A Concentration Inequality for the Facility Location Problem

Sandeep Silwal
MIT *

Abstract

We give a concentration inequality for a stochastic version of the facility location problem. We show the objective $C_n = \min_{F \subseteq [0,1]^2} |F| + \sum_{x \in X} \min_{f \in F} \|x - f\|$ is concentrated in an interval of length $O(n^{1/6})$ and $\mathbb{E}[C_n] = \Theta(n^{2/3})$ if the input X consists of i.i.d. uniform points in the unit square. Our main tool is to use a geometric quantity, previously used in the design of approximation algorithms for the facility location problem, to analyze a martingale process. Many of our techniques generalize to other settings.

Keywords: Facility Location; Concentration Inequality; Stochastic Optimization

1 Introduction

Let X be a set of n points in $D \subset \mathbb{R}^d$. The (minimum) facility location problem (with uniform demands) is the problem of finding a set of points $F \subset D$ (called facilities or centers) to minimize the objective

$$C_n(X) = \min_{F \subset D} |F| + \sum_{x \in X} \min_{f \in F} \|x - f\|. \tag{1}$$

The facility location problem is a well studied combinatorial optimization problem and is NP-hard in general. As is the case of many other NP-hard combinatorial optimization problems, stochastic versions of these problems have been studied (see [5, 25, 17, 21] and the book [34] for examples in TSP, MST, and many other problems). In this paper, we study the stochastic version of the facility location problem. In particular, if our domain D is the unit square $[0,1]^2$ in $\mathbb{R}^2$, our result presented in Theorem 3.3 states that C_n is concentrated in an interval of length $O(n^{1/6})$ and satisfies the following concentration bound

$$\Pr(|C_n - \mathbb{E}[C_n]| \geq t n^{1/6}) \leq \exp(-ct^2)$$

where $\mathbb{E}[C_n] = \Theta(n^{2/3})$. However, our techniques are more general and can be extended to other domains and distributional assumptions.

To give more context to our result, we compare our bound against Rhee and Talagrand's concentration result for the k-median problem [31]. The k-median problem is a related optimization problem where only the second term of the objective in (1) appears and where we are constrained to $|F| = k$. Rhee and Talagrand showed in [31] that the cost of the objective function for the k-median problem concentrates on an interval of length $O(\sqrt{n/k})$. This follows from the following theorem.

*77 Massachusetts Avenue Cambridge, MA 02139. silwal@mit.edu

Theorem 1.1 (Theorem B in [31]). *Let Q_n be the random variable which denotes the cost of the k-median problem on the unit square in $\mathbb{R}^2$ where n points are drawn independently and uniformly at random from the unit square. There exists a constant $c > 0$ such that*

$$\Pr(|Q_n - \mathbb{E}[Q_n]| \geq t) \leq \exp(-ct^2 k/n).$$

While their techniques aren't applicable in our setting, we can interpret our results as 'plugging in a specific value' of $k = n^{2/3}$ even though $|F|$ is a random variable in our case.

Our proof strategy relies on standard martingale tools but uses a more geometric and 'local' representation of C_n that allows us to better track the objective cost as new random points are drawn. This geometric formulation is stated in Section 2 and has been previously used in algorithmic works related to the facility location problem [28, 4]. We also present a weaker concentration result using Talagrand's concentration inequality in Theorem 3.2 which we conjecture gives us the optimal concentration result for a variety of settings. We leave it as an interesting open problem to verify this conjecture.

Lastly, we note that while many of our techniques can be adapted to more general distributions and domains, we mainly stick to the uniform distribution on the unit square in $\mathbb{R}^2$ for our presentation due to simplicity and clarity since this case already conveys our ideas. Furthermore, the uniform distribution is the most well-studied case for stochastic combinatorial optimization problems in general and has lead to a wide array of results and influential techniques; for example, in matchings [35, 38, 26, 16], minimum-spanning trees [18, 33, 23, 29, 19, 20, 10, 15], the traveling salesman problem [6, 22, 27, 36, 9], bin packings [8, 7, 12, 24, 32], k-SAT [2, 3, 1, 13, 14], and many more problems. See the references within the cited papers, the book [34] and the excellent set of notes in [37] for further examples.

1.1 Related Work

Piersma considered a different formulation of stochastic facility location [30]. In their work, they consider a capacitated version of facility location where each facility is only allowed to 'serve' a fixed number of points. Their formulation is given by an integer program with randomly drawn coefficients for their linear constraints. In contrast, our input points are random and the cost to connect a point to a facility is given by Euclidean distances rather than randomly drawn values. This leads to their integer program having a non zero probability of being infeasible whereas in our setting it is always possible to find a solution.

In addition, the 'scaling' of the cost for our formulation is naturally on the order of $n^{2/3}$ whereas in [30], the scaling is n. Furthermore, the goal of [30] is to mostly study the convergence of the cost of their formulation using central limit type theorems whereas for us we are more concerned with concentration. Finally, our formulation is more geometric and closely related to the stochastic k-median problem studied previously in [31].

2 Preliminaries

Our points are given by $X = (X_1, \cdots, X_n)$ and all our asymptotics are as $n \to \infty$.

A key point about Rhee and Talagrand's method is that it relies heavily on the fact that the k-median objective is only composed of 'local' terms (representing the cost incurred by every input point). Local behaviour is often the key to getting concentration bounds for complicated processes, even in other settings such as random graphs, since they allow tools such as bounded difference or Lipschitz concentration inequalities to be used. On first glance, the facility location problem

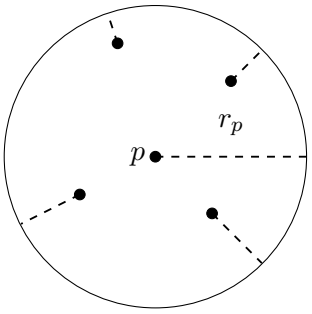

Figure 1: For each point p, we compute a radius r_p such that the dotted lines add to 1.

does not seem to have nice local structures due to the additional global $|F|$ term. However in the algorithmic literature about facility location, a suitable local geometric quantity has been considered which will form the basis of our analysis. This is additionally interesting since it's an example of an algorithmic result being used in probability. The geometric quantity is defined as follows.

Let $B(p, r)$ denote the ball of radius r centered at p. For each $p \in X$, define radius $r_p > 0$ to satisfy the following relation.

$$\sum_{q \in B(p,r_p) \cap X} (r_p - \|p - q\|) = 1. \tag{2}$$

We record some properties of r_p, some of which were used in previous algorithmic works [28, 4].

Lemma 2.1 (Lemma 1 in [4]). *Every $p \in X$ satisfies $r_p \geq 1/|B(p, r_p) \cap X|$.*

Proposition 2.2. *Let $q \in B(p, r_p) \cap X$. Then $r_q \leq 3r_p$.*

Proof. Any point $q' \in B(p, r_p) \cap X$ satisfies $\|q - q'\| \leq 2r_p$ from the triangle inequality. If we consider the ball $B(q, 3r_p)$ then the sum of the dashed lines in Figure 1 contributed by points from $B(p, r_p) \cap X$ is at least r_p each. This is because all points $q' \in B(p, r_p) \cap X$ must also be in $B(q, 2r_p)$. The result follows from noting that $|B(p, r_p) \cap X| \geq 1/r_p$ due to Lemma 2.1. □

Proposition 2.3. *In the optimal solution of (1), every point p must have some $f \in F$ at distance at most $3r_p$.*

Proof. Suppose that a point p does not have a center $f \in F$ within distance $3r_p$. We show in this case that the cost can be reduced. We know from Lemma 2.1 that $|B(p, r_p) \cap X| \geq 1/r_p$. Let m be the number of points in $B(p, r_p) \cap X$ *excluding* p. It follows that these points don't have an f within distance $2r_p$. Therefore in total, the contribution of the points in $|B(p, r_p) \cap X|$ to the objective function is at least $2mr_p + 3r_p$. Now if we put a new f at the point p, then the m points all have a facility within distance r_p and therefore, the cost of the solution decreases by at least

$$(2mr_p + 3r_p) - (1 + mr_p) = (m + 3)r_p - 1 = ((m + 1)r_p - 1) + 2r_p > 0$$

where the last inequality follows from the fact that $(m + 1)r_p \geq 1$ due to Equation (2). Thus it follows that the optimal solution must have some $f \in F$ that is within distance $3r_p$ of p. □

Lemma 2.4. *There exists constants $c, C > 0$ such that $C \sum_{p \in X} r_p \geq C_n \geq c \sum_{p \in X} r_p$.*

Proof. From [28, 4] we know that $\sum_{p \in X} r_p$ is a constant factor approximation of C_n *when we restrict* the set F to be a subset of the points X. In our case, we want to study a more general version

where the set F can come from the entire space. Previous results readily extend to our desired upper bound since not restricting F only decreases the value of the objective function.

For the lower bound, we denote C'_n as the optimal cost of the objective where F is restricted to points in X. Consider the optimal solution for Problem (1) (whose objective cost is C_n) and denote its set of facilities as F^*. For each $f \in F^*$, consider the set of X that it *serves*: for each f we have disjoint subsets $X_f \subseteq X$ such that f is the closest point in F^* to points in X_f, breaking ties arbitrarily. Move each f to its closest point in X_f. This increases the cost of the objective in (1) by at most $\sum_{x \in X} \min_{f \in F^*} \|x - f\|$ since the distance from each point $p \in X_f$ to f increased by at most $\|p - f\|$. Furthermore, we have that this new configuration is a valid solution for the objective where we restrict the set of facilities to come from the points in X and therefore, serves as an upper bound for C'_n. Altogether, we have

$$2C_n \geq 2|F^*| + 2\sum_{x \in X} \min_{f \in F^*} \|x - f\| \geq |F^*| + 2\sum_{x \in X} \min_{f \in F^*} \|x - f\| \geq C'_n \geq c \sum_{p \in X} r_p$$

where the last relation follows from [4]. Adjusting the constants gives us our desired bound. □

Lastly we calculate the expected value of C_n for uniformly random inputs.

Theorem 2.5. *The expected value of the objective* (1) *for i.i.d. uniform points in* $[0,1]^2$ *satisfies* $\mathbb{E}[C_n] = \Theta(n^{2/3})$.

Proof. We know from Lemma 2.4 that $\sum_{p \in X} r_p$ is a constant factor approximation to the objective given in (1). Therefore, we fix our attention to calculating $\mathbb{E}[r_p]$. Fix a point p and let $r = n^{-1/3}$. The number of points that fall in $B(p, r)$ is distributed as $\text{Bin}(n, cr^2)$ for some constant c. By a standard binomial concentration, we know that $|B(p, r) \cap X| = \Theta(n^{1/3})$ with probability at least $1 - e^{-\Theta(n^{1/3})}$. For example, this follows from Theorems 2.3 and 2.4 in [11].

Conditioning on this event $\mathcal{E}$, we see that from the geometric interpretation of r_p in Figure 1 that increasing r by $Cn^{-1/3}$ for some sufficiently large constant C will imply $r_p = O(n^{-1/3})$. Thus,

$$\mathbb{E}[r_p] \leq \mathbb{E}[r_p \mid \mathcal{E}] + \Pr(\mathcal{E}^c)\mathbb{E}[r_p \mid \mathcal{E}^c] = O(n^{-1/3}) + e^{-\Theta(n^{1/3})} = O(n^{-1/3}).$$

For the lower bound, we consider the same approach as above but let $r = c'n^{-1/3}$ for a sufficiently small constant c'. In this case, we see that $|B(p, r) \cap X| \leq c''n^{1/3}$ with probability at least $1 - e^{-\Theta(n^{1/3})}$ for a sufficiently small constant c''. Again conditioning on this event $\mathcal{E}$, we see that to make r_p as small as possible, the worst configuration is where all the points in $|B(p, r) \cap X|$ are located at p. In that case, we see that $r_p \geq 1/(c''n^{1/3})$. Then we calculate that

$$\mathbb{E}[r_p] \geq \Pr(\mathcal{E})\mathbb{E}[r_p \mid \mathcal{E}] = \Omega(n^{-1/3}).$$

The final result follows by linearity of expectations. □

Remark 2.6. Theorem 2.5 is essentially the only place where the uniform distribution assumption and our domain assumption of the unit square in $\mathbb{R}^2$ are used as they allow for an easy calculation of $\mathbb{E}[r_p]$. Most of our concentration arguments in Section 3 generalize to arbitrary distributions and arbitrary domains where the appropriate value of $\mathbb{E}[r_p]$ is used.

2.1 A Heuristic Derivation of the Concentration Bound

As stated in the introduction, our main result of C_n being concentrated in an interval of length $O(n^{1/6})$ for the case where the input points are i.i.d. uniform on the unit square in $\mathbb{R}^2$ can be interpreted as picking a suitable choice of k in Rhee and Talagrand's bound. Indeed, consider the k-median problem where k is some parameter specified later. Heuristically, it makes sense to pick the k facilities in a uniform grid of squares of dimension $1/\sqrt{k} \times 1/\sqrt{k}$. In such a case, the distance from any point to its nearest facility is at most $\Theta(1/\sqrt{k})$ and there are k facilities. Thus, the facility location problem objective is $\Theta(n/\sqrt{k}) + k$. Minimizing this as a function of k, we see that $k = \Theta(n^{2/3})$. Now Rhee and Talagrand's concentration bound states that the cost of the random k-median concentrates on an interval of length $O(\sqrt{n/k})$. 'Plugging in' $k = n^{2/3}$ we get an interval of $O(n^{1/6})$ which matches the bound given by Theorem 3.3.

Of course, the above justification is pure heuristics and not rigorous. In addition, Rhee and Talagrand's proof is substantially different than ours. In their work, they exploit the fact that the k-median objective is composed of only 'local' terms whereas we have a 'global' term $|F|$. However, we rely on the geometric properties of the radii r_p outlined above.

Note that the sum of the radii r_p only serves as a constant factor approximation to the objective value. Therefore, it is *not sufficient* to understand the concentration of the sum of the radii values if we really want to get concentration on the order of $o(\mathbb{E}[C_n])$. Nevertheless, we are able to leverage their properties to provide such a concentration bound for the objective value C_n.

3 Concentration

We prove our main concentration inequality in this section. First, we present a suboptimal concentration inequality that follows from Talagrand's inequality. It is interesting to note that an application of this inequality is not sufficient to provide us with the best concentration bound, which is a rare occurrence. Nonetheless, we conjecture that a sharper analysis of our proof using Talagrand's inequality should result in the optimal concentration bound.

We first recall Talagrand's concentration inequality for 'non uniform' differences.

Theorem 3.1 (Talagrand's Concentration Inequality)**.** *Let f be a function on the product space $\Omega = \prod_{i=1}^{n} \Omega_i$ such that for every $x \in \Omega$, there exists $\alpha_i(x) \geq 0$ with*

$$f(x) \leq f(y) + \sum_{i:x_i \neq y_i} \alpha_i(x)$$

for all $y \in \Omega$. Let M denote the median of f and

$$c = \sup_{x \in \Omega} \sum_{i=1}^{n} \alpha_i(x)^2.$$

Then,

$$\Pr(|f - M| \geq t) \leq 2e^{-t^2/4c}.$$

Using Theorem 3.1, we can prove a weaker concentration result that states that C_n is concentrated in an interval of length $n^{1/3}$ in the case of uniform inputs in $[0, 1]^2$.

Theorem 3.2 (Weak Concentration)**.** *Suppose the points $X = (X_1, \cdots, X_n)$ are chosen independently from some domain $D \subset \mathbb{R}^d$. For each point $p \in X$, define the radius r_p according to (2). Let C_n denote the cost of the objective function (1). C_n satisfies the concentration inequality*

$$\Pr(|C_n - \text{Med}(C_n)| \geq t) \leq e^{-t^2/s}$$

where s is any upper bound on the quantity $\sum_{p\in X} r_p^2$ for ***any*** *set of n points chosen from D. In particular, if $D = [0,1]^2$ and the points in X are chosen independently in $[0,1]^2$ (not necessarily from the uniform distribution), we have*

$$\Pr(|C_n - \mathrm{Med}(C_n)| \geq t) \leq e^{-t^2/O(n^{2/3})}$$

where $\mathrm{Med}(C_n)$ denotes the median value of C_n, i.e., C_n is concentrated in an interval of length $n^{1/3}$.

Proof. Fix an arbitrary collection of points $X = (X_1, \cdots, X_n)$. We define our vector α by letting the ith coordinate of α be equal to Cr_i for a suitably large constant C. Now given an optimal clustering of a different set of points $Y = (Y_1, \cdots, Y_n)$, we want to extend it to a clustering of X by only using additional 'budget' given by $\sum_{X_i \neq Y_i} \alpha_i(X)$.

Take the set of facilities for Y. Our goal is to show that we can find a facility for every point p in $X \setminus Y$ within distance $O(r_p)$ where the constant in the O doesn't depend on any parameters of the problem. To do this, we first consider the following two cases:

Case 1: At least $1/2$ of the points of $B(p, r_p) \cap X$ are in Y.

Let q be any such point in the intersection. Define r_q^Y be the radius of q calculated according to (2) but using only the points in Y. We claim that $r_q^Y = O(r_p)$. To show this, we know from Lemma 2.1 that $|B(p, r_p) \cap X| \geq 1/r_p$ so at least $1/2r_p$ points are in $|B(p, r_p) \cap X \cap Y|$. If we go radius $O(r_p)$ away from q, then the sum (2) in Y will be more than 1, which implies $r_q^Y = O(r_p)$. Thus from Proposition 2.3, we know that some facility of Y will be within distance $O(r_p)$ from p.

Case 2: At least $1/2$ of the points of $B(p, r_p) \cap X$ are not in Y.

In this case, we want to find enough points in $B(p, r_p)$ that are in X but not in Y to 'pay for a new center' using their radii (that's the budget we are allowed from Theorem 3.1). Pick a large constant C. We can assume that every $w \in B(p, Cr_p) \cap X$ doesn't fall in case 1, i.e., the ball $B(w, r_w)$ contains at least $1/2$ of its points from X. Indeed, otherwise, p will have a facility in radius $O(r_p)$ from the observation that any w satisfies $r_w = O(r_p)$ from Proposition 2.2.

Now consider the w in the ball $B(p, r_p) \cap X$ with the smallest radius r_w. If $r_w \geq r_p/2$ then we can pay for a new facility from the points in $B(p, r_p) \cap X$ that are not in Y because we know there are at least $1/2r_p$ such points and they all contribute radii $\Omega(r_p)$. If $r_w \leq r_p/2$, then we recurse into the ball $B(w, r_w)$. If every $w' \in B(w, r_w) \cap X$ satisfies that $r'_w \geq r_w/2$ then we are again done by the same argument. Otherwise, we again recurse. We know this process ends since we only have n points and when it ends, we are at distance at most $r_p(1 + 1/2 + 1/4 + \cdots) \leq 2r_p$ away from p. Therefore, we can use the entries of α for the points of $B(p, r_p) \cap X$ that are not in Y to pay for a new facility.

Now to finish our argument for all points, we just repeat the above cases iteratively: We start with a clustering of Y and its facilities. For every point $p \in X$, if it has a facility near $C'r_p$ for a large constant C' then we are done. Otherwise, we consider $B(p, r_p)$ and perform one of the above two cases.

Applying Theorem 3.1, we get that C_n satisfies a concentration inequality of the form

$$\Pr(|C_n - \mathrm{Med}(C_n)| \geq t) \leq e^{-t^2/O(\sum_{p\in X} r_p^2)}.$$

Therefore, the value C_n is concentrated in an interval of length $O(\sqrt{\sum_{p\in X} r_p^2})$. To bound this we note that $r_p^2 \leq r_p$ since $r_p \leq 1$ always since p is included in the sum in Equation (2). We now

claim that $\sum_p r_p = O(n^{2/3})$. This is because $\sum_p r_p$ is constant factor upper bound on facility location cost from Lemma 2.4 and for every configuration, we can *deterministically* achieve this cost by considering the following construction: place k points in a uniform grid. Then the cost is $n/\sqrt{k} + k$ since each point is within distance $O(1/\sqrt{k})$ from any center. Optimizing for k we get $\sum_{p \in X} r_p = O(n^{2/3})$. □

We conjecture that the above analysis actually gives us a tighter concentration bound. To show this, we would need a way to control the value of $\sum_{p \in X} r_p^2$ which would depend on the domain D and the how the points in X are drawn.

We now present an argument that gives a much sharper concentration bound using less sophisticated tools.

Theorem 3.3 (Strong Concentration)**.** *Let $D = [0,1]^2$ and suppose the points in X are i.i.d uniform in $[0,1]^2$. Then,*

$$\Pr(|C_n - \mathbb{E}[C_n]| \geq t) \leq e^{-t^2/O(n^{1/3})},$$

i.e., C_n is concentrated in an interval of length $n^{1/6}$.

Proof. The main ideas of the proof generalize to beyond the uniform in the unit square assumption. We explicitly point out where we assume this in the proof. First, let S be a set of points in D. We first claim that for any $p \notin S$,

$$C(S \cup \{p\}) \leq C(S) + O(r_p^{S \cup \{p\}}) \tag{3}$$

where $r_p^{S \cup \{p\}}$ means we calculate the radius (2) with respect to the points $S \cup \{p\}$ and $C(\cdot)$ denotes the facility location problem cost. To show this, we either have $r_p^{S \cup \{p\}} = 1$, in which case we can just put a new facility located at p, or otherwise, there must exist some some point $q \in B(p, r_p^{S \cup \{p\}}) \cap S$. That point must have been served in $C(S)$ so by Proposition 2.3, there must exist a facility near q within distance $3r_q^S$ where we calculate the radius of q with respect to the points in S only.

Our goal is to show that $r_q^S = O(r_p^{S \cup \{p\}})$. Indeed, if it is the case that $r_q^S \leq \|q - p\|$ then clearly $r_q^S = r_q^{S \cup \{p\}}$ since q's radius doesn't change. Else, $p \in B(q, r_q^S)$ in which case $r_q^{S \cup \{p\}}$ is potentially smaller than r_q^S. However, considering the geometric interpretation of the radii given in Figure 1, we know that the distance contributed by p towards $r_q^{S \cup \{p\}}$ is at most half of the other distances (in other words, the dotted line stemming from p contributes total length at most half to the computation of $r_q^{S \cup \{p\}}$ since there is also a dotted line stemming from q). Thus, $r_q^{S \cup \{p\}} \geq r_q^S/2$. Finally from Proposition 2.2, it follows that $r_q^S = O(r_p^{S \cup \{p\}})$.

We now use our above observation to perform a martingale analysis. Consider the Doob martingale $\Lambda_i = \mathbb{E}[C_n \mid X_1, \ldots, X_i]$ for $1 \leq i \leq n$. We analyze the martingale difference $\Delta_i = \Lambda_i - \Lambda_{i-1}$ which can be written as

$$\Delta_i = \mathbb{E}[C_n(X_1, \ldots, X_i, \ldots, X_n) - C_n(X_1, \ldots, X_i', \ldots, X_n) \mid X_1, \ldots, X_i]$$

where X_i' is an independent copy of X_i. Defining $S = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$, we see that

$$\Delta_i = \mathbb{E}[C_n(S \cup \{X_i\}) - C_n(S \cup \{X_i'\}) \mid X_1, \ldots, X_i]$$

and therefore,

$$|\Delta_i| \leq \mathbb{E}[r_{X_i}^{S \cup \{X_i\}} + r_{X_i'}^{S \cup \{X_i'\}} \mid X_1, \ldots, X_i]$$

by (3). Now crucially, we know that the radius defined in (2) *can only decrease* as more points are added. Therefore, we bound each of the expectations above using only the randomness of the remaining $n - i$ points.

Note that everything we have stated so far in the proof is perfectly valid over general domains in $\mathbb{R}^d$ for any d and any choice of distribution that the input points are drawn from, as long as we draw points independently. Now if

$$|\Delta_i|^2 \leq f(i)$$

for some real valued function f, then we immediately arrive at

$$\Pr(|C_n - \mathbb{E}[C_n]| \geq t) \leq e^{-t^2/\sum_{i=1}^{n} f(i)}$$

from the Azuma-Hoeffding inequality. The uniform assumption makes the calculation of $f(i)$ particularly tractable. In general, we can calculate $f(i)$ by obtaining a version of Theorem 2.5 for an alternative distribution of choice as we will see shortly.

We now use our assumptions that the domain is $D = [0, 1]^2$ and X consists of i.i.d. uniform points on the unit square to bound $|\Delta_i|$. This is the main part of the proof where we use the distribution and domain assumptions and the argument which follows can be adapted for other distributional and domain assumptions by finding a suitable bound on $|\Delta_i|$.

From a similar analysis as in Theorem 2.5 (except for a slight caveat that will be addressed in a bit), we know that each of the expectations in our martingale difference can be bounded by $O((n - i)^{-1/3})$ and so it follows that $|\Delta_i| = O((n - i)^{-1/3})$. We now calculate $\sum_i |\Delta_i|^2$. We have that

$$\sum_{i=1}^{n} (n - i)^{-2/3} \sim \int_1^n x^{-2/3}\, dx = O(n^{1/3}) \tag{4}$$

and so by the Azuma-Hoeffding inequality, we get the concentration bound

$$\Pr(|C_n - \mathbb{E}[C_n]| \geq t) \leq e^{-t^2/O(n^{1/3})},$$

as desired.

To tie up the loose ends, we note that the upper bound for the expectation given in Theorem 2.5 might not hold if $n-i$ is too small. However, we don't care about this case since we can substitute the deterministic bound $|\Delta_i| = O(1)$ which always holds since the unit square is bounded. In particular, we can use the deterministic bound say when $n-i = O(n^{1/3})$ in which case the 'variance' calculation of (4) still gives us the same asymptotics. □

An interesting open question is if the concentration bound of Theorem 3.3 is tight for the uniform unit square case. This is possibly a much harder question but we suspect the answer to be yes. This is because we can show the variance of the quantity $\sum_{p \in X} r_p$ is at least $\Omega(n^{1/3})$ (by performing similar calculations as in Theorem 2.5) which implies that the quantity $\sum_{p \in X} r_p$ truly fluctuates on an interval of length $\Omega(n^{1/6})$ (the standard deviation). Since this quantity has been very influential in designing algorithms for the facility location problem [28, 4], it hints that our bound is close to the truth. Of course this doesn't formally imply anything for the facility location problem case since $\sum_{p \in X} r_p$ is only a constant factor approximation to the cost. In addition, a $\Omega(n^{1/6})$ fluctuation matches the fluctuation when we 'plug in' $k = n^{2/3}$ into the Rhee and Talagrand's bound as done in Section 2.1, again hinting that 3.3 has the right concentration.

Other Open Problems. In this paper we analyzed a well known combinatorial optimization problem under stochastic inputs by borrowing algorithmic ideas that introduce a 'local' property. Its plausible that other complicated algorithm problems that allow for local or greedy algorithms can also be analyzed under random inputs. We leave this as an interesting open direction.

Acknowledgements. Research supported by the NSF Graduate Research Fellowship under Grant No. 1122374.

References

[1] D. Achlioptas and C. Moore. The asymptotic order of the random k-sat threshold. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 779–788. IEEE, 2002.

[2] D. Achlioptas and C. Moore. Random k-sat: Two moments suffice to cross a sharp threshold. *SIAM Journal on Computing*, 36(3):740–762, 2006.

[3] D. Achlioptas and Y. Peres. The threshold for random k-sat is 2k (ln 2-o (k)). In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 223–231, 2003.

[4] M. Badoiu, A. Czumaj, P. Indyk, and C. Sohler. Facility location in sublinear time. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, editors, *Automata, Languages and Programming*, pages 866–877, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[5] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[6] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[7] J. Bentley, D. S. Johnson, F. T. Leighton, C. C. McGeoch, and L. A. McGeoch. Some unexpected expected behavior results for bin packing. In *STOC '84*, 1984.

[8] J. Bentley, D. S. Johnson, T. Leighton, and C. C. McGeoch. An experimental study of bin packing. 1983.

[9] D. Bertsimas and L. H. Howell. Further results on the probabilistic traveling salesman problem. *European Journal of Operational Research*, 65(1):68–95, 1993.

[10] A. Beveridge, A. M. Frieze, and C. McDiarmid. Random minimum length spanning trees in regular graphs. *Combinatorica*, 18:311–333, 1998.

[11] F. Chung and L. Lu. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. American Mathematical Society, USA, 2006.

[12] E. G. Coffman, K. So, M. Hofri, and A. C.-C. Yao. A stochastic model of bin-packing. *Inf. Control.*, 44:105–115, 1980.

[13] A. Coja-Oghlan. A better algorithm for random k-sat. *SIAM Journal on Computing*, 39(7):2823–2864, 2010.

[14] A. Coja-Oghlan. A better algorithm for random k-sat. *SIAM Journal on Computing*, 39(7):2823–2864, 2010.

[15] C. Cooper, A. M. Frieze, N. Ince, S. Janson, and J. H. Spencer. On the length of a random minimum spanning tree. *Combinatorics, Probability and Computing*, 25:89 – 107, 2015.

[16] M. Dyer, A. Frieze, and C. McDiarmid. Partitioning heuristics for two geometric maximization problems. *Operations research letters*, 3(5):267–270, 1984.

[17] A. Frieze. On the value of a random minimum spanning tree problem. *Discrete Applied Mathematics*, 10(1):47 – 56, 1985.

[18] A. M. Frieze. On the value of a random minimum spanning tree problem. *Discret. Appl. Math.*, 10:47–56, 1985.

[19] A. M. Frieze and C. McDiarmid. On random minimum length spanning trees. *Combinatorica*, 9:363–374, 1989.

[20] A. M. Frieze, M. Ruszinkó, and L. Thoma. A note on random minimum length spanning trees. *Electron. J. Comb.*, 7, 2000.

[21] A. M. Frieze and J. E. Yukich. *Probabilistic Analysis of the TSP*, pages 257–307. Springer US, Boston, MA, 2007.

[22] P. Jaillet. *Probabilistic traveling salesman problems.* PhD thesis, Massachusetts Institute of Technology, 1985.

[23] S. Janson. The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph. *Random Struct. Algorithms*, 7:337–356, 1995.

[24] N. Karmarkar. Probabilistic analysis of some bin-packing problems. In *FOCS 1982*, 1982.

[25] R. M. Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Mathematics of Operations Research*, 2(3):209–224, 1977.

[26] R. M. Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Mathematics of operations research*, 2(3):209–224, 1977.

[27] T. Leipälä. On the solutions of stochastic traveling salesman problems. *European Journal of Operational Research*, 2(4):291–297, 1978.

[28] R. R. Mettu and C. G. Plaxton. The online median problem. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 339–348, Nov 2000.

[29] M. D. Penrose. Random minimal spanning tree and percolation on the n -cube. *Random Structures and Algorithms*, 12:63–82, 1998.

[30] N. Piersma. A Probabilistic Analysis of the Capacitated Facility Location Problem. *Journal of Combinatorial Optimization*, 3(1):31–50, July 1999.

[31] W. T. Rhee and M. Talagrand. A concentration inequality for the k-median problem. *Mathematics of Operations Research*, 14(2):189–202, 1989.

[32] P. W. Shor. The average-case analysis of some on-line algorithms for bin packing. *Combinatorica*, 6:179–200, 1986.

[33] J. M. Steele. On frieze's $\chi(3)$ limit for lengths of minimal spanning trees. *Discret. Appl. Math.*, 18:99–103, 1987.

[34] J. M. Steele. *Probability theory and combinatorial optimization*. Society for Industrial and Applied Mathematics, 1997.

[35] K. J. Supowit, E. M. Reingold, and D. A. Plaisted. The travelling salesman problem and minimum matching in the unit square. *SIAM Journal on Computing*, 12(1):144–156, 1983.

[36] K. J. Supowit, E. M. Reingold, and D. A. Plaisted. The travelling salesman problem and minimum matching in the unit square. *SIAM Journal on Computing*, 12(1):144–156, 1983.

[37] J. Wästlund. Notes on random optimization problems. http://www.math.chalmers.se/~wastlund/RandOptNotes.pdf, 2008.

[38] M. Weber and T. M. Liebling. Euclidean matching problems and the metropolis algorithm. *Zeitschrift für Operations Research*, 30(3):A85–A110, 1986.