Approximating Global Optimum for Probabilistic Truth Discovery*

Shi Li, Jinhui Xu, and Minwei Ye

State University of New York at Buffalo
{shil,jinhui,minweiye}@buffalo.edu

Abstract. The problem of truth discovery arises in many areas such as database, data mining, data crowdsourcing and machine learning. It seeks trustworthy information from possibly conflicting data provided by multiple sources. Due to its practical importance, the problem has been studied extensively in recent years. Two competing models were proposed for truth discovery, weight-based model and probabilistic model. While $(1+\epsilon)$-approximations have already been obtained for the weight-based model, no quality guaranteed solution has been discovered yet for the probabilistic model. In this paper, we focus on the probabilistic model and formulate it as a geometric optimization problem. Based on a sampling technique and a few other ideas, we achieve the first $(1+\epsilon)$-approximation solution. The general technique we developed has the potential to be used to solve other geometric optimization problems.

Keywords: geometric optimization, truth discovery, high-dimension, data mining

1 Introduction

Truth discovery has received a great deal of attention in recent years in databases, data crowdsourcing, machine learning and data mining [16,13,9,10,14]. It emerges from various practical scenarios such as copying detection [5], data fusion [3] and conflicting information resolving on the web [16]. In a typical scenario, the unknown truth for one or multiple objects can be viewed as a vector in a high-dimension space. The information about the truth vector may come from multiple sources. Those sources may be inaccurate, conflicting or even biased from the beginning if they come from subjective evaluation. Our goal is to infer the truth vector from these noisy information.

A naive method for this problem is to take the average of all the vectors from sources as the the ground truth (for coordinates correspondent to categorical data, take the majority vote). However, this approach, which inherently treats all sources as equally important, is vulnerable to unreliable and malicious sources.

* The research of the first author was supported in part by NSF grants CCF-1566356 and CCF-1717134. The research of the last two authors was supported in part by NSF through grants CCF-1422324, IIS-1422591, and CCF-1716400.

Such sources can provide information that pulls the average away from the truth. A more robust type of approaches is to give weights to sources to indicate their reliability and use the weighted average or weighted majority as the ground truth. However, since the weights are often unknown, the goal of finding the ground truth is coupled with the task of reliability estimation. This type of approaches is referred as a *truth discovery* approach. Among all, there are two competing and sometimes complementary frameworks that are widely accepted and used for different data types.

Weight-based Truth Discovery In this framework, both the truth and the weights are treated as variables. An objective function is defined on these variables [10]. Then an alternating minimization algorithm can be used to solve the problem. In each iteration, the algorithm fixes one set of variables (either the truth variables, or the weight variables) and optimizes the other. This procedure continues until a stable solution is reached. Many existing methods [16,4,7,11] follow this framework and justify themselves by experimenting with different types of real-world datasets. However, none of these methods provides any theoretical guarantee regarding the quality of solution. Recently, Ding et al. [2] gave the first algorithm that achieves a theoretical guarantee (*i.e.,* a $(1+\epsilon)$-approximation) for a well-known weight-based model of truth discovery introduced in [10]. Later, Huang et al. [19] further improved the running time to near quadratic.

Probabilistic Truth Discovery Probabilistic models lie in a different category of models for truth discovery. They were also studied extensively in the literature [17,15,12,18]. Instead of giving weights to indicate the reliability of all sources, these models assume that the information for each source is generated independently from some distribution that depends on the truth and the reliability of the source. Then the goal under these models is to find the truth that maximizes the likelihood of the generated information from all sources. The probabilistic models have been shown to outperform the weight-based methods on numerical data [17]. They also prevail other models in the case where sources come from subjective evaluation [13]. For the qualify of the optimization, [15] gave an iterative algorithm with guaranteed fast convergence to a local optimum.

1.1 Our Results

We propose a probabilistic truth discovery model, reformulate it as an optimization problem and give a PTAS (Polynomial-Time Approximation Scheme) to solve it. We assume that each observation of a source is generated around the truth vector with variance corresponding to the reliability of the source. Then, the goal of finding the truth vector with the maximum likelihood can be formulated as an optimization problem. Instead of directly solving the optimization problem, we convert it to the following more general geometric optimization problem:

$$\text{Given } \{p_1, p_2, \cdots, p_n\} \subset \mathbb{R}^d, \text{ find } x \in \mathbb{R}^d \text{ to minimize } \sum_{i=1}^{n} f(\|x - p_i\|),$$

where f is a function satisfying some reasonable properties.

This general problem encloses as special cases the classic 1-median and 1-mean problems, and the more general problem of minimizing p-th power of distances. Moreover, by considering the corresponding functions with an upper-threshold, i.e, $f(\ell) = \min\{\ell, B\}$, $f(\ell) = \min\{\ell^2, B\}$ and $f(\ell) = \min\{\ell^p, B\}$, one can capture the outlier versions of all these problems.

We give a sampling-based method that solves the above optimization problem up to a factor of $1+\epsilon$ for any $\epsilon > 0$ in quadratic running time. Thus, it not only solves our truth discovery problem but also gives a unified approach to solve all the above problems under this framework.

1.2 Our Techniques

One property that we *do not* impose on the function f is convexity. Requiring f to be convex will make our problem too restrictive. For example, the cost function f_{truth}(defined later) is non-convex in our truth discovery problem. The threshold functions that are used to model the outlier versions of the 1-center problems are also non-convex. Without the convexity property, iterative approaches such as gradient descent and EM do not guarantee the global optimality. General coreset technique (such as the one in [6]) which reduces the size of the problem will not work, either. The dimensionality is not reduced by those techniques so that the problem is still hard even for the coreset.

Instead of using methods in continuous optimization or general sampling technique, our algorithm is based on the elegant method Badoiu, Har-Peled and Indyk developed to give fast algorithms for many clustering problems [1,8]. Roughly speaking, [1] showed that a small set of sample points X can guarantee that the affine subspace span(X) contains a $(1+\epsilon)$ approximate solution for these clustering problems. Therefore both the size and the dimensionality can be reduced.

Directly applying [1] does not work for non-convex cost function. In this paper, we extend [1] to a more general family of cost functions, including the non-convex cost function for our truth discovery problem. We will elaborate the challenges in Section 2.2.

2 Problem formulation and Main Results

2.1 Probabilistic Truth Discovery

We first set the stage for the problem. The unknown truth can be represented as a d dimensional vector p^*, as justified in [10]. There are n sources, and the observation/evaluation made by the i-th source is denoted as p_i which also lies in the d dimensional space $\mathbb{R}^d$. In our model, we assume that each observation/evaluation is a random variable following a multi-variate Gaussian distribution centered at

the truth p^* with covariance $\sigma_i^2 I_d$.[1] Each unknown parameter $\sigma_i \geq 0$ represents the reliability of the source; the smaller the variance, the more reliable the source is.

We formulate the problem as finding the $(p^*, \sigma = (\sigma_i)_{i\in[n]})$ that maximizes the likelihood of the random procedure generating p^*.We impose a hyper-parameter $\sigma_0 > 0$ and require $\sigma_i \geq \sigma_0$ for every $i \in [n]$. It is naturally interpreted as an upper bound of the reliability of all sources, but there is another interpretation that we will discuss later.

Given the set of observation $P = \{p_i\}_{i=1}^n \subset \mathbb{R}^d$ under this probabilistic model and a hyper-parameter σ_0, we need to find a point x that maximizes the following likelihood function:

$$\prod_{i=1}^{n} \mathcal{N}(p_i \mid x, \sigma_i^2 I_d) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^d \exp\left[-\frac{\|p_i - x\|^2}{2\sigma_i^2}\right].$$

Taking negative logarithm and optimizing the quantity over all valid vectors $\sigma = (\sigma_i)_{i\in[n]}$, we obtain the following optimization problem:

$$\min_{x\in\mathbb{R}^d,\sigma} \left\{ \frac{nd}{2}\ln(2\pi) + \sum_{i=1}^{n}\left(d\ln\sigma_i + \frac{\|p_i - x\|^2}{2\sigma_i^2}\right)\right\}, \quad \text{s.t.} \quad \sigma_i \geq \sigma_0, \forall i \in [n]. \tag{1}$$

Lemma 1. *For a fixed $x \in \mathbb{R}^d$, the following vector σ minimizes the objective function in* (1)*:*

$$\sigma_i = \max\left\{\sigma_0, \|p_i - x\|/\sqrt{d}\right\}, \quad \forall i \in [n].$$

Applying Lemma 1, the optimization problem now only depends on the point $x \in \mathbb{R}^d$:

$$\min_{x\in\mathbb{R}^d} \left\{ \frac{nd}{2}\ln(2\pi) + \sum_{\|p_i - x\| < \sigma_0\sqrt{d}} \left(\frac{\|p_i - x\|^2}{2\sigma_0^2} + d\ln\sigma_0\right) + \sum_{\|p_i - x\| \geq \sigma_0\sqrt{d}} \left(\frac{d}{2} + d\ln\frac{\|p_i - x\|}{\sqrt{d}}\right)\right\}.$$

Notice that scaling x, σ_0 and all points p_i by a fact of c only changes the value of the function by a constant additive term $(nd\ln c)$. For simplicity, we will apply a scaling to the triple $(x, \sigma_0, \{p_i\}_{i=1}^n) \mapsto (x', \sigma_0'.\{p_i'\}_{i=1}^n)$ so that $\sigma_0' = 1/\sqrt{d}$ and

[1] For categorical data, the Gaussian distribution may cause fractional answers, which can be viewed as a probability distribution over possible truths. In practice, variance for different coordinates of the truth vector may be different and there might be some non-zero covariance between different coordinates; however, up to a linear transformation, we may assume the covariance matrix is $\sigma_i^2 I_d$.

drop the prime symbol if there is not ambiguity. The objective function becomes:

$$\min_{x\in\mathbb{R}^d}\left\{\frac{nd}{2}\ln(2\pi)+\sum_{\|p_i-x\|<1}\left(\frac{d\|p_i-x\|^2}{2}-\frac{d\ln d}{2}\right)\right.$$
$$\left.+\sum_{\|p_i-x\|\geq 1}\left(d\left(\frac{1}{2}+\ln\|p_i-x\|\right)-\frac{d\ln d}{2}\right)\right\}.$$

Moreover, we can drop the constant term $\frac{nd}{2}\ln(2\pi)-\frac{nd}{2}\ln d$, and then divide the whole function by $d/2$, the final optimization problem becomes:

$$\min_{x\in\mathbb{R}^d}\sum_{i=1}^{n} f_{truth}(\|x-p_i\|) \quad \text{where} \quad f_{truth}(\ell)=\begin{cases}\ell^2 & 0\leq\ell<1\\ 1+\ln\ell^2 & \ell\geq 1\end{cases}. \tag{2}$$

This objective function can be seen as the summation of costs from each individual point. The cost function f for each p_i is quadratic when its distance to the variable p is close, and it grows logarithmically when p_i is far away.

The function $\sum_{i=1}^{n} f_{truth}(\|x-p_i\|)$ can be served as an alternative way of evaluating the solution's quality other than the negative log-likelihood since:
(1) It has non-negative objective function value so that multiplicative approximation factor can be properly defined, which serves as a criterion of the solution's quality.
(2) The $(1+\epsilon)$ approximation of $\sum f_{truth}$ gives the following guarantee. Let $Q_0=\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^d$ be the maximum possible likelihood for the optimum solution of *any instance* with n points and d dimensions. Let Q^* be the likelihood for the optimum solution to the given instance. If $Q^*=Q_0e^{-t}$, then we shall give a solution with likelihood at least $Q_0e^{-(1+\epsilon)t}$.

Interpretation of the Parameter σ_0 σ_0 in our model is introduced to reflect the overall reliability of the dataset. If each σ_i is unconstrained, or in other words $\sigma_0=0$, then quantity (1) can tend to $-\infty$ by letting $p_i=x$ and $\sigma_i\to 0$ for some $i\in[n]$. At this point, it may seem that the introduction of the parameter σ_0 is a little bit unnatural. However, we argue that this issue caused by the singular solutions does not only exist in our model; it comes with the truth discovery problem itself. If one does not impose any assumption on the reliability of the sources, then a solution (in any model) can be: one source is 100% reliable, all the other sources are not reliable at all and the truth is the data given by the reliable source. Such a model will not be general enough. Any meaningful model needs to be able to capture more than this type of solutions.

With the understanding that σ_0 gives an upper bound on the reliability of the sources, we can discuss how σ_0 affects the optimum solution of our problem. In one extreme, σ_0 is very small, meaning that any source can be very reliable. Then in our final optimization problem (**??**), the points p_i's are far away from each other. (Recall that to obtain (**??**), we scaled the original $p_i\mapsto p_i'=\frac{p_i}{\sqrt{d}\sigma_0}$.) Then for a typical center point x, most p_i's will have large $\|p_i-x\|$. For these points, the f values are logarithmic in their distances to the center and thus are

very insensitive to the location of the center. In this case, the optimum solution x will be very close to some input point p_i.

Consider the other extreme where σ_0 is very large. Then, the points p_i are close to each other. In this case, the cost function will be distance square when x is close to all points. The problem then becomes the classic 1-mean problem. This coincides with our intention of setting the "overall confidence" σ_0: σ_0 being very large indicates that all sources are unreliable when considered alone, and it is wiser to take the average than to favor a particular source.

It might seem unreasonable to set a hyper parameter in "truth discovery" problem because "truth" is usually assumed to be invariant to some hyper-parameter we select in our model. Indeed, the truth should be invariant if it is a numerical fact such as the height of a mountain or today's weather forecast at some location. But if we are talking about the rating of a movie or evaluation of an instructor, it is presumptuous to suggest that there exists some "truth discovery" model which can somehow "calculate" such truth exactly or approximately. In such setting, the best we can guarantee is providing a model that can rule out some outliers for the users. The hyper-parameter is provided for the users to decide how much portion of the sources are outliers to him/her.

Here we present our main result for probabilistic truth-discovery problem. It is directly implied by our main theorem, Theorem 2.

Theorem 1. *Let $0 < \epsilon \leq 1$. Let P be a set of n points in $\mathbb{R}^d$ and $G(x) = \sum_{p \in P} f_{truth}(\|x - p\|)$. A $(1+\epsilon)$-approximate solution can be obtained in time $O(2^{(1/\epsilon)^{O(1)}} d + n^2 d)$.*

3 Solution for General 1-Center Optimization Problem

3.1 General description of the algorithm

The following notations are used throughout this section. Given the point set $P \subset \mathbb{R}^d$, a cost function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$, let $G(x) = \sum_{p \in P} f(\|x - p\|)$ denote the objective function. We reuse the variable p_{opt} as the optimizer of $G(x)$.
We show in advance the following three properties that a general cost function f need to satisfy in order to apply our extended sampling method.

Property 1. (Regularity) f is a continuous, non-negative, monotonically increasing function.

Property 2. (Sub-proportionality)[2] $\exists \alpha \geq 1 : f(kx) \leq k^\alpha f(x)$ for any $k \geq 1$,$x \geq 0$. We say α is the *proportional degree* of f if it is the smallest α satisfies such property.

Property 3. The function f can be computed in polynomial time with respect to the size of the input. The inverse of f, defined as $f^{-1}(y) = \sup_x\{x : f(x) = y\}$, should also be able to calculate in polynomial time w.r.t to the size of x when $y \leq 2f(x)$.

[2] Also referred as polynomial growing function or Log-Log Lipschitz function in literature.

Remark 1. The only place property **??** is used is in Theorem 2. It is imposed to ensure a polynomial running time in arithmetic calculation.

Remark 2. Continuity can be implied by property **??** by taking $k \to 1$. Also, by taking $x = 0$ in property **??**, one can infer that $f(0) \geq 0$. With the fact that f is non-decreasing, one can also infer that f is non-negative.

Thus essentially the first two properties are (i) monotonically increasing, which is a common assumption when a function is referred as a "cost" function; (ii) sub-proportionality, which can be roughly thought of as requiring the function not growing exponentially. Intuitively speaking, an equivalent statement is that for every $a > 0$, the graph of the unique function $g(x) = Cx^{\alpha}$ going through $(0, 0)$ and $(a, f(a))$ is completely above (can overlap) the graph of $f(x)$ when $x \geq a$.

From now on, these three properties are always assumed for a cost function f unless stated otherwise.

To approximate the optimizer p_{opt} of $G(x)$, we generalize an existing result from Badoiu, Har-Peled and Indyk [1] (for convex functions) to our problem where the function can be non-convex. The key idea is to sample a core-set X from the input points P such that the affine subspace $\text{span}(X)$ contains a $(1+\epsilon)$-approximate solution. We summarize the method in a general way in the following procedures:

1. The value L is chosen so that the following two things can both happen:
   (a) It's possible to sample a few points and guarantee that with constant probability, the Euclidean distance from one of the sample is close enough to the optimizer p_{opt}, i.e. $\|s_i - p_{opt}\| \leq L$ for some sample s_i.
   (b) If the distance from p' to p_{opt} is $O(\epsilon L)$ for sufficiently small constant in this big O notation, p' is guaranteed to be $(1+\epsilon)$ approximate solution.
2. Continue the sampling in batches so that for each batch of samples, either the $(1+\epsilon)$-approximate solution is already in the affine subspace spanned by the sampled points, or the subspace becomes closer to p_{opt} by a factor about $1+\epsilon$. It is also required that the size of each batch is $\text{poly}(1/\epsilon)$.
3. Repeat step 2 until the distance from $\text{span}(X)$ to p_{opt} is smaller than $O(\epsilon L)$, where X is the set of sampled points.
4. Inside $\text{span}(X)$, draw a grid around each point in X. The radius of the grid is $2L$ and the side length is ϵL. Then there is an $(1+\epsilon)$ approximate solution in these grid points.

Remark 3. To be able to shorten from the initial gap L to the desired gap $O(\epsilon L)$, the number of batches required on average is bounded by $\text{poly}(1/\epsilon)$, which means it only depends on the approximation factor. Since each batch contains $\text{poly}(1/\epsilon)$ many samples, in total the sample set X is of size $\text{poly}(1/\epsilon)$.

Remark 4. Notice that in Step 4 we need to approximately know the value L to perform the actual algorithm. This is guaranteed in our algorithm for general cost function f, as we showed in the next section.

3.2 The choice of L

Let us first focus on the choice of L for a general cost function f. Denote $\texttt{AVG} = G(p_{opt})/n$. If $f(x) = x$, L can be chosen to be $2G(p_{opt})/n$ in Step 1, as shown in [1]. We can think of this L as the average cost contributed from points in P. So for condition (b), it is trivial that $\epsilon L/2$ is the necessary distance from p' to p_{opt} to make p' a $(1+\epsilon)$-approximated solution. At the same time, L is also roughly the "average" of Euclidean distance from each point in P to p_{opt} since the cost function f is an identity function. So for condition (a), a point $s \in P$ such that $\|s - p_{opt}\| \leq L$ can be regarded as an "average" case. An average case is easy to approximate using sampling.

However, such coincidence will not happen for general f. If f is a slowly growing function (e.g. $log(x)$, $1 - 1/x$) and L is chosen like above, condition (b) still holds but L is far from the "average" of Euclidean distances to p_{opt} in some of worse cases. To compromise, we do not require L to be "average". We only require roughly ϵn points in P satisfying that the distance from them to p_{opt} is less than L. Then on average, we can obtain such point after $O(1/\epsilon)$ samples. Consequently, condition (a) and (b) can both be satisfied again. The following lemma shows the exact choice of the value L, the unknown variables A and B will be removed later:

Lemma 2. *Let $0 < \epsilon \leq 1$. Let $P \subset \mathbb{R}^d$ and $|P| = n$, $G(x) = \sum_{p \in P} f(\|x - p\|)$ with α as the proportion degree of f. Suppose $\tilde{p}$ is the $\lceil \epsilon n \rceil$-th closest point to the optimal solution p_{opt} among the points in P. Choose L accordingly if the following two cases apply:*
(i) If we know a value A such that $f(\|\tilde{p} - p_{opt}\|) \in [A, (1+\epsilon/3)A)$, choose $L = f^{-1}((1+\epsilon)A)$.
(ii) If $f(\|\tilde{p} - p_{opt}\|) \leq \epsilon \texttt{AVG}/B$ for some constant $B \geq 3$, choose a value $L \in [f^{-1}(\epsilon \texttt{AVG}/B),\ f^{-1}(\epsilon \texttt{AVG}/3)]$.
Then p' is a $(1+\epsilon)-$approximate solution of G if $\|p' - p_{opt}\| \leq \epsilon L/(4\alpha)$.

Proof. We prove case (i) first. Let $P = \{p_1, p_2, \cdots, p_n\}$ so that $\|p_1 - p_{opt}\| \leq \|p_2 - p_{opt}\| \leq \cdots \leq \|p_n - p_{opt}\|$. Then $i \geq \lceil \epsilon n/4 \rceil$ implies $f(\|p_i - p_{opt}\|) \geq A$ since f is non-decreasing. By Markov's inequality the value A can not be greater than $\texttt{AVG}/(1 - \epsilon/4)$. This is a fact we are going to use in the following argument and later in Lemma 5. Now assume p' is a point satisfies $\|p' - p_{opt}\| \leq \epsilon L/(4\alpha)$. For p_i with $i < \lceil \epsilon n/4 \rceil$, the total increase of cost by moving p_{opt} to p' is at most

$$\begin{aligned}
&\sum_{i < \lceil \epsilon n/4 \rceil} f(\|p_i - p'\|) \leq \epsilon \frac{n}{4} f(L + \frac{L\epsilon}{4\alpha}) \\
&\leq \epsilon \frac{n}{4}(1 + \frac{\epsilon}{4\alpha})^{\alpha} f(L) \leq \epsilon \frac{n}{4}(1 + \frac{\epsilon}{3})(1 + \epsilon/3)A \\
&\leq \epsilon \frac{n}{4}(1 + \frac{\epsilon}{3})(1 + \epsilon/3)\frac{\texttt{AVG}}{(1 - \epsilon/4)} < \frac{16}{27}\epsilon G(p_{opt})
\end{aligned}$$

The second inequality comes from the sub-proportionality of f. For the remaining points, If $\|p_i - p_{opt}\| < L - \epsilon L/(4\alpha)$ but $\|p_i - p_{opt}\| \geq \|\tilde{p} - p_{opt}\|$, then

$$f(\|p_i - p'\|) \leq f(\|p_i - p_{opt}\| + \|p_{opt} - p'\|) \leq f(L) = (1 + \epsilon/3)A$$

With the fact that $f(\|p_i - p_{opt}\|) \geq A$, we have

$$f(\|p_i - p'\|) - f(\|p_i - p_{opt}\|) \leq \frac{\epsilon}{3}A \leq \frac{\epsilon}{3}f(\|p_i - p_{opt}\|)$$

If $\|p_i - p_{opt}\| \geq L - \epsilon L/(4\alpha)$, the cost from moving p_{opt} to p' is increased by a factor of at most $(1 + 11\epsilon/27)$:

$$\begin{aligned} f(\|p_i - p'\|) &\leq f(\|p_i - p_{opt}\| + \|p_{opt} - p'\|) \leq f(\|p_i - p_{opt}\| + \frac{\epsilon L}{4\alpha}) \\ &\leq (1 + \frac{\epsilon}{4\alpha - \epsilon})^{\alpha} f(\|p_i - p_{opt}\|) \leq e^{\epsilon/3} f(\|p_i - p_{opt}\|) \\ &\leq (1 + \frac{11}{27}\epsilon) f(\|p_i - p_{opt}\|) \end{aligned}$$

In sum, the total difference between $G(p') = \sum_i f(\|p_i - p'\|)$ and $G(p_{opt})$ is at most $\epsilon G(p_{opt})$, therefore p' is a $(1 + \epsilon)$-approximate solution of G.

For case (ii), for $i < \lfloor \epsilon n/4 \rfloor$, in other words, $\|p_i - p'\| \leq \|\tilde{p} - p'\|$, the total increase of cost by moving p_{opt} to p' is at most:

$$\sum_{i < \lceil \epsilon n/4 \rceil} f(\|p_i - p'\|) \leq \epsilon \frac{n}{4}(1 + \frac{\epsilon}{4\alpha})^{\alpha} f(L) \leq \epsilon \frac{n}{4}(1 + \frac{\epsilon}{3})\frac{\epsilon \texttt{AVG}}{3} < \frac{1}{9}\epsilon G(p_{opt})$$

When $\|\tilde{p} - p_{opt}\| \leq \|p_i - p_{opt}\| < L - \epsilon L/(4\alpha)$ we have:

$$f(\|p_i - p'\|) \leq f(\|p_i - p_{opt}\| + \|p_{opt} - p'\|) \leq f(L) = \epsilon \texttt{AVG}/3$$

Lastly, if $\|p_i = p_{opt}\| \geq L - \epsilon L/(4\alpha)$, the argument is the same as in case(i):

$$f(\|p_i - p'\|) \leq (1 + \frac{11}{27}\epsilon) f(\|p_i - p_{opt}\|)$$

In sum, the total difference between $G(p')$ and $G(p_{opt})$ is $< \epsilon G(p_{opt})$. So p' is a $(1 + \epsilon)$-approximate solution of G. □

The above lemma shows that if we choose L in this way, condition (b) of Step 1 is satisfied. Furthermore, the following lemma indicates that condition (a) can also be achieved.

Lemma 3. *Let ϵ,P,G,f, $\tilde{p}$, L be defined as in Lemma 2. By uniformly sampling $|X| = O(1/\epsilon)$ points in P, there will be a point $s \in X$ satisfying inequality $\|s - p_{opt}\| \leq L$ with constant probability.*

Proof. Since for both case(i) and case(ii) there are at least $\lfloor \epsilon n \rfloor$ points in P having $\|p_i - p_{opt}\| \leq \|\tilde{p} - p_{opt}\| \leq L$, after $2/\epsilon$ samples there will be at least one point falling in this set of points with probability $\geq 1/2$ by Markov's inequality. □

3.3 Main result

In this subsection, we omit the details of most of the proofs due to the space limit. First we present the theorem which guarantees the correctness of Step 2. For a set of points $X \subset \mathbb{R}^d$, we denote by $\text{span}(X)$ the affine subspace spanned by the set of points in X.

Theorem 2 (Core-set). *Let $0 < \epsilon < 1$. Let P be a point set in $\mathbb{R}^d$. $G(x) = \sum_{p \in P} f(\|x - p\|)$ with α as the proportion degree of f. L is chosen as in Lemma 2. If X is a set of points obtained from sampling $O(\log(1/\epsilon)/\epsilon^{3+\alpha})$ points in P, then with constant probability, the following two events happen: (i) The distance from the affine subspace $span(X)$ to the optimizer p_{opt} is at most $\epsilon L/(8\alpha)$, and (ii) X contains a point in distance $\leq L$ from p_{opt}.*

The above theorem gives the existence of a $(1+\epsilon)$-approximate solution in the affine subspace of a small sample. To actually find the solution is the final issue. We provide one of the possible approaches in the following.

The lemma below shows that we know a value $t = \Theta(\texttt{AVG})$. It also shows that trust the best source alone gives a constant approximate factor solution.

Lemma 4 (a $2^\alpha-$approximated solution). *Let P be a set of n points in $\mathbb{R}^d$ and $G(x) = \sum_{p \in P} f\|x - p\|$ with α as the proportion degree of f. We can try every point in P to achieve a 2^α-approximate solution for the function G, and the total running time is $O(n^2 d)$.*

Proof. Let $p' \in P$ be the one closest to the optimal point p_{opt}. Then

$$\begin{aligned} G(p') &= \sum_{p \in P} f(\|p - p'\|) \leq \sum_{p \in P} f(\|p - p_{opt}\| + \|p' - p_{opt}\|) \\ &\leq \sum_{p \in P} f(2\|p - p_{opt}\|) \leq 2^\alpha \cdot G(p_{opt}). \end{aligned}$$

The last inequality comes from the sub-proportionality of f. The minimum among $G(p_1), G(p_2), \cdots, G(p_n)$ must be less than $G(p')$. The function G can be evaluated in $O(nd)$ time. Therefore, the $2^\alpha-$approximate solution can be found in $O(n^2 d)$ time. □

There are more efficient ways to bound the value of AVG for special f. For example, when $f(x) = x$, it is shown [8] that AVG can be approximated in linear time.

Now we settle the unknown variables A and B in Lemma 2. We will show that if choosing B properly, A is approximately bounded in the way that $A = \Theta_\epsilon(\texttt{AVG})$. Thus the search of the value A takes at most $\text{poly}(1/\epsilon)$ time. The effect on the whole algorithm is a multiplicative factor of $\text{poly}(1/\epsilon)$, which is small comparing to the time for drawing grid points.

Lemma 5. *Let ϵ,P,G,f be defined as in Lemma 2. Let $\tilde{p}$ be the $\lceil \epsilon n \rceil$-th closest point to the optimal solution p_{opt} among the points in P. There exists a set $\mathcal{L}$ of size $O(\log(1/\epsilon)/\epsilon)$ such that for every possible values of $f(\|\tilde{p} - p_{opt}\|)$, there is a member $L \in \mathcal{L}$ such that it satisfies condition (a) and (b) in Step 1.*

The next theorem summarizes the complete algorithm.

Theorem 3. *Let $0 < \epsilon \leq 1$. Let P be a set of n points in $\mathbb{R}^d$ and $G(x) = \sum_{p \in P} f(\|x - p\|)$ with α as the proportion degree of f. Let X be a set of random samples from P of size $O(\log(1/\epsilon)/\epsilon^{3+\alpha})$. We can construct a set of grid points Y of size $O(2^{(1/\epsilon)^{O(1)}})$ such that with constant probability there is at least one point p' in Y being a $(1+\epsilon)$-approximate solution of G. The time complexity is $O(2^{(1/\epsilon)^{O(1)}} d + n^2 d)$ for the construction of Y.*

Synopsis of the proof: For each $L \in \mathcal{L}$, denote Y_L as the union of the grid points around each $x \in X$, where the diameter of the grid is $4L$ and the side length is roughly $O(\epsilon L)$. Let $Y = \cup_{L \in \mathcal{L}} Y_L$. Then Theorem 1 guarantees the desired result.

References

1. M. Bādoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 250–257. ACM, 2002.
2. H. Ding, J. Gao, and J. Xu. Finding global optimum for truth discovery: Entropy based geometric variance. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 51. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
3. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.
4. X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.
5. X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *Proceedings of the VLDB Endowment*, 2(1):562–573, 2009.
6. D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.
7. A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 131–140. ACM, 2010.
8. A. Kumar, Y. Sabharwal, and S. Sen. Linear time algorithms for clustering problems in any dimensions. In *International Colloquium on Automata, Languages, and Programming*, pages 1374–1385. Springer, 2005.
9. F. Li, M. L. Lee, and W. Hsu. Entity profiling with varying source reliabilities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1146–1155. ACM, 2014.
10. Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM, 2014.

11. J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885. Association for Computational Linguistics, 2010.
12. P. Welinder, S. Branson, S. J. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, volume 23, pages 2424–2432, 2010.
13. J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.
14. H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1935–1944. ACM, 2016.
15. H. Xiao, J. Gao, Z. Wang, S. Wang, L. Su, and H. Liu. A truth discovery approach with theoretical guarantee. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1925–1934. ACM, 2016.
16. X. Yin, J. Han, and S. Y. Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.
17. B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 2012.
18. B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.
19. H. D. Ziyun Huang and J. Xu. Faster algorithm for truth discovery via range cover. In *Proceedings of Algorithms and Data Structures Symposium (WADS 2017)*, pages 461– 472, 2017.