cover by 3-sets (X3C) problem also can't be solved in $2^{o(m)}$ time (if ETH holds). Then, using the reduction from the previous section implies that PCC also can't be solved in $2^{o(n)}$ time. Thus, any query based algorithm for PCC needs to make atleast $\Omega(n)$ queries where $n = |X|$ is the number of vertices in the graph.

Definition 3 (3-SAT)**.** .
Input: A boolean formulae ϕ in 3CNF with n literals and m clauses. Each clause has exactly three literals.
Output: YES if ϕ is satisfiable, NO otherwise.

Exponential Time Hypothesis
There does not exist an algorithm which decides 3-SAT and runs in $2^{o(m)}$ time.

Definition 4 (3DM)**.** .
Input: Sets W, X and Y and a set of matches $M \subseteq W \times X \times Y$ of size m.
Output: YES if there exists $M' \subseteq M$ such that each element of W, X, Y appears exactly once in M'. NO otherwise.

To prove that (X3C) is NP-Hard, the standard We will reduce 3-SAT to 3-dimensional matching problem. 3DM is already known to be NP-Hard. However, the standard reduction of 3-SAT to 3DM constructs a set with $|M| \in \Theta(m^2n^2)$. Hence, using the standard reduction, the exponential time hypothesis would imply there does not exist an algorithm for 3DM which runs in $\Omega(m^{\frac{1}{4}})$. Our reduction is based on the standard reduction. However, we make some clever optimizations especially in the way we encode the clauses. This helps us improve the lower bound to $\Omega(m)$.

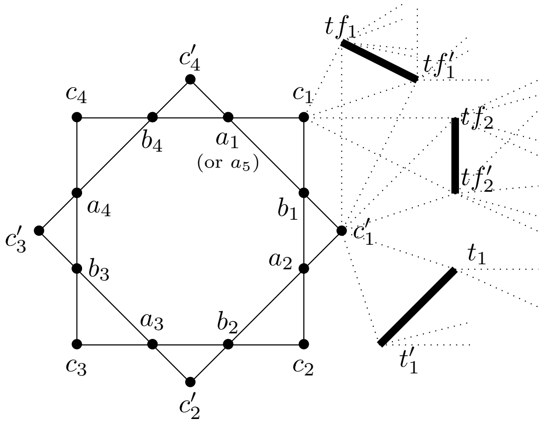

Figure 1: Part of graph G constructed for the literal x_1. The figure is an illustration for when x_1 is part of four different clauses. The triangles (or hyper-edge) (a_i, b_i, c_i) capture the case when x_1 is true and the other triangle (b_i, c'_i, a_{i+1}) captures the case when x_1 is false. Assuming that a clause $C_j = \{x_1, x_2, x_3\}$, the hyper-edges containing tf_i, tf'_i and t_1, t'_1 capture different settings. The hyper-edges containing t_1, t'_1 ensure that atleast one of the literals in the clause is true. The other two ensure that two variables can take either true or false values.

Our gadget is described in Fig. 1. For each literal x_i, let m_i be the number of clauses in which the the literal is present. We construct a "truth-setting" component containing $2m_i$ hyper-edges (or triangles). We add the following hyper-edges to M.

$$\{(a_k[i], b_k[i], c_k[i]) : 1 \leq k \leq m_i\} \\ \cup \{(a_{k+1}[i], b_k[i], c'_k[i]) : 1 \leq k \leq m_i\}$$

Note that one of (a_k, b_k, c_k) or (a_{k+1}, b_k, c'_k) have to be selected in a matching M'. If the former is selected that corresponds to the variable x_i being assigned true, the latter corresponds to false. This part is the same as the standard construction.

For every clause $C_j = \{x_1, x_2, x_3\}$ we add three types of hyper-edges. The first type ensures that atleast one of the literals is true.

$$\{(c_k[i], t_1[j], t'_1[j]) : x'_i \in C_j\} \cup \{(c'_k[i], t_1[j], t'_1[j]) : x_i \in C_j\}$$

The other two types of hyper-edges (conected to the tf_i's) say that two of the literals can be either true or false. Hence, we connect them to both c_k and c'_k

$$\{(c_k[i], tf_1[j], tf'_1[j]) : x'_i \text{ or } x_i \in C_j\} \\ \cup \{(c_k[i], tf_2[j], tf'_2[j]) : x_i \text{ or } x'_i \in C_j\} \\ \cup \{(c'_k[i], tf_1[j], tf'_1[j]) : x'_i \text{ or } x_i \in C_j\} \\ \cup \{(c'_k[i], tf_2[j], tf'_2[j]) : x_i \text{ or } x'_i \in C_j\}$$

Note that in the construction k refers to the index of the clause C_j in the truth-setting component corresponding to the literal x_i. Using the above construction, we get that

$$W = \{c_k[i], c'_k[i]\} \\ X = \{a_k[i]\} \cup \{t_1[j], tf_1[j], tf_2[j]\} \\ Y = \{b_k[i]\} \cup \{t'_1[j], tf'_1[j], tf'_2[j]\}$$

Hence, we see that $|W| = 2\sum_i m_i = 6m$. Now, $|X| = |Y| = \sum_i m_i + 3m = 6m$. And, we have that $|M| = 2\sum_i m_i + 15m = 21m$. Thus, we see that this construction is linear in the number of clauses.

Now, if the 3-SAT formula ϕ is satisfiable then there exists a matching M' for the 3DM problem. If a variable $x_i = T$ in the assignment then add $(c_k[i], a_k[i], b_k[i])$ to M' else add $(c'_k[i], a_{k+1}[i], b_k[i])$. For every clause C_j, let x_i (or x'_i) be the variable which is set to true in that clause. Add $(c'_k[i], t_1[j], t'_1[j])$ (or $(c_k[i], t_1[j], t'_1[j])$) to M'. For the remaining two clauses, add the hyper-edges containing $tf_1[j]$ and $tf_2[j]$ depending upon their assignments. Clearly, M' is a matching.

Now, the proof for the other direction is similar. If there exists a matching, then one of (a_k, b_k, c_k) or (a_{k+1}, b_k, c'_k) have to be selected in a matching M'. This defines a truth assignment of the variables. Now, the construction of the clause hyper-edges ensures that every clause is satisfiable.

Theorem 5. *If the exponential time hypothesis holds then there does not exist an algorithm which decides the three dimensional matching problem 3DM and runs in time $2^{o(m)}$.*

Corollary 6. *If the exponential time hypothesis holds then there does not exist an algorithm which decides exact cover by 3-sets problem (X3C) and runs in time $2^{o(m)}$.*

Hence, from the discussion in this section, we know that X3C is not only NP-Hard but the running time is lower bounded by $\Omega(2^m)$. Now, using the same reduction of X3C to PCC as before, gives the same lower bound on the running time of PCC. Using this, we can now lower bound the number of queries required by PCC.

For the sake of contradiction, let us assume that there exists an algorithm which solves PCC in polynomial time by making $o(n)$ same-cluster queries (n is the number of vertices). Then by simulating all possible answers for the oracle, we get a non-query algorithm which solves PCC in $2^{o(n)}$. However, combining Cor. 6 with the reduction of X3C to PCC, we get that any algorithm that solves PCC takes $\Omega(2^n)$. Hence, no such query algorithm exists.

3 Sampling positive pairs

Lemma 7. *Given set (X, d), a C^*-oracle and parameter λ. Let d be (α, β)-informative w.r.t λ and C^*. Then the sampling procedure $\mathcal{P}_1$ induces a distribution T over $X^{[2]}$ such that for any labelling function h over $X^{[2]}$ we have that*

$$\left| \mathbf{P}_{(x,y)\sim P^+} \left[h(x,y) = 0\right] - \mathbf{P}_{(x,y)\sim T} \left[h(x,y) = 0\right] \right| \leq 2\alpha.$$

Proof. Let $K = \{(x,y) : d(x,y) \leq \lambda\}$ and D be a distribution over K defined by $D(x,y) := \frac{|S_x|}{\sum_{x'} |S_{x'}|} \cdot \frac{1}{|S_x|} = \frac{U^2(x,y)}{U^2(K)}$. Let $K^+ = \{(x,y) : d(x,y) \leq \lambda \text{ and } C^*(x,y) = 1\}$. Let T be the distribution induced by $\mathcal{P}_1$. It's easy to see that for $(x,y) \notin K^+$, $T(x,y) = 0$. For $(x,y) \in K^+$, let $D(x,y) = p$ and $D(K^+) = q$. Then, $T(x,y) = p + (1-q)p + \ldots = \frac{p}{q} = \frac{D(x,y)}{D(K^+)} = \frac{U^2(x,y)}{U^2(K^+)}$. Using Defn. 1, we know that

$$1 - \alpha \leq \mathbf{P}_{(x,y)\sim U^2}[d(x,y) \leq \lambda \mid C^*(x,y) = 1] \\ = \frac{\mathbf{P}_{(x,y)\sim U^2}[d(x,y) \leq \lambda, C^*(x,y) = 1]}{\mathbf{P}_{(x,y)\sim U^2}[C^*(x,y) = 1]} = \frac{U^2(K^+)}{U^2(X^{[2]+})} \tag{3}$$

Now, we will use the above inequality to prove our result.

$$\begin{aligned}
&\mathop{\mathbf{P}}_{(x,y)\sim T}\big[h(x,y)=0\big] = \sum_{(x,y)\in K^+} T(x,y)\mathbf{1}_{h(x,y)=0}\\
&= \sum_{(x,y)\in K^+} \frac{U^2(x,y)}{U^2(K^+)}\mathbf{1}_{h(x,y)=0}\\
&\le \frac{1}{1-\alpha}\sum_{(x,y)\in K^+}\frac{U^2(x,y)}{U^2(X^{[2]+})}\mathbf{1}_{h=0}\\
&\le (1+2\alpha)\sum_{(x,y)\in X_+^2} P^+(x,y)\mathbf{1}_{h(x,y)=0}\\
&= (1+2\alpha)\mathop{\mathbf{P}}_{(x,y)\sim P^+}\big[h(x,y)=0\big]
\end{aligned}$$

Now, for the other direction, we have that

$$\begin{aligned}
&\mathop{\mathbf{P}}_{(x,y)\sim P^+}\big[h(x,y)=0\big] = \sum_{(x,y):X^{[2]+}} P^+(x,y)\mathbf{1}_{h(x,y)=0}\\
&= \sum_{(x,y)\in K^+} \frac{U^2(x,y)}{U^2(X^{[2]+})}\mathbf{1}_{h(x,y)=0}\\
&\qquad + \sum_{(x,y)\in X_+^2\setminus K^+} \frac{U^2(x,y)}{U^2(X^{[2]+})}\mathbf{1}_{h=0}\\
&\le \sum_{(x,y)\in K^+} \frac{U^2(x,y)}{U^2(K^+)}\mathbf{1}_{h(x,y)=0}\\
&\qquad + \sum_{(x,y)\in X^{[2]+}\setminus K^+} \frac{U^2(x,y)}{U^2(X^{[2]+})}\mathbf{1}_{h=0}\\
&\le \mathop{\mathbf{P}}_{(x,y)\sim T}\big[h(x,y)=0\big] + \sum_{(x,y)\in X^{[2]+}\setminus K^+} \frac{U^2(x,y)}{U^2(X^{[2]+})}\\
&\le \mathop{\mathbf{P}}_{(x,y)\sim T}\big[h(x,y)=0\big] + \alpha
\end{aligned}$$

Hence, we have shown that both the directions hold and this completes the proof of the lemma. Note that this shows that our sampling procedure approximates the distribution P^+. It is easy to see that pre-computing S_x for all x takes $|X|^2$ time. Once the pre-computation is done, the sampling can be done in constant time. □

4 Sample and query complexity of RCC

Theorem 8. *Given metric space (X,d), a class of clusterings $\mathcal{F}$ and a threshold parameter λ. Given $\epsilon,\delta \in (0,1)$ and a C^*-oracle. Let d be (α,β)-informative and X be γ-skewed w.r.t λ and C^*. Let $\mathcal{A}$ be the ERM-based approach as described in Alg.* **??** *and $\hat{C}$ be the output of $\mathcal{A}$. If*

$$m_-, m_+ \;\; \ge a\frac{\text{VC-Dim}(\mathcal{F})+\log(\frac{2}{\delta})}{\epsilon^2} \qquad (4)$$

where a is a global constant then with probability atleast $1-\delta$ (over the randomness in the sampling procedure), we have that

$$L_{C^*}(\hat{C}) \;\;\le\;\; \min_{\mathcal{C}\in\mathcal{F}} L_{C^*}(\mathcal{C}) + 3\alpha + \epsilon$$

Proof. Let T_0 be the distribution induced by $\mathcal{P}_0$ and T_1 be the distribution induced by $\mathcal{P}_1$. Denote by $E(h) = \mathop{\mathbf{P}}_{(x,y)\sim P^+}\big[h(x,y) = 0\big]$ and by $G(h) = \mathop{\mathbf{P}}_{(x,y)\sim P^-}\big[h(x,y)=1\big]$.

Using Thm. 16, we know that if $m_+ > a\frac{\text{VC-Dim}(\mathcal{F})+\log(\frac{1}{\delta})}{\epsilon^2}$ then with probability atleast $1-\delta$, we have that for all h

$$\begin{aligned}
&|\hat{E}(h) - \mathop{\mathbf{P}}_{(x,y)\sim T_1}[h(x,y)=0]| \le \epsilon\\
\implies &\hat{E}(h) \le \epsilon + \mathop{\mathbf{P}}_{(x,y)\sim T_1}[h(x,y)=0] \le \epsilon + 2\alpha + E(h) \quad \text{and}\\
&E(h) - 2\alpha - \epsilon \le \hat{E}(h) \qquad (5)
\end{aligned}$$

Note that we obtain upper and lower bounds for $\mathop{\mathbf{P}}_{(x,y)\sim T_1}[h(x,y) = 0]$ using Lemma 7. Similarly, if $m_- > a\frac{\text{VC-Dim}(\mathcal{F})+\log(\frac{1}{\delta})}{\epsilon^2}$, then with probability atleast $1-\delta$, we have that for all h,

$$\begin{aligned}
&|\hat{G}(h) - \mathop{\mathbf{P}}_{(x,y)\sim T_0}[h(x,y)=1]| \le \epsilon\\
\implies &\hat{G}(h) \le \epsilon + G(h) \quad \text{and} \quad G(h) - \epsilon \le \hat{G}(h) \qquad (6)
\end{aligned}$$

Combining Eqns. 5 and 6, we get that with probability atleast $1-2\delta$, we have that for all $C \in \mathcal{F}$

$$\begin{aligned}
&\hat{L}(C) \le \mu[\epsilon + E(h) + 2\alpha] + (1-\mu)(\epsilon + G(h))\\
&\quad \le L(h) + \epsilon + 2\alpha\\
&\text{And } \hat{L}(C) \ge \mu(E(h) - \epsilon - \alpha) + (1-\mu)(G(h) - \epsilon)\\
&\quad \ge L(h) - \epsilon - \alpha
\end{aligned}$$

Now, let $\hat{C}$ be the output of $\mathcal{A}$ and let $\hat{C}^*$ be $\arg\min_{C\in\mathcal{F}} L(C)$. Then, we have that with probability atleast $1-2\delta$

$$L(\hat{C}) \le \hat{L}(\hat{C}) + \alpha + \epsilon \le \hat{L}(\hat{C}^*) + \alpha + \epsilon \le L(\hat{C}^*) + 2\epsilon + 3\alpha$$

Choosing $\epsilon = \frac{\epsilon}{2}$ and $\delta = \frac{\delta}{2}$ throughout gives the result of the theorem. □

Theorem 9. *[Query Complexity] Let the framework be as in Thm. 8. With probability atleast $1 - \exp\big(-\frac{\nu^2 m_-}{4}\big) - \exp\big(-\frac{\nu^2 m_+}{4}\big)$ over the randomness in the sampling procedure, the number of same-cluster queries q made by $\mathcal{A}$ is*

$$q \le (1+\nu)\bigg(\frac{m_-}{(1-\gamma)} + \frac{m_+}{\beta}\bigg)$$

Proof. Let q_+ denote the number queries to sample the set S_+. We know that $\mathbf{E}[q_+] \le \frac{1}{\beta}$. Given that the expectation is bounded as above, using Thm. 17, we get that $q_+ \le \frac{(1+\nu)m_+}{\beta(1-\epsilon)}$ with probability atleast $1 - \exp(\frac{-\nu^2 m_+}{4})$. Similarly, we get that with probability atleast $1 - \exp(\frac{-\nu^2 m_-}{4})$, $q_- \le \frac{(1+\nu)m_-}{(1-\gamma)}$. □

4.1 VC-dimension of common classes

Theorem 10. *Given a finite set $\mathcal{X}$ and a finite class $\mathcal{F} = \{C_1, \ldots, C_s\}$ of clusterings of $\mathcal{X}$.*

$$\text{VC-Dim}(l_{\mathcal{F}}) \le g(s)$$

where $g(s)$ is the smallest integer n such that $B_{\sqrt{n}} \ge s$ where B_i is the i^{th} bell number [A000108,].

Proof. Let n be as defined in the statement of the theorem. Let $M^2 \subseteq \mathcal{X}^2$ be a set of size $> n$. Define $M := \{x : (x,y) \in M^2 \text{ or } (y,x) \in M^2\}$. We know that $|M| > \sqrt{n}$. The number of clusterings (partitions) on n elements is given by the n^{th} bell number. Thus, for $s \le B_{\sqrt{n}}$ there exists a clustering $C' \notin \mathcal{F}$ of the set $\mathcal{X}$. Hence, $l_{\mathcal{F}}$ can't shatter any set of size $> n$. □

Lemma 11. *Let $\mathcal{X}$ be a finite set, $S \subseteq \mathcal{X}$ be a set of n points and T be any hierarchical clustering tree of $\mathcal{X}$. There exists a set $\mathcal{C} = \{C_1, \ldots, C_s\}$ where each C_i is a clustering of S with the following properties*

- $|\mathcal{C}| \ge \frac{n!}{\lfloor n/2 \rfloor !\ 2^{\lfloor n/2 \rfloor}}$
- *T contains atmost one clustering from $\mathcal{C}$.*

Proof. Consider clusterings C_i of S of the following type. Each cluster in C_i contains exactly two points (except possibly one cluster which contains one point if n is odd). One such clustering along with a tree $T$is shown in Fig. 2. Let $\mathcal{C}$ be the set of all such clusterings C_i. The number of such clusterings $|\mathcal{C}|$ is

$$\begin{cases} \dfrac{n!}{2^{\frac{n-1}{2}}\frac{n-1}{2}!} & n \text{ is odd} \\ \dfrac{n!}{2^{\frac{n}{2}}\frac{n}{2}!} & n \text{ is even} \end{cases} = \frac{n!}{2^{\lfloor \frac{n}{2} \rfloor}(\lfloor \frac{n}{2} \rfloor)!}$$

For the sake of contradiction, assume that T is a hierarchical clustering tree T of $\mathcal{X}$ which contains C_i and C_j. Since $C_i \neq C_j$, there exists points s_1, s_2 and s_3 such that the following happens. (i) s_1, s_2 are in the same cluster in C_i. s_2, s_3 as well as s_1, s_3 are in different clusters in C_i. (ii) s_1, s_3 are in the same cluster in C_j. s_2, s_3 as well as s_1, s_2 are in different clusters in C_j.

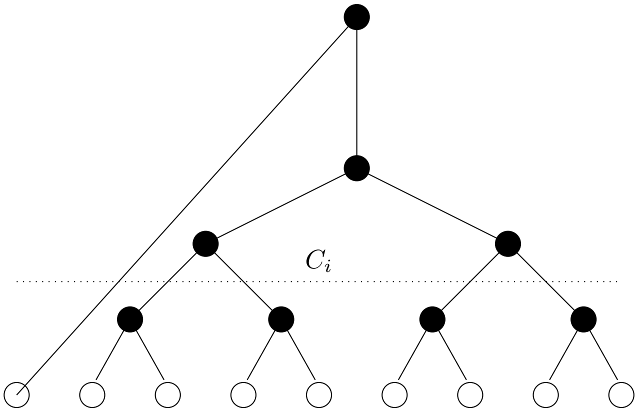

Figure 2: A hierarchical clustering tree of $n = 9$ points. This tree contains the clustering C_i described in the proof of Lemma 11.

Now, T contains C_i. Hence, there exists a node v such that $s_1, s_2 \in C(v)$ but $s_3 \notin C(v)$. T also contains C_j. Hence, there exists a node u such that $s_1, s_3 \in C(u)$ and $s_2 \notin C(u)$. Both u and v contain the point s_1. Hence, either u is a descendant of v or the other way around. Observe that $s_2 \in C(v)$ but $s_2 \notin C(u)$. Hence, v is not a descendant of u. Similarly, $s_3 \in C(u)$ and $s_3 \notin C(v)$ so u is not a descendant of v. This leads to a contradiction. Hence, no such tree T can exist. □

Theorem 12. *Given a finite set $\mathcal{X}$ and a finite class $\mathcal{F} = \{T_1, \ldots, T_s\}$ where each T_i is a hierarchical clustering over $\mathcal{X}$. Then*

$$\text{VC-Dim}(\mathcal{F}) \le g(s)$$

where $g(s)$ is the smallest integer n such that $\frac{\sqrt{n}!}{\lfloor \sqrt{n}/2 \rfloor !\ 2^{\lfloor \sqrt{n}/2 \rfloor}} \ge s$

Proof. Let n be as defined in the statement of the theorem. Let $M^2 \subseteq \mathcal{X}^2$ be a set of size $> n^2$. Define $M := \{x : (x,y) \in M^2 \text{ or } (y,x) \in M^2\}$. We know that $|M| > n$. Using lemma 11, there exists a set of clusterings $\mathcal{C} = \{C_1, \ldots, C_{s'}\}$ of size $s' > \frac{n!}{\lfloor n/2 \rfloor !\ 2^{\lfloor n/2 \rfloor}} \ge s$ such that each $T_i \in \mathcal{F}$ contains atmost one $C_j \in \mathcal{C}$. Thus, there exists a clustering C_j which is not captured by any $T_i \in \mathcal{F}$. Hence, $l_{\mathcal{F}}$ can't shatter any set of size $> n^2$. □

References

[A000108,] A000108, S. The on-line encyclopedia of integer sequences. *published electronically at https://oeis.org, 2010.*

[Blumer et al., 1989] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.

[Brown, 2011] Brown, D. G. (2011). How i wasted too long finding a concentration inequality for sums of geometric variables. *Found at https://cs. uwaterloo. ca/˜ browndg/negbin. pdf*, 6.

[Mitzenmacher and Upfal, 2005] Mitzenmacher, M. and Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[Vapnik and Chervonenkis, 2015] Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer.

A Technical lemmas and theorems

Theorem 13 (Multiplicative Chernoff bound [Mitzenmacher and Upfal, 2005])**.** *Let $X_1, \ldots, X_n$ be i.i.d random variables in $\{0, 1\}$ such that $\mu = E[X_i]$. Let $X = \frac{\sum X_i}{n}$. Then for any $0 < \epsilon < 1$*

$$P\big[\ X > (1+\epsilon)\mu\ \big] \ \ \le\ \ \exp\left(\frac{-\epsilon^2\mu n}{3}\right)$$

Theorem 14 (Multiplicative Chernoff bound [Mitzenmacher and Upfal, 2005])**.** *Let $X_1, \ldots, X_n$ be i.i.d random variables in $\{0, 1\}$ such that $\mu = E[X_i]$. Let $X = \frac{\sum X_i}{n}$. Then for any $0 < \epsilon < 1$*

$$P\big[\ X < (1-\epsilon)\mu\ \big] \ \ \le\ \ \exp\left(\frac{-\epsilon^2\mu n}{2}\right)$$

Theorem 15 (Vapnik and Chervonenkis [Vapnik and Chervonenkis, 2015])**.** *Let X be a domain set and D a probability distribution over X. Let H be a class of subsets of X of finite VC-dimension d. Let $\epsilon, \delta \in (0, 1)$. Let $S \subseteq X$ be picked i.i.d according to D of size m. If $m > \frac{c}{\epsilon^2}(d \log \frac{d}{\epsilon} + \log \frac{1}{\delta})$, then with probability $1 - \delta$ over the choice of S, we have that $\forall h \in H$*

$$\left|\frac{|h \cap S|}{|S|} - P(h)\right| < \epsilon$$

Theorem 16 (Fundamental theorem of learning [Blumer et al., 1989])**.** *Here, we state the theorem as in the book [Shalev-Shwartz and Ben-David, 2014]. Let H be a class of functions $h : \mathcal{X} \to \{0, 1\}$ of finite VC-Dimension, that is* $\text{VC-Dim}(H) = d < \infty$. *Let D be a probability distribution over X and h^* be some unknown target function. Given $\epsilon, \delta \in (0, 1)$. Let err_D be the $\{0, 1\}$-loss function $err : H \to [0, 1]$. That is $err_D(h) = \underset{x \in D}{\mathbf{P}}[h(x) \neq h^*(x)]$. Sample a set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ according to the distribution D. Define $err_S(h) = \sum_{i=1}^m \frac{\mathbf{1}_{[h(x_i) \neq h^*(x_i)]}}{m}$. If $m \ge a\frac{d + \log(1/\delta)}{\epsilon^2}$, then with probability atleast $1 - \delta$ over the choice of S, we have that for all $h \in H$*

$$|err_D(h) - err_S(h)| \le \epsilon$$

where a is an absolute global constant.

Theorem 17 (Concentration inequality for sum of geometric random variables [Brown, 2011])**.** *Let $X = X_1 + \ldots + X_n$ be n geometrically distributed random variables such that $\mathbf{E}[X_i] = \mu$. Then*

$$\mathbf{P}[X > (1+\nu)n\mu] \le \exp\left(\frac{-\nu^2\mu n}{2(1+\nu)}\right)$$