Robust Statistical Geolocation of Internet Hosts

Zheng Wang and Brian L. Mark
Dept. of Electrical and Computer Engineering
George Mason University, MS 1G5
4400 University Drive, Fairfax, VA
email: zwang23@gmu.edu; bmark@gmu.edu

***Abstract*—Measurement-based approaches to geolocation of Internet hosts typically employ active probes sent from a predetermined set of landmarks to the target with unknown location. A significant improvement of such approaches can be made by constructing kernel density estimators using measurement data of end-to-end delays among all landmarks within the set. However, obtaining the data for this statistical approach to geolocation can be time-consuming, whereas the measurements are not all of high quality due to the time-varying characteristics of the network. To overcome these drawbacks, an ordinal optimization scheme is proposed to determine a subset of the landmarks that will yield improved geolocation accuracy with substantially less overhead. Given a set of kernel density estimates obtained by the landmarks, an optimization-based approach, with differential evolution as the search engine, is developed to find the location of the target. Numerical results demonstrate the efficiency, accuracy, and robustness of the proposed geolocation scheme.**

***Index Terms*—IP geolocation, ordinal optimization, kernel density estimator, differential evolution**

I. Introduction

A popular class of Internet geolocation techniques employs active measurements to infer the location of unknown targets using a set of fixed nodes, which are known as landmarks. The landmarks are distributed geographically and reside in the cloud. Thus, IP geolocation may be viewed as a cloud service or application. Delay measurements obtained by landmarks are used to estimate the location of the target Internet host. However, the delay measurements may lead to significant geolocation errors due to the nonlinearity and time-varying characteristics of the network. Shortest Ping [1], [2] is an early approach to geolocation whereby the location of the target is estimated by the location of the landmark with the minimum measured end-to-end delay to the target. However, using the measurement data from the landmarks without preprocessing can lead to a significant bias in the location estimates. A more advanced approach, GeoPing, improves upon Shortest Ping by introducing a group of passive landmarks. Nevertheless, GeoPing is still sensitive to the quality of the measurement data. As the number of passive landmarks grows, it will take much longer to acquire the measurements.

Another class of approaches formulates IP geolocation as a constrained optimization problem [3]. This requires a sufficient number of measurements from each landmark for the baseline-fitting problem, and the estimation accuracy is sensitive to both the accuracy and quantity of the measurements. In the CBDG scheme [4], circular constraints around the landmarks are defined through the measurements, which are maintained in a storage cloud and allocated to the landmarks. In [5], a two-stage scheme is proposed to map the passive landmarks into predetermined groups. In the first step, the approach tries to determine the possible groups that contain the target, then a mapping operation is applied to map the estimate to one of the landmarks. To do this, a large number of passive landmarks is required and the performance of the scheme depends heavily on the selection of the groups of landmarks. A database containing all the information of the landmarks needs to be maintained as well. The *Octant* framework [6] tries to define the possible region where the target may be located. Some additional constraints are introduced to find those regions in which the target is not likely to be located. Since the bounds of the target are defined, the approach can achieve higher accuracy. However, additional data related to DNS naming patterns and hop-delays are required.

The aforementioned Internet geolocation techniques may be classified as *deterministic* approaches, and may incur significant errors. To improve estimation accuracy and achieve lower complexity in algorithm implementation on real test platforms, a *statistical* approach to geolocation was proposed in [7]. This approach is based on the idea of kernel density estimators (KDEs) and relies solely on measurement data. In the first step, a set of KDEs is constructed using end-to-end delay measurements from all landmarks to some given landmarks. The KDEs are then maximized to obtain estimates of the distance from these landmarks to the unknown target [7], [8]. A force-directed algorithm is applied to search for the location of the target. Along similar lines, a spring-based approach was presented in [9], which aims to find an equilibrium state for the final estimate of the target estimation.

Statistical geolocation schemes can achieve relatively high efficiency and geolocation accuracy. However, they have several drawbacks: (1) To obtain a satisfactory estimation, a large amount of measurement data needs to be obtained before each KDE can be constructed, which incurs a significant overhead. (2) Not all of the landmarks provide measurement data of the same quality. The delay measurements may have large fluctuations even for the same landmark. Furthermore, our numerical results have shown that the force-directed algorithm proposed in [7] can have convergence issues and may introduce a substantial bias in the target location estimate.

In this paper, we propose a more robust approach to statistical geolocation that achieves higher estimation accuracy while incurring lower processing overhead. This approach consists of two phases. In the first phase of acquiring measurement data, ordinal optimization is applied to effectively cut down the redundant measurements and extract high-quality measurement data. In the second phase, an optimization-based approach with a differential evolution (DE) search engine is applied to search for the location of the target. The rest of the paper is organized as follows: Section II introduces the ordinal optimization algorithm and the framework of the proposed statistical geolocation scheme. In Section III, the optimization-based DE search scheme is developed and compared to the force-directed algorithm proposed in [7]. Section IV presents numerical results to demonstrate the performance of the proposed geolocation scheme, with comparisons to the statistical scheme in [7]. Concluding remarks are given in Section V.

II. KDE Construction via Ordinal Optimization

In this section, we apply a technique called ordinal optimization (OO) to effectively reduce the number of measurements required for IP geolocation and select the data with better quality to construct KDEs. The OO technique has been applied successfully to large-scale optimization problems with considerable uncertainties [10]. The OO approach follows two main principles: (1) Due to uncertainties in the experiments, we need to find some criteria to quantify the performance of these solutions in order to rank them. (2) If the candidate solutions can be ranked and sorted, we regard the higher-ranked solutions as the best solutions and neglect the lower-ranked solutions. In measurement-based IP geolocation schemes, servers called landmarks collect delay measurement data which is used to localize a given target host. The OO approach can be used to identify the best set of delay measurements to use for IP geolocation of a given host.

With respect to the geolocation problem, landmarks that are closer to the target are more likely to provide reliable measurements, since the geographic distance is much shorter. In other words, the quality of measurements depends strongly on the location of the corresponding landmark. End-to-end delays are assumed to be roughly proportional to geographic distances [9], [11]. This assumption has been verified in some regions such as North America [1], and it is most likely that the shortest delays come from the nearest landmarks [5]. We regard the landmarks that have lower end-to-end delay measurements to the target as ‘better’ landmarks. Another issue that needs to be considered is the stability of delay measurements. If the delay measurements of one landmark change very often from time to time, i.e., the variance of the measurements is large, we do not regard the landmark as good with respect to localizing the target. Hence, if we can pick such a subset of ‘good’ landmarks and use their measurement data to do the estimation, the performance of the statistical geolocation approach can be improved significantly.

Since both the value and stability of the delay measurements are critical to performance evaluation, the following analytical expression is proposed to represent the performance of the landmarks with respect to the geolocation problem:

$$Q = k_1 d_{avg} + k_2 \sigma^2, \tag{1}$$

where d_{avg} is the average landmark-to-target delay, σ^2 is the variance, and k_1 and k_2 are two properly selected constant scalars. This equation takes both the distance and the stability of measurements into consideration to quantify the overall performance and imposes a balance between them. Basically, landmark-to-landmark and landmark-to-target delay measurements are both acquired in the first phase. The statistics of the latter measurements indicate the geographic distance to the target and the stability of measurements. Thus, it is reasonable to choose the landmark-to-target measurements to compute Q. As the number of measurements grows, the statistics of a certain landmark should converge to steady-state values, and hence Q is suitable as a performance indicator in the ordinal optimization process.

In each iteration of the geolocation algorithm, landmarks with lower Q value are given a higher rank after each iteration, and the delay measurements from the top-ranking landmarks are used to construct KDEs after all the measurements are assigned. The KDE for a given landmark i is computed as follows:

$$\hat{f}_{i,\mathbf{H}}(g,d) = \frac{1}{M\det(\mathbf{H})} \sum_{j\in\mathcal{L}\setminus\{i\}} \sum_{l=1}^{m} \kappa((g-g_{ij}, d-d_{ij})\mathbf{H}^{-1}), \tag{2}$$

where $\mathcal{L}$ is the set of landmarks, M is the total number of measurements taken from landmark i, g_{ij} is the geographic distance between landmarks i and j, d_{ij} is the estimated delay between landmarks i and j, and $\mathbf{H}$ is a nonsingular bandwidth matrix given by Scott’s rules-of-thumb [12], [13]. The Gaussian kernel κ has the following form:

$$\kappa(g,d) = \frac{1}{2\pi} e^{-\frac{1}{2}(g^2+d^2)}, \tag{3}$$

More details on KDEs for IP geolocation can be found in [7].

In the statistical geolocation approach of [7], each landmark takes a large number of measurements, which is a time-consuming process. For a certain landmark, the time interval between every pair of consecutive measurements is fixed. To reduce the time overhead and select the better landmarks for geolocation, we propose to integrate OO into the statistical geolocation procedure. The general idea is to gradually allocate a predetermined measurement budget to these landmarks. At an initial stage, a certain number of measurements is assigned to each landmark to obtain a rough estimate, and an additional number of measurements is assigned in every following iteration to calibrate the value of Q. The loop terminates when the measurement budget is achieved. A typical allocation scheme is proposed in [14]. As the allocation process proceeds, the ranking of the landmarks stabilizes such that the best landmarks are selected.

Note that in the allocation process of OO, the landmarks are assigned different numbers of measurements. A larger portion of the measurements is allocated to the more promising landmarks. Since the landmarks can make delay measurements simultaneously, the duration of each iteration is limited by the maximum number of measurements assigned to any one of the landmarks. The measurement budget, the measurements at the initial stage, and the additional budget in the subsequent iterations can be regarded as control parameters for OO and they should be selected properly to avoid high time overhead. Some additional measurements can be performed for the top-ranked landmarks so as to reach the same maximum number in each iteration, since these will not introduce additional time overhead.

Once we have the data from the selected subset of landmarks, a KDE for each landmark in this subset can be constructed. Typically, a KDE can be illustrated on a distance-delay plane as a contour plot. By maximizing the KDEs, we can derive the corresponding distance estimates to the target. For a given landmark, if the distance estimate is g, then the target is assumed to lie on the border of the circle with its center at the landmark and a radius of g. In practical scenarios, the statistical nature of the measurements must be taken into account. The OO-based method for constructing the KDEs for statistical geolocation is summarized in Algorithm 1.

Algorithm 1 Phase 1: OO-based construction of KDEs.

Initialize $\mathcal{L}$ to be the set of all landmarks.
Initialize $\mathcal{B}$ to be the empty set.
Let N be the number of best landmarks to be chosen.
Initialize control parameters for OO.
while stopping criterion not met **do**
  Execute measurements as allocated to each landmark.
  for each landmark $i \in \mathcal{L}$ **do**
    Update the quality indicator Q_i
  end for
  Rank landmarks within $\mathcal{L}$ according to their Q values.
  Select N top-ranked landmarks to form the set $\mathcal{B}$.
  Compute measurement budgets for all landmarks in $\mathcal{L}$.
end while
for each landmark $i \in \mathcal{B}$ **do**
  Construct KDE$_i$.
  Compute average landmark-to-target delay d_i.
  Find maximum of KDE$_i$ for a fixed d_i.
end for

III. Optimization-Based Search Strategy

Once the KDEs for all of the best landmarks have been constructed, the second phase of our statistical geolocation approach applies a search strategy to find the location of the target. We propose an optimization-based search procedure, which employs differential evolution as the search engine. In this section, we first briefly discuss the force-directed algorithm originally proposed in [7], and then introduce the optimization-based approach.

A. Force-directed algorithm

In the force-directed algorithm (FDA) proposed in [7], the target location estimate is initialized to a suitable starting location, e.g., the landmark with the shortest mean delay to the target. A similar idea was also applied in a 'spring-based' approach [9] which uses forces to push the target estimate to a final position. In FDA, we iteratively apply a force on the target estimate proportional to the gradient of the estimated conditional distribution of distance from each landmark to the target given the delay. In each iteration, the resultant of the forces from all landmarks is calculated. The target location estimate is then moved according to the resultant force vector. The step sizes for the gradient ascent algorithm form a sequence that converges to zero. The canonical FDA of [7] suffers from the uncertainty of the measurements and may have convergence issues. An initial guess far from the selected landmarks can easily move away from the actual location of the target.

B. Optimization-based approach with differential evolution

The basic idea of the optimization-based approach is to maximize the conditional KDEs for each landmark simultaneously, which leads to a multi-objective optimization problem. To achieve a tradeoff between these objectives, a weight vector is used to construct an objective function, and this replaces the multi-objective problem with a single objective optimization problem. Such an objective function is given as follows:

$$\max_{\phi,\lambda} f(\phi,\lambda) = \sum_{i=1}^{|B|} w_i(d_{i\tau}) \cdot f_{G_i|D_i}(g_i|d_{i\tau}) \tag{4}$$

where $w_i(d_{i\tau})$ is the ith term of the weight vector, and the geographic distance g_i can be derived using the inverse Vincenty formula (see [7]). It is clear that the weight vector must be selected properly. In our experiments, the weight $w_i(d_{i\tau})$ is set to a value inversely proportional to the mean delay, $(d_{i\tau})$, from the ith landmark to the target. For simplicity, we have set $w_i(d_{i\tau}) = 1/d_{i\tau}$. The objective function thus becomes:

$$\max_{\phi,\lambda} f(\phi,\lambda) = \sum_{i=1}^{|B|} \frac{1}{d_{i\tau}} \cdot f_{G_i|D_i}(g_i|d_{i\tau}) \tag{5}$$

Next we discuss the search engine employed in the searching process. Since our aim is to determine the longitude and latitude estimate of the target within a 2-dimensional continuous searching space, standard nonlinear optimizers such as gradient descent can be applied. However, as the landscape of the objective function can be rather complex in practical scenarios, heuristic approaches such as evolutionary computation (EC) becomes a more competitive choice. Differential evolution (DE) is one of the most powerful and robust search engines in EC class for continuous optimization problems [15]. The DE method uses differential operators to generate new candidate solutions for further searching. A population of N candidate vectors $(x^1, x^2, ...x^N)$ is maintained in each iteration. Each candidate vector consists of d dimensions. One strategy in the family of DE algorithms, called DE/best/1, first produces a donor vector v as follows:

$$v = x^* + F \cdot (x^{r_1} - x^{r_2}), \tag{6}$$

where r_1 and r_2 are two mutually different indices randomly selected from the set $1, 2, ..., N$, and x^* is the best vector within the current population. The vector $x^{r_1} - x^{r_2}$ is called the mutant vector, and F is referred to as the scaling factor. Then a crossover operator is applied to generate the offspring vector u:

$$u_j = \begin{cases} v_j, & \text{if } U \leq \gamma \text{ or } j = J, \\ x_j, & \text{otherwise}, \end{cases} \tag{7}$$

where u_j denotes the jth dimension of u, U is a uniform random variable within range $[0, 1]$, γ is known as the crossover rate, and J is random number chosen from $1, \cdots, d$. After crossover, a pairwise selection is performed to choose the better population vector between x^r and u. More guidance for parameter settings and strategy selection for the DE method can be found in [15].

IV. Experimental Results

A. Network model and parameter settings

In a real network, the end-to-end delay experienced by a packet sent from one node to another is a random variable with an unknown distribution. The variability is caused by several factors, including the presence of multiple alternative paths and random queueing delays. In a moderately connected area, the routers and links are assumed to be uniformly distributed and have a sufficiently high density, and the distance-delay relation is expected to exhibit a linear characteristic. Hence, from a statistical view, we can roughly model the delay measurement T_d as follows:

$$T_d = \frac{K \times g}{c} + W, \tag{8}$$

where g is the actual geographic distance from one node to another, c is the speed of light, and K and W are Gaussian random variables. Due to the homogeneous propagation property of optical fiber, the delay is assumed to be approximately proportional to the actual distance. This latency model is also adopted in some other approaches [9], [11], [16]. The deterministic baseline proposed in [3] has a similar expression as well. These approaches often try to estimate the parameters in the model by doing curve fitting. However, in our approach, such estimation is not necessary and we can directly use the measurements to search for the target. The measurement results are accumulated on each of the landmarks for a future selection process.

A real-world network for IP geolocation is depicted in Fig. 1, in which the asterisks represent the positions of landmarks distributed over the continental U.S. These landmarks are real Internet servers which are interconnected through other nodes and routers in the network. To experimentally study the performance of the proposed approach through simulations, we randomly generate a total number of $N = 50$ landmarks within a 2-dimensional searching area and apply the proposed OO-assisted statistical geolocation approach to find the position of the target. For the delay model, the parameter settings are $K \sim \mathcal{N}(1.2, 0.2/3)$ and $W \sim \mathcal{N}(1, 1/3)$. For

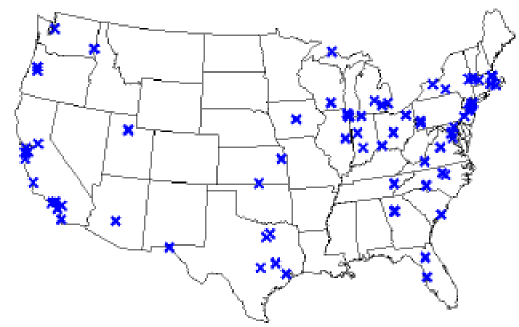

Fig. 1. The distribution of landmarks in U.S. continent.

simplicity, we take $c = 3$. For the ordinal optimization process, we take $k_1 = 0.8$ and $k_2 = 0.2$, and the maximum number of landmarks within the best set is set to $B = 10$. The KDEs are constructed using the measurement data from these 10 landmarks. The distribution of the landmarks and target is shown in Fig. 2, in which the landmarks are denoted by asterisks, and the target is denoted by a red circle. The landmarks shown in red are those selected by the OO technique.

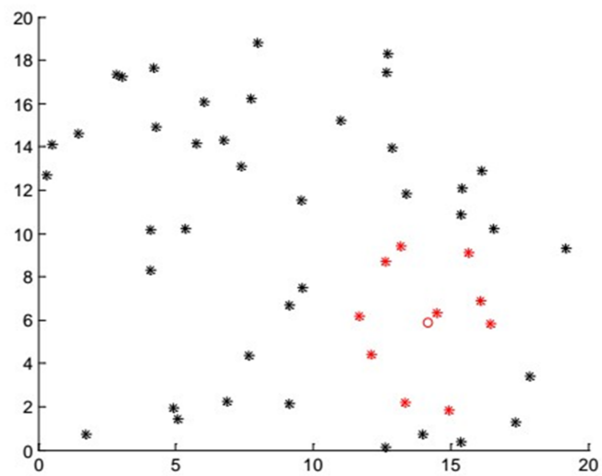

Fig. 2. Distribution of landmarks and target.

A typical contour plot of a KDE is given in Fig. 3. In this plot, we show the KDE constructed using one landmark. The lower figure is the KDE constructed from all measurement data, while the upper one is the KDE constructed from the 10 best landmarks. We can see that the KDE constructed from full measurement data has one local optimum and one global optimum. The global optimum of the KDE derived from the OO-based approach is very close to that of the original statistical approach, and very little degradation is incurred. In the above experiment without an OO procedure, we set the number of measurements for each landmark as 1000. Thus, a total number of $50 \cdot 50 \cdot 1000 = 2.5 \times 10^6$ measurements are needed to construct all of the KDEs. The number of measurements with an OO technique is only $10 \cdot 10 \cdot 350 = 3.5 \times 10^4$. This greatly reduces the time overhead and computational expense for the searching phase, and the time spent on obtaining this data is approximately 35% compared with the approach proposed in [7].

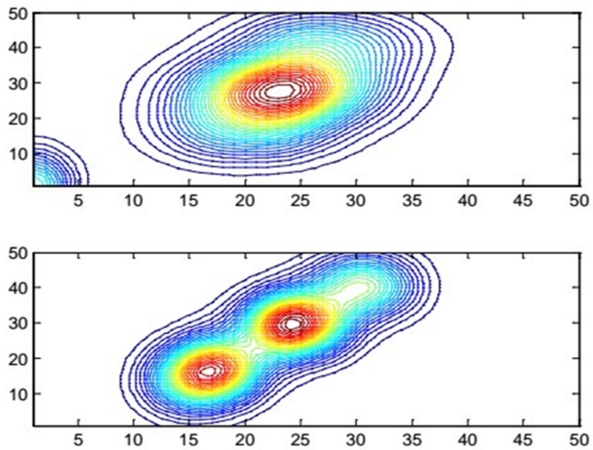

Fig. 3. Typical contour plot of a KDE.

B. Results using force-directed algorithm

For the second phase of the entire searching process, we first investigate the performance of the FDA strategy. By fixing the delay as the average landmark-to-target delay, the conditional KDEs of the 10 best landmarks are derived and are shown in Fig. 4. It is clear that each conditional KDE has a peak and the target is assumed to lie on the circle with a radius that maximizes the corresponding conditional KDE. We randomly chose 100 starting points for the target location estimate. The trajectories of the location estimates are shown in Fig. 5. In this experiment, the trajectories corresponding to 12 out of the 100 starting points did not converge to a point near the target. The mean and variance of the location estimation area for the remaining tests were 8.46 and 3.47×10^{-4}, respectively.

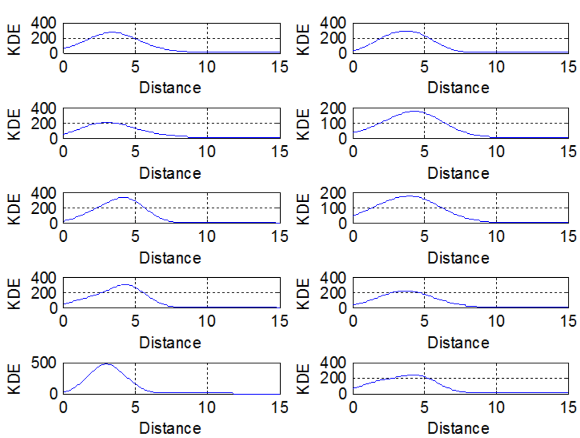

Fig. 4. Conditional distribution of distance in KDEs of all the best landmarks.

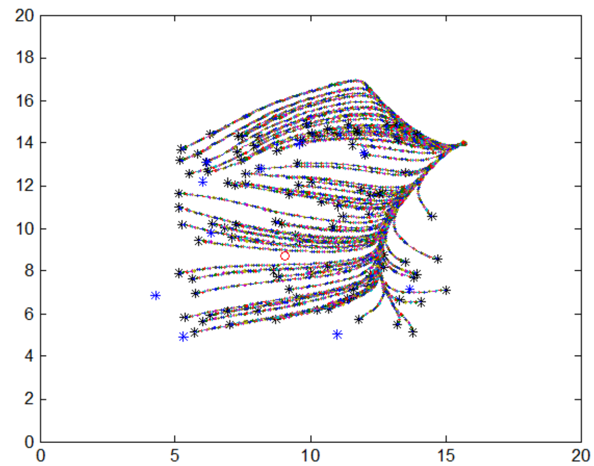

Fig. 5. Convergence curves for 100 different starting points using force-directed algorithm.

C. Results using optimization-based approach

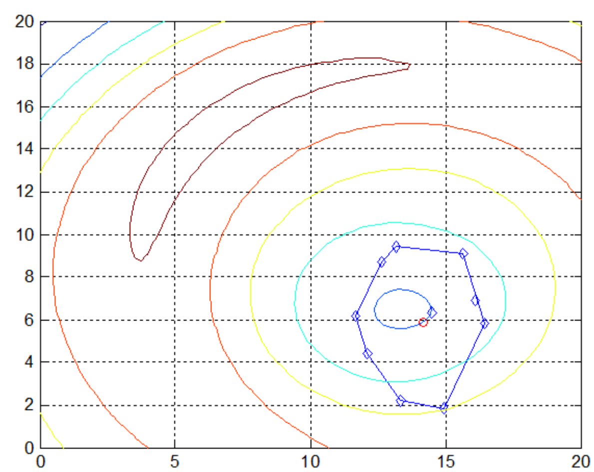

Fig. 6. Contour plot of weighted linear objective function.

In this approach, we first test the same group of landmark settings and measurement data as in FDA. The contour plot of the weighted objective function given by (5) is shown in Fig. 6. It is clear that the global optimum is quite close to the true target location, and this shows the effectiveness of the weighted objective function. The DE algorithm using the DE/best/1 strategy is applied to search for the target location. In all of our experiments, the DE method converged on the global optimum solution. Since the DE method is a population-based search engine, the time spent on a single run is longer than that FDA, but much shorter relative to the first phase.

We performed a total number of 10 runs using the optimization-based approach. In each run, the maximum number of DE iterations was set to 100. In contrast to the FDA

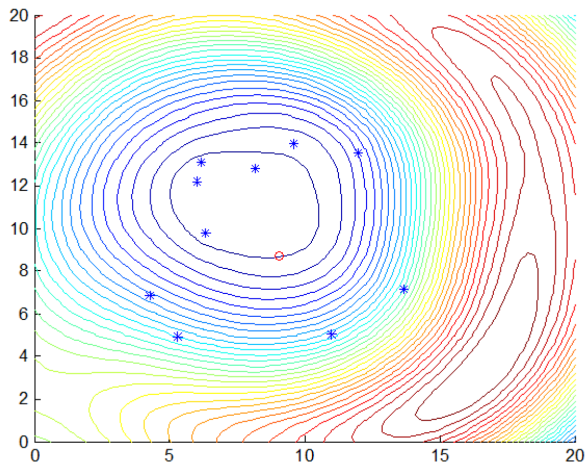

Fig. 7. Contour plot of weighted linear objective function for the second group of settings.

approach, *all* of the 10 runs converged to an estimate close to the true target location. The mean and variance of the estimation error for these 10 runs were 2.34 and 7.33×10^{-7}, respectively. The contour plot of the objective function for a second experiment with randomized landmark/target settings is shown in Fig. 7. In this case, the global optimum is much closer to the real target compared to the convergence region of the FDA approach when it is applied with the same landmark/target settings.

D. Comments

Our numerical experiments have shown that the FDA approach frequently does not converge when initialized from random starting points, whereas the optimization-based approach always converges. When the FDA approach does converge, the mean geolocation error achieved is typically similar to or worse than that of the proposed optimization-based approach. On the other hand, the variance of the geolocation error of the FDA approach was always found to be larger than that of the optimization-based approach.

V. Conclusion

We proposed a two-phase scheme for robust statistical IP geolocation. The first phase incorporates ordinal optimization to reduce the time and computation overhead of collecting statistical measurement data from landmarks in the network. The second phase consists of a differential evolution engine which solves a global optimization problem formulated to an estimate of the target location based on the statistical measurement data. Our numerical experiments show that the proposed approach uses only a small portion of the total measurements used in the statistical geolocation scheme of [7] and requires about 35% of the total computation time. In searching for the target location, the weighted optimization-based approach obtains results of similar or better estimation accuracy regardless of the starting point and the measurement data used. Unlike the force-directed algorithm proposed in [7], the differential evolution solver always converges to a fixed target location estimate.

References

[1] W. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *ACM SIGCOMM'01*, San Diego, CA, Aug. 2001.

[2] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Internet Measurement Conference (IMC'09)*, Rio de Janeiro, Brazil, Oct. 2006.

[3] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Trans. Networking*, vol. 14, no. 6, pp. 1219–1232, Dec. 2006.

[4] M. Gondree and Z. N. Peterson, "Geolocation of data in the cloud," in *Proceedings of the third ACM conference on Data and application security and privacy (CODASPY'12)*, San Antonio, TX, Feb. 2013.

[5] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "IP-geolocation mapping for moderately connected Internet regions," *IEEE Trans. on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 381–391, Feb. 2013.

[6] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A comprehensive framework for the geolocalization of internet hosts," in *4th USENIX Symposium on Networked Systems Design and Implementation (NSDI'07)*, Cambridge, MA, Apr. 2007.

[7] I. Youn, B. L. Mark, and D. Richards, "A statistical approach to geolocation of Internet hosts," in *Proc. 18th IEEE Int. Conf. on Computer Communications and Networks (ICCCN'09)*, San Francisco, CA, Aug. 2009.

[8] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," in *Proc. Int'l Conf. Passive and Active Measurement (PAM'10)*, Zurich, Switzerland, Apr. 2010.

[9] S. M. Gunther, J. Schlamp, and G. Carle, "Spring-based geolocation," in *IEEE/IFIP Network Operations and Management Symposium (NOMS'12)*, Maui, HI, Apr. 2012.

[10] Y. C. Ho, Q. C. Zhao, and Q. S. Jia, *Ordinal Optimization: Soft Optimization for Hard Problems*. Springer, 2007.

[11] S. Laki, P. Matray, P. Haga, T. Sebok, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *The 24th International Conference on Advanced Information Networking and Applications (AINA'10)*, Perth, Australia, Apr. 2010.

[12] B. Silverman, *Density estimation for statistics and data analysis*. Chapman Hall/CRC, 1986.

[13] D. Scott, *Multivariate ensity estimation: theory, practice, and visualization*. Wiley-Interscience, 1992.

[14] D. Chen, J. Lin, E. Yucesan, and S. Chick, "Simulation budget allocation for further enhancing the efficiency of ordinal optimization," *Discrete Events Dynamic Systems*, vol. 10, no. 3, pp. 251–270, Feb. 2000.

[15] K. Price, R. Storn, and J. Lampinen, *Differential Evolution: a Practical Approach to Global Optimization*. Springer-Verlag New York Inc., 2005.

[16] M. J. Arif, S. Karunasekera, S. Kulkami, A. Gunatilaka, and B. Ristic, "Internet host geolocation using maximum likelihood estimation technique," in *The 24th International Conference on Advanced Information Networking and Applications (AINA'10)*, Perth, Australia, Apr. 2010.