A Bad Instance for k-Means++[⋆]

Tobias Brunsch and Heiko Röglin

Department of Computer Science, University of Bonn, Germany
brunsch@cs.uni-bonn.de, heiko@roeglin.org

Abstract. k-means++ is a seeding technique for the k-means method with an expected approximation ratio of $O(\log k)$, where k denotes the number of clusters. Examples are known on which the expected approximation ratio of k-means++ is $\Omega(\log k)$, showing that the upper bound is asymptotically tight. However, it remained open whether k-means++ yields an $O(1)$-approximation with probability $1/\text{poly}(k)$ or even with constant probability. We settle this question and present instances on which k-means++ achieves an approximation ratio of $(2/3 - \varepsilon) \cdot \log k$ only with exponentially small probability.

1 Introduction

In the *k-means problem* we are given a set of *data points* $X \subseteq \mathbb{R}^d$ and the objective is to group these points into k mutually disjoint *clusters* $C_1, \ldots, C_k \subseteq X$. Each of those clusters should contain only 'similar points' that are close together in terms of Euclidean distance. In order to evaluate the quality of a clustering, we assign a *cluster center* $c_i \in \mathbb{R}^d$ to each cluster C_i and consider the potential $\Phi = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$. The goal of the k-means problem is to find clusters and cluster centers that minimize this potential.

Aloise et al. showed that the k-means problem is $\mathcal{NP}$-hard, even for $k = d = 2$ [2]. To deal with this problem in practice, several heuristics have been developed over the past decades. Probably the "most popular" one [5] is Lloyd's algorithm [7], usually called the *k-means method* or simply *k-means*. Starting with k arbitrary cluster centers, each data point is assigned to its nearest center. In the next step each center is recomputed as the center of mass of the points assigned to it. This procedure is repeated until the centers remain unchanged.

Though Vattani showed that the running time of k-means can be exponential in the number of input points [8], speed is one of the most important reasons for its popularity in practice. This unsatisfying gap between theory and practice was narrowed by Arthur, Manthey, and Röglin who showed that the running time of k-means is polynomially bounded in the model of smoothed analysis [3].

Another problem is that k-means may yield poor results if the initial centers are badly chosen. The approximation ratio can be arbitrarily large, even for small input sets and $k = 2$. To improve the quality of the solutions found by the k-means method, Arthur and Vassilvitskii proposed the following seeding technique called k-means++, which has an expected approximation ratio of $O(\log k)$ [4].

[⋆] A part of this work was done at Maastricht University and was supported by a Veni grant from the Netherlands Organisation for Scientific Research.

M. Ogihara and J. Tarui (Eds.): TAMC 2011, LNCS 6648, pp. 344–352, 2011.

1. Choose center c_1 uniformly at random from the input set X.
2. For $i = 2$ to k do:
   Let $D_i^2(x)$ be the square of the distance between point x and the nearest already chosen center $c_1, \dots, c_{i-1}$. Choose the next center c_i randomly from X, where every $x \in X$ has a probability of $\frac{D_i^2(x)}{\sum_{y \in X} D_i^2(y)}$ of being chosen.

Previous work. In [4] instances are given on which k-means++ yields in expectation an $\Omega(\log k)$-approximation, showing that the bound of $O(\log k)$ for the approximation ratio of k-means++ is asymptotically tight. However, the expected approximation ratio of a heuristic is not the only useful quality criterion if the variance is large. If, for example, an $O(1)$-approximation is obtained with probability $1/\mathrm{poly}(k)$, then after a polynomial number of restarts an $O(1)$-approximation is reached with high probability even if the expected approximation ratio is $\Omega(\log k)$. Interestingly, k-means++ achieves a constant factor approximation for the instance given in [4] with constant probability, and no instance was known on which k-means++ does not yield an $O(1)$-approximation with constant probability.

For this reason Aggarwal, Deshpande, and Kannan called the lower bound of $\Omega(\log k)$ "misleading" [1]. In the same paper they showed that sampling $O(k)$ instead of k centers with the k-means++ seeding technique and selecting k good points among them yields an $O(1)$-approximation with constant probability. Unfortunately the selection step is done with LP-based algorithms, which makes this approach less simple and efficient than k-means++ in practice.

Therefore, both Arthur and Vassilvitskii [4] and Aggarwal, Deshpande, and Kannan [1] raise the question whether k-means++ yields an $O(1)$-approximation with constant probability. Aggarwal et al. call this a "tempting conjecture" which "would be nice to settle". So far the only known result in this direction is due to Arthur and Vassilvitskii who mention that the probability to achieve an $O(1)$-approximation is at least $c \cdot 2^{-k}$ for some constant $c > 0$ [4].

Our contribution We modify the instances given in [4] and show that it is very unlikely that k-means++ achieves an approximation ratio of $(2/3 - \varepsilon) \cdot \log k$ on this modified example.

Theorem 1. *Let $r\colon \mathbb{N} \to \mathbb{R}^+$ be a real function.*

1. *If $r(k) = \delta^* \cdot \ln(k)$ for a fixed real $\delta^* \in (0, 2/3)$, then there is a class of instances on which k-means++ achieves an $r(k)$-approximation with probability at most $\exp(-k^{1-3/2 \cdot \delta^* - o(1)})$.*
2. *If $r = o(\log k)$, then there is a class of instances on which k-means++ achieves an $r(k)$-approximation with probability at most $\exp(-k^{1-o(1)})$.*

2 Construction and Analysis of a Bad Instance

2.1 Construction

Throughout the paper "log" denotes the natural logarithm. Let $r = r(k) > 0$ be a function where $r(k) = \delta^* \cdot \log k$ for a fixed real $\delta^* \in (0, 2/3)$ or $r = o(\log k)$.

Without loss of generality let $r(k) \to \infty$ in the latter case. Additionally, let $\delta = \delta(k) := r(k)/\log k$ be the ratio of $r(k)$ and $\log k$. Based on the function r, we introduce a parameter $\Delta = \Delta(k)$. In Section 2.3 we describe the details of this choice. In this section we present the instances used for proving Theorem 1, which are a slight modification of the instances given in [4].

We first choose k centers $c_1, \dots, c_k$, each with squared distance $\Delta^2 - (k-1)/k$ to each other. For each point c_i we construct a regular $(k-1)$-simplex with center c_i and with side length 1. We denote the vertices of this simplex by $x_1^{(i)}, \dots, x_k^{(i)}$, and we assume that the simplices for different points c_i and $c_{i'}$ are constructed in orthogonal dimensions. Then we get

$$\|x_j^{(i)} - c_i\|^2 = \frac{k-1}{2k}\,, \tag{1}$$

and for $x_j^{(i)} \neq x_{j'}^{(i')}$ we get

$$\|x_j^{(i)} - x_{j'}^{(i')}\|^2 = \begin{cases} 1 & : \quad i = i'\,, \\ \Delta^2 & : \quad i \neq i'\,, \end{cases} \tag{2}$$

due to the fact that for $i \neq i'$ the squared distance between $x_j^{(i)}$ and $x_{j'}^{(i')}$ is

$$\|x_j^{(i)} - x_{j'}^{(i')}\|^2 = \|x_j^{(i)} - c_i\|^2 + \|c_i - c_{i'}\|^2 + \|x_{j'}^{(i')} - c_{i'}\|^2 = \Delta^2$$

because of orthogonality and Equation (1). Let $C_i = \{x_1^{(i)}, \dots, x_k^{(i)}\}$ for $i = 1, \dots, k$. As input set for our k-means problem we consider the union $X = \bigcup_{i=1}^k C_i$ of these sets.

In the remainder we show that, with a good choice of Δ, X is a bad instance for `k-means++`. Note that the only relevant difference to the example given in [4] is the choice of Δ. While in [4] it was sufficient to choose Δ large enough, we have to tune Δ much more carefully to prove Theorem 1.

2.2 Reduction to a Markov Chain

We consider the k-clustering $C^* = (C_1, \dots, C_k)$ induced by the centers $c_1, \dots, c_k$. Note that for small Δ this might be a non-optimal solution, but its potential is an upper bound for the optimal potential. Due to Equation (1), the potential Φ^* of C^* is

$$\Phi^* = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 = k^2 \cdot \frac{k-1}{2k} \leq \frac{k^2}{2}$$

as for any point of C_i the nearest center is c_i. Now let C' be a clustering with distinct centers $c'_1, \dots, c'_t$, $1 \leq t \leq k$, chosen from X. For each center c'_i let l_i be the index of the set C_{l_i} that c'_i belongs to. Let $s := |\{l_1, \dots, l_t\}|$ denote the number of *covered* sets C_i and let $X_u := X \setminus \bigcup_{i=1}^t C_{l_i}$ denote the set of the points of *uncovered* sets. Furthermore, let Φ denote the potential of X induced by the centers $c'_1, \dots, c'_t$ and let $\Phi(X_u)$ be the part of Φ contributed by the uncovered sets. Applying Equations (1) and (2) we get $\Phi(X_u) = (k-s) \cdot k \cdot \Delta^2$ and $\Phi = (s \cdot k - t) \cdot 1^2 + \Phi(X_u) \geq (s-1) \cdot k + \Phi(X_u)$.

The inequality $\frac{\Phi}{\Phi^*} \leq r$ is necessary for C' being an r-approximation. This implies

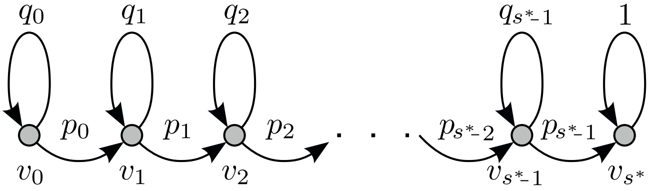
$$r \geq \frac{\Phi}{\Phi^*} \geq \frac{\Phi(X_u)}{\Phi^*} \geq \frac{2(k-s) \cdot \Delta^2}{k},$$

i.e. at least $s^* := \left\lceil k \cdot \left(1 - \frac{r}{2\Delta^2}\right) \right\rceil$ of the k sets C_i have to be *covered* to get an r-approximation.

Let us assume that we are in step 2 of `k-means++` (see introduction) and let s denote the number of covered sets C_i. The probability of covering an uncovered set in this step is

$$\frac{\Phi(X_u)}{\Phi} \leq \frac{\Phi(X_u)}{(s-1) \cdot k + \Phi(X_u)} = \frac{1}{1 + \frac{s-1}{(k-s) \cdot \Delta^2}} =: p_s\,. \tag{3}$$

Hence, we can upper-bound the probability that `k-means++` yields an r-approximation by the probability of reaching vertex v_{s^*} within k steps in the following Markov chain, starting from vertex v_0.



Here, p_s are the probabilities defined in Inequality (3), $p_0 = 1$ and $q_s = 1 - p_s$.

2.3 How to Choose Δ?

Arthur and Vassilvitskii [4] have shown that choosing Δ large enough results in instances on which `k-means++` has an expected approximation ratio of $\Omega(\log k)$. This does not suffice for proving Theorem 1 because if we choose Δ too large, the probability that we do not cover every cluster becomes small. Hence, if we choose Δ too large, we have a good probability of covering every cluster and thus of obtaining a constant-factor approximation. On the other hand, if we choose Δ too small, already a single covered cluster might suffice to obtain a constant-factor approximation.

We first define a function $\varepsilon\colon \mathbb{N} \to (0,1)$ as follows:

$$\varepsilon = \varepsilon(k) := \begin{cases} 1/3 & : \quad r = o(\log k)\,, \\ \frac{2}{3} \cdot \frac{\log r}{r} & : \quad r = \delta^* \cdot \log k\,. \end{cases}$$

Now we set $\tilde{\Delta} = \tilde{\Delta}(k) := \sqrt{r} \cdot e^{r \cdot (1+\varepsilon)/4} = \exp(\Theta(r))$ and $\Delta := \lceil \tilde{\Delta} \rceil$. In the analysis of the Markov chain in the following section, we will assume that k is chosen sufficiently large such that the following inequalities hold:

$$\Delta^2 > r\,, \tag{4}$$

$$\frac{\Delta^2}{k} \leq \frac{r}{2}\,, \tag{5}$$

$$\left(\frac{r+2}{2}\right)^{\Delta} \geq \left(\frac{2\Delta^2}{r}\right)^2, \tag{6}$$

$$\Delta^6 \leq \frac{19}{18} r^3 \cdot e^{3r(1+\varepsilon)/2}\,, \tag{7}$$

$$k - 1 \geq \left(1 - \frac{\varepsilon}{9}\right) \cdot k\,, \tag{8}$$

$$r + 2 \leq \left(1 + \frac{\varepsilon}{3}\right) \cdot r\,, \tag{9}$$

$$\frac{r}{2\Delta^2} + \frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right) \cdot \left(\frac{r}{2\Delta^2}\right)^2 \leq \left(\frac{\varepsilon}{3}\right)^2, \tag{10}$$

$$\log r \leq \frac{3}{2}\varepsilon \cdot r\,. \tag{11}$$

In the appendix we show that, for our choice of ε, Inequalities (4) to (11) are satisfied for every sufficiently large k. We have not made any attempt to simplify these inequalities as they appear in exactly this form in the analysis in the next section.

2.4 Analysis of the Markov Chain

Now we concentrate on bounding the probability to reach vertex v_{s^*} in the Markov chain above. For this we introduce geometrically distributed random variables $X_0, \ldots, X_{s^*-1}$. Variable X_s describes the number of trials that are required to move from vertex v_s to vertex v_{s+1}. We would like to show that the expected value of $X := \sum_{s=0}^{s^*-1} X_s$ is much greater than k and then conclude that it is unlikely to reach v_{s^*} within k steps. Unfortunately, Hoeffding's Inequality [6] which is often used for drawing such a conclusion requires random variables with bounded domain. So we make a technical detour by introducing additional random variables $Y_s := \min\{X_s, \Delta\}$, $s = 0, \ldots, s^* - 1$, and $Y := \sum_{s=0}^{s^*-1} Y_s$. We will see that the differences caused by truncating the variables X_s are negligible for our purpose.

The expected value of X_s is $1/p_s$, the expected value of Y_s is $(1 - q_s^{\Delta})/p_s$ (see Appendix A). If we express p_s as $p_s = \frac{1}{1+\frac{1}{z_s}}$ for $z_s = \frac{(k-s)\cdot\Delta^2}{s-1}$, then

$$1 - \frac{\mathbf{E}[Y_s]}{\mathbf{E}[X_s]} = q_s^{\Delta} = (1 - p_s)^{\Delta} = \left(1 - \frac{1}{1 + \frac{1}{z_s}}\right)^{\Delta} = \left(\frac{1}{z_s + 1}\right)^{\Delta}.$$

As z_s is decreasing with s and $s \leq s^* - 1 \leq k \cdot \bigl(1 - \frac{r}{2\Delta^2}\bigr)$, we can bound z_s for $s \geq 1$ by

$$z_s \geq \frac{(k - k \cdot \bigl(1 - \frac{r}{2\Delta^2}\bigr)) \cdot \Delta^2}{k \cdot \bigl(1 - \frac{r}{2\Delta^2}\bigr) - 1} = \frac{\frac{r}{2}}{1 - \frac{r}{2\Delta^2} - \frac{1}{k}} \geq \frac{r}{2} \,.$$

The non-negativity of the second last denominator follows from Inequalities (4) and (5). By applying Inequality (6), we get

$$\frac{\mathbf{E}[Y_s]}{\mathbf{E}[X_s]} = 1 - \left(\frac{1}{z_s + 1}\right)^{\Delta} \geq 1 - \left(\frac{1}{\frac{r}{2} + 1}\right)^{\Delta} \geq 1 - \left(\frac{r}{2\Delta^2}\right)^2 . \qquad (12)$$

Due to Inequality (12) a lower bound for $\mathbf{E}[X]$ implies a lower bound for $\mathbf{E}[Y]$. The former one can be bounded as follows.

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{s=0}^{s^*-1} \mathbf{E}[X_s] = \sum_{s=0}^{s^*-1} \frac{1}{p_s} = 1 + \sum_{s=1}^{s^*-1} \left(1 + \frac{s-1}{(k-s) \cdot \Delta^2}\right) \\
&= s^* + \sum_{i=k-s^*+1}^{k-1} \frac{k-i-1}{i \cdot \Delta^2} = s^* - \frac{s^*-1}{\Delta^2} + \frac{k-1}{\Delta^2} \cdot \sum_{i=k-s^*+1}^{k-1} \frac{1}{i} \\
&\geq s^* \cdot \left(1 - \frac{1}{\Delta^2}\right) + \frac{k-1}{\Delta^2} \cdot \log\left(\frac{k}{k - s^* + 1}\right) .
\end{aligned}$$

Using $s^* \geq k \cdot \bigl(1 - \frac{r}{2\Delta^2}\bigr)$, we can lower bound this by

$$\begin{aligned}
\mathbf{E}[X] &\geq k \cdot \left(1 - \frac{r}{2\Delta^2}\right) \cdot \left(1 - \frac{1}{\Delta^2}\right) + \frac{k-1}{\Delta^2} \cdot \log\left(\frac{k}{k - k \cdot \bigl(1 - \frac{r}{2\Delta^2}\bigr) + 1}\right) \\
&\geq k \cdot \left[\left(1 - \frac{r+2}{2\Delta^2}\right) + \frac{k-1}{k\Delta^2} \cdot \log\left(\frac{\Delta^2}{\frac{r}{2} + \frac{\Delta^2}{k}}\right)\right] .
\end{aligned}$$

Inequalities (5), (8), (9) and the choice of Δ yield

$$\begin{aligned}
\frac{\mathbf{E}[X]}{k} &\geq 1 - \frac{\bigl(1 + \frac{\varepsilon}{3}\bigr) \cdot r}{2\Delta^2} + \frac{1 - \frac{\varepsilon}{9}}{\Delta^2} \cdot \log\left(\frac{\Delta^2}{r}\right) \\
&\geq 1 - \frac{\bigl(1 + \frac{\varepsilon}{3}\bigr) \cdot r}{2\Delta^2} + \frac{1 - \frac{\varepsilon}{9}}{\Delta^2} \cdot r \cdot \frac{1 + \varepsilon}{2} \\
&\geq 1 + \frac{r}{2\Delta^2} \cdot \frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right) ,
\end{aligned}$$

where the last inequality holds because $\varepsilon \in (0, 1)$. Applying Inequality (12), we can show that even the expected value of Y is significantly larger than k.

$$\begin{aligned}
\frac{\mathbf{E}[Y]}{k} &\geq \left(1 - \left(\frac{r}{2\Delta^2}\right)^2\right) \cdot \frac{\mathbf{E}[X]}{k} \geq \left(1 - \left(\frac{r}{2\Delta^2}\right)^2\right) \cdot \left(1 + \frac{r}{2\Delta^2} \cdot \frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right)\right) \\
&= 1 + \frac{r}{2\Delta^2} \cdot \left(\frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right) - \frac{r}{2\Delta^2} - \frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right) \cdot \left(\frac{r}{2\Delta^2}\right)^2\right) \\
&= 1 + \frac{r}{2\Delta^2} \cdot \left(\frac{\varepsilon}{3} + \left(\frac{\varepsilon}{3}\right)^2 - \left(\frac{r}{2\Delta^2} + \frac{\varepsilon}{3} \cdot \left(1 + \frac{\varepsilon}{3}\right) \cdot \left(\frac{r}{2\Delta^2}\right)^2\right)\right) .
\end{aligned}$$

Hence, we get $\mathbf{E}[Y] \geq k \cdot (1 + \frac{r}{2\Delta^2} \cdot \frac{\varepsilon}{3}) = k + k \cdot f$ for $f = f(k) = \frac{\varepsilon r}{6\Delta^2}$ because of Inequality (10). Using Hoeffding's Inequality [6], we can now bound the probability to reach vertex v_{s^*} within k steps in the Markov chain above.

$$\begin{aligned}\mathbf{Pr}[X \leq k] \leq \mathbf{Pr}[Y \leq k] \leq \mathbf{Pr}[\mathbf{E}[Y] - Y \geq k \cdot f] &\leq \exp\left(-\frac{2 \cdot (k \cdot f)^2}{s^* \cdot \Delta^2}\right) \\ \leq \exp\left(-\frac{2k^2 f^2}{k \cdot \Delta^2}\right) &= \exp\left(-k \cdot \frac{2f^2}{\Delta^2}\right) .\end{aligned}$$

Because of Inequalities (7) and (11) we can bound the fraction $2f^2/\Delta^2$ by

$$\begin{aligned}\frac{2f^2}{\Delta^2} &= \frac{\varepsilon^2 r^2}{18\Delta^6} \geq \frac{\varepsilon^2 r^2}{19 r^3 \cdot e^{3r(1+\varepsilon)/2}} = \frac{\varepsilon^2}{19} \cdot \frac{1}{e^{3r(1+\varepsilon)/2 + \log r}} \geq \frac{\varepsilon^2}{19} \cdot \frac{1}{e^{(3/2+3\varepsilon)\cdot r}} \\ &= \frac{\varepsilon^2}{19} \cdot \frac{1}{e^{(3/2+3\varepsilon)\cdot\delta\cdot\log k}} = \frac{\varepsilon^2}{19} \cdot k^{-(3/2+3\varepsilon)\cdot\delta} .\end{aligned}$$

If $r = o(\log k)$, then $\delta \in o(1)$ and $\mathbf{Pr}[X \leq k] \leq \exp\left(-k^{1-o(1)}\right)$. If $r = \delta^* \cdot \log k$ for some fixed real $\delta^* \in (0, 2/3)$, then we get

$$\begin{aligned}\mathbf{Pr}[X \leq k] &\leq \exp\left(-k^{-o(1)} \cdot k^{1-\left(\frac{3}{2}+\frac{2\log r}{r}\right)\cdot\delta^*}\right) \\ &= \exp\left(-k^{1-\frac{3}{2}\delta^* - o(1)} \cdot k^{-\frac{2\log r}{\log k}}\right) \\ &= \exp\left(-k^{1-\frac{3}{2}\delta^* - o(1)}\right) .\end{aligned}$$

This concludes the proof of Theorem 1.

3 Conclusion

We proved that, in general, `k-means++` yields an $o(\log k)$-approximation only with negligible probability. The proof of this result is based on instances with fairly high dimension. Since we constructed the simplices in orthogonal dimensions, our instances have dimension $\Theta(k^2)$.

It remains open how `k-means++` behaves on instances in small dimensions. One intriguing question is whether there exists an upper bound for the expected approximation ratio of `k-means++` that depends only on the dimension of the instance. Currently we cannot exclude the possibility that the expected approximation ratio of `k-means++` is $O(\log d)$ where d is the dimension of the instance.

References

1. Aggarwal, A., Deshpande, A., Kannan, R.: Adaptive sampling for k-means clustering. In: Dinur, I., Jansen, K., Naor, J., Rolim, J. (eds.) APPROX 2009. LNCS, vol. 5687, pp. 15–28. Springer, Heidelberg (2009)

2. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of Euclidean sum-of-squares clustering. Machine Learning 75(2), 245–248 (2009)
3. Arthur, D., Manthey, B., Röglin, H.: k-means has polynomial smoothed complexity. In: Proc. of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp. 405–414 (2009)
4. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027–1035. SIAM, Philadelphia (2007)
5. Berkhin, P.: Survey of Clustering Data Mining Techniques. Technical report, Accrue Software (2002)
6. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301), 13–30 (1963)
7. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–136 (1982)
8. Vattani, A.: k-means requires exponentially many iterations even in the plane. In: Proc. of the 25th ACM Symposium on Computational Geometry (SoCG), pp. 324–332. ACM, New York (2009)

A The Expected Value of Truncated Geometrically Distributed Random Variables

Let X be a geometrically distributed random variable with parameter p, let $q := 1 - p$ and let M be a non-negative integer. The expected value of the truncated random variable $Y := \min\{X, M\}$ is

$$\begin{aligned}\mathbf{E}[Y] &= \sum_{i=1}^{\infty} \min\{i, M\} \cdot p \cdot q^{i-1} = \sum_{i=1}^{\infty} i \cdot p \cdot q^{i-1} - \sum_{i=M+1}^{\infty} (i - M) \cdot p \cdot q^{i-1} \\ &= \mathbf{E}[X] - q^M \cdot \sum_{i=1}^{\infty} i \cdot p \cdot q^{i-1} = \left(1 - q^M\right) \cdot \mathbf{E}[X] = \frac{1 - q^M}{p} .\end{aligned}$$

B Inequalities (4) to (11)

Throughout this section any inequality $f(k) \le g(k)$ is a short hand for $f(k) \le g(k)$ for sufficiently large k. First note that regardless of the choice of r the inequalities $r \le \frac{2}{3} \log k$ and $\frac{2}{3}(\log r)/r \le \varepsilon \le 1$ hold. The latter one immediately implies Inequality (11).

- Inequality (4) follows from $\Delta^2 \ge \tilde{\Delta}^2 \ge r \cdot \exp\left(\frac{r}{2}\right)$ which is greater than r because $r > 0$.
- As $\tilde{\Delta} \le \sqrt{r} \cdot \exp(r/2) \le \sqrt{r} \cdot \exp\left((\log k)/3\right) = \sqrt{r} \cdot \sqrt[3]{k} \le \sqrt{r} \cdot \sqrt{k/2} - 1$, we get Inequality (5): $\Delta^2 \le (\tilde{\Delta} + 1)^2 \le r \cdot k/2$.
- Due to the fact $\Delta \to \infty$ we get $2^{\Delta} \ge \Delta^4$. The inequalities $(r + 2)/2 \ge 2$ and $2\Delta^2/r \le \Delta^2$ then imply the correctness of Inequality (6).

- Inequality (7) is a consequence of $\tilde{\Delta} \to \infty$. This yields $\tilde{\Delta} + 1 \leq \sqrt[6]{19/18} \cdot \tilde{\Delta}$ and hence $\Delta^6 \leq (\tilde{\Delta} + 1)^6 \leq \frac{19}{18}\tilde{\Delta}^6 = \frac{19}{18}r^3 \cdot \exp(3r \cdot (1 + \varepsilon)/2)$.
- Inequalities (8) and (9) hold if $\varepsilon \geq 9/k$ and $\varepsilon \geq 6/r$. This is true since $\varepsilon = \Omega\left((\log r)/r\right)$, $k = \exp(\Omega\left(r\right))$ and $1/r = O(1/r)$.
- Let us consider Inequality (10). As $\Delta^2 > r$ (see Inequality (4)), $r/(2\Delta^2) + \varepsilon/3 \cdot (1 + \varepsilon/3) \cdot (r/(2\Delta^2))^2 \leq r/(2\Delta^2) + 4/9 \cdot r/(2\Delta^2) \leq r/\Delta^2$. The correctness follows from $\Delta^2 \geq r \cdot \exp(r/2)$, i.e. $r/\Delta^2 \leq 1/\exp(r/2)$, whereas $(\varepsilon/3)^2 = \Omega\left(1/r^2\right)$.