A De-singularity Subgradient Approach for the Extended Weber Location Problem

Zhao-Rong Lai\textsuperscript{1}, Xiaotian Wu\textsuperscript{2}, Liangda Fang\textsuperscript{2} and Ziliang Chen\textsuperscript{3,4}

\textsuperscript{1}Department of Mathematics, College of Information Science and Technology, Jinan University
\textsuperscript{2}Department of Computer Science, College of Information Science and Technology, Jinan University
\textsuperscript{3}Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory
\textsuperscript{4}Guangdong Institute of Smart Education, Jinan University

\{laizhr, wxiaotian, fangld\}@jnu.edu.cn, c.ziliang@yahoo.com

Abstract

The extended Weber location problem is a classical optimization problem that has inspired some new works in several machine learning scenarios recently. However, most existing algorithms may get stuck due to the singularity at the data points when the power of the cost function $1 \leq q < 2$, such as the widely-used iterative Weiszfeld approach. In this paper, we establish a de-singularity subgradient approach for this problem. We also provide a complete proof of convergence which has fixed some incomplete statements of the proofs for some previous Weiszfeld algorithms. Moreover, we deduce a new theoretical result of superlinear convergence for the iteration sequence in a special case where the minimum point is a singular point. We conduct extensive experiments in a real-world machine learning scenario to show that the proposed approach solves the singularity problem, produces the same results as in the non-singularity cases, and shows a reasonable rate of linear convergence. The results also indicate that the $q$-th power case ($1 < q < 2$) is more advantageous than the 1-st power case and the 2-nd power case in some situations. Hence the de-singularity subgradient approach is beneficial to advancing both theory and practice for the extended Weber location problem.

1 Introduction

The extended Weber location problem is a classical optimization problem [Ostresh, 1978; Brimberg and Love, 1993; Beck and Sabach, 2015] that has been introduced to some new machine learning scenarios recently [Aftab et al., 2015; Li et al., 2016; Lai et al., 2018c; Lai et al., 2020; Lai and Yang, 2023]. It finds the point that minimizes the $q$-th power of the Euclidean distances from $m$ fixed data points $x_1, \ldots, x_m \in \mathbb{R}^d$ [Weber, 1909; Cooper, 1968; Chen, 1984]:

$$x_* = \arg \min_y C_q(y) = \arg \min_y \sum_{i=1}^m \|y - x_i\|^q,$$

(1)

where $\| \cdot \|^q$ denotes the $q$-th power of the Euclidean distance, and $C_q(y)$ is the $q$-th power cost function at the point $y \in \mathbb{R}^d$. The $q$-th power median $x_*$ (especially when $1 \leq q < 2$) has recently been found useful in rotation averaging and gives more robust results when outliers exist [Aftab et al., 2015]. We will further show its advantage in online portfolio selection [Li et al., 2016; Lai and Yang, 2023] in this paper.

1.1 The Singularity Problem

In order to have a quick insight into the aim of this paper, we first compute the gradient of $C_q(y)$ w.r.t. $y$:

$$\nabla C_q(y) = \sum_{i=1}^m q\|y - x_i\|^{q-2}(y - x_i).$$

(2)

It is well-defined when $q \geq 2$ and one could easily find some gradient descent methods to work out the minimum [Afsari et al., 2013], which is beyond the main concern of this paper. However, it is singular at the data points $\{x_i\}_{i=1}^m$ when $1 \leq q < 2$, since at least one of the weights $\|y - x_i\|^{q-2}$ becomes undefined (see Figure 1). This paper mainly addresses this case and we assume $1 \leq q < 2$ in the rest of this paper if not specified.

Since $C_q(y)$ is convex and continuous in $\mathbb{R}^d$ including these data points $\{x_i\}_{i=1}^m$, we can still turn to the subgradient approach for solutions.

Definition 1 ([Rockafellar and Wets, 2009]). The Fréchet subdifferential of $C_q$ at $y$, denoted by $\partial C_q(y)$, is the set of all vectors $v \in \mathbb{R}^d$ satisfying

$$\partial C_q(y) \triangleq \left\{ v \in \mathbb{R}^d : \liminf_{z \to y} \frac{C_q(z) - C_q(y) - v^\top(z - y)}{\|z - y\|_2} \geq 0 \right\}.$$

(3)

Since $C_q$ is also proper (dom $(C_q) \neq \emptyset$ and $C_q(y) \neq -\infty$, $\forall y \in \mathbb{R}^d$), its Fréchet subdifferential is equivalent to the classical subdifferential in the literature of convex analysis.
Weiszfeld algorithm (5) is actually a gradient descent method (6), which shall fail if \( y(p) \) hits one of \( \{x_i\}_{i=1}^m \) and \( 1 \leq q < 2 \).

1.2 The Significance and Difficulty of This Singularity Problem

This singularity problem actually happens frequently and unexpectedly. We can easily give a simple example. Let \( q = 1.1 \) and 6 data points be

\[ \{x_i\}_{i=1}^6 \triangleq \{(-2, 0), (-1, 0), (1, 0), (2, 0), (0, 1), (0, -1)\} \]

The starting point \( y(0) \triangleq (1.68645, 0) \) is chosen distinct from \( \{x_i\}_{i=1}^6 \). The existing algorithm \( q \text{PWA} \) gets stuck with just 1 iteration since it hits one data point \( x_3 = (1, 0) \) and cannot proceed. The resulted cost function is \( C_q = 9.4201 \), which is not the true minimum. On the contrary, the proposed algorithm \( q \text{PWAWS} \) successfully escapes from the singular point \( x_3 \) and finds the true solution \( (0, 0) \) with the true minimum \( C_q = 8.2871 \). [Chandrasekaran and Tamir, 1989; Vardi and Zhang, 2000] further point out that the bad starting point set \( \{y(0)\} \) itself may constitute a continuum set that can be dense in an open region of \( \mathbb{R}^d \) with \( d \geq 2 \), even the entire \( \mathbb{R}^d \). Hence this singularity problem is quite significant, if not serious.

<table>
<thead>
<tr>
<th>( q )</th>
<th>1.1</th>
<th>Iter</th>
<th>( x )</th>
<th>( C_q(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q \text{PWA} )</td>
<td>1 (get stuck)</td>
<td>(1, 0)</td>
<td>9.4201</td>
<td></td>
</tr>
<tr>
<td>( q \text{PWAWS} ) (ours)</td>
<td>24</td>
<td>(0, 0)</td>
<td>8.2871</td>
<td></td>
</tr>
</tbody>
</table>

This singularity problem cannot be radically eliminated by straightforward treatments, such as perturbations and random restarts. A treatment is to find a starting point with a smaller cost than the costs of any data points \( \{x_i\}_{i=1}^m \) by computing \( \{C_q(x_i)\}_{i=1}^m \) at first [Aftab et al., 2015], but it is rather exhaustive especially when the number of data points \( m \) is large. It requires at least \( m \) iterates to find a starting point, but sometimes we only need fewer iterates to find a minimum point.

Table 3 shows that our method requires only 9.31 iterates in average to find the minimum in the “NYSE(N), \( m = 10 \), \( q = 1.9 \)” case, while the above treatment conducts \( m = 10 \) iterates only to find a starting point. Furthermore, if \( x_k \) is exactly the minimum, we have to check whether \( 0 \in \partial C_q(x_k) \). After all, once we figure out \( \partial C_q(x_k) \), we can get rid of these troubles. In fact, there is no need to introduce additional treatments because our de-singularity approach can immediately replace the ordinary gradient step without increasing computational complexity, which is a great advantage.

In the \( q = 1 \) case, there is a remedy that removes the singular term in the gradient [Kuhn, 1973; Vardi and Zhang, 2000], but it does not investigate the subgradient \( \partial C_q(x_k) \). More importantly, there are some incomplete statements in the proof of convergence, which will be fixed in Section 3.4. In the \( 1 < q \leq 2 \) case, neither has the subgradient \( \partial C_q(x_k) \) been figured out nor has the proof of convergence been completed, yet.

1.3 Our Results

We mainly establish a complete “de-singularity” subgradient approach for the extended Weber location problem (1) \( 1 \leq q < 2 \). The key contributions fall into four aspects:

1. By removing the singular term, we propose a de-singularity subgradient of \( C_q \) at each singular point \( x_k \in \{C_q(x_i)\}_{i=1}^m \). It can replace the ordinary gradient without increasing computational complexity.

2. If \( x_k \) is not a minimum point, we pull the stuck iterate in the de-singularity subgradient descent direction that can reduce the cost \( C_q \), so that the iterates can monotonically converge to the exact minimum. In the rest of this pa-
per, we call the new algorithm q-th Power Weiszfeld Algorithm without Singularity (qPWAWS).

3. We present a complete proof of convergence for qPWAWS, which has fixed some incomplete statements of the proofs for some previous Weiszfeld algorithms.

4. If one of \( \{x_i\}_{i=1}^m \) is the minimum, we prove that qPWAWS enjoys a superlinear convergence for the iteration sequence when \( 1 < q < 2 \), which is a new theoretical result. If none of \( \{x_i\}_{i=1}^m \) is the minimum, we show that qPWAWS enjoys a reasonable rate of linear convergence for the iteration sequence with computational experiments.

The proposed algorithm qPWAWS can use any starting point in \( \mathbb{R}^d \) and guarantee a monotonic convergence to the exact minimum. Moreover, this de-singularity subgradient approach can also be adopted in other gradient-algorithm problems for Problem (1).

2 Related Works on the Weiszfeld Algorithms

In this section, we assume that the data points \( \{x_i\}_{i=1}^m \) are distinct and non-collinear (i.e., they do not lie in a hyper-plane). Thus the cost function \( C_q(y) \) in (1) is strictly convex and has a unique minimum. We review some related works on how to derive the Weiszfeld algorithms.

2.1 Convergence in the Non-singularity Case

For the basic qPW (5), it has been proven that \( C_q(y_{(p+1)}) < C_q(y_{(p)}) \) if \( y_{(p+1)} \neq y_{(p)} \), and that as long as the sequence of iterates does not hit any data points, it monotonically converges to the exact minimum [Kuhn, 1973; Cooper, 1968; Chen, 1984; Aftab et al., 2015], while some incomplete statements are to be fixed in Section 3.4. The rate of convergence for qPWA (1 < q < 2) has not been deduced, yet.

2.2 Iterative Re-weighted Least Squares (IRLS) Interpretation

Based on the current iterate \( y_{(p)} \), a weighted 2-norm power cost function \( \tilde{C}_q(y) \) can be created to approximate \( C_q(y) \):

\[
\tilde{C}_q(y) = \sum_{i=1}^m \|y_{(p)} - x_i\|^{-q/2}\|y - x_i\|^2.
\]

By setting the gradient of \( \tilde{C}_q(y) \) as zero, we have the same update formula as (5). It is a kind of Iterative Re-weighted Least Squares (IRLS) techniques [Chartrand and Yin, 2008; Daubechies et al., 2010; Eldar and Mishali, 2009; Aftab et al., 2015].

2.3 1-st Power Weiszfeld Algorithm without Singularity

A remedy for the L1-median is studied in [Kuhn, 1973; Vardi and Zhang, 2000; Ostresh, 1978]. First, we define a de-singularity Weiszfeld transform that excludes the singular point as follows:

\[
\tilde{T}(y) = \frac{\sum_{x_i \neq y} \|y - x_i\|^{-1} x_i}{\sum_{x_i \neq y} \|y - x_i\|^{-1}}.
\]

Then the next iterate is the combination of \( \tilde{T}(y_{(p)}) \) and the current iterate \( y_{(p)} \):

\[
y_{(p+1)} = (1 - \lambda)\tilde{T}(y_{(p)}) + \lambda y_{(p)},
\]

where \( 0 \leq \lambda \leq 1 \) is a mixing parameter.

In order to drag the iterate out of the singular point and reduce the cost simultaneously, we adopt the 1-st power de-singularity subgradient (a formal definition will be given in Definition 5) and set \( \lambda \) as follows:

\[
\lambda = \begin{cases} 0 & \text{if } y_{(p)} \notin \{x_i\}_{i=1}^m \\
\min \left\{ 1, \frac{1}{\|\nabla D_1(y_{(p)})\|} \right\} & \text{if } y_{(p)} \in \{x_i\}_{i=1}^m
\end{cases}
\]

Compared with (2), \( \nabla D_1(y_{(p)}) \) removes the singular term in \( \nabla C_1(y_{(p)}) \). If \( y_{(p)} \) does not hit any data point, then (9) is just the same as (5). Else, \( \lambda \) ensures \( C_1(y_{(p+1)}) \leq C_1(y_{(p)}) \) and the iterates monotonically converge to the exact minimum, which are explained by [Vardi and Zhang, 2000].

3 q-th Power Weiszfeld Algorithm without Singularity

Although there is a remedy for the 1-st power case, it is non-trivial to deal with the general q-th power (1 < q < 2) case. We will see that the 1-st power case and the q-th power (1 < q < 2) case are different in the characterization of minimum, the minimizing strategy, and the rate of convergence, since they have different smoothness. Nevertheless, we still present a unified q-th power (1 < q < 2) framework for the integrity of the approach.

In the rest of this paper, we generalize the q-th power (1 < q < 2) cost function in (1) to:

\[
C_q(y) = \sum_{i=1}^m \xi_i \|y - x_i\|^q,
\]

where \( \xi_i > 0 \) is the multiplicity of the data point \( x_i \). It can deal with duplication and collinearity of data points. For example, if \( x_1 = x_2 = x_3 \) and the other points are distinct, then we can merge \( x_1, x_2, x_3 \) and set \( \xi_1 = 3, \xi_i = 1 (i \geq 4) \). The collinear data points can also be merged to be non-collinear ones by the same way. Hence, we assume that the data points are distinct and non-collinear in the rest of this paper. Thus \( C_q(y) \) is strictly convex on \( y \) and there is a unique minimum point \( M \). To be convenient for illustrations, we set \( \eta_i \triangleq \xi_i^{1/2} \) and further change (12) to:

\[
C_q(y) = \sum_{i=1}^m \eta_i \|y - x_i\|^q.
\]

The illustration of qPWAWS consists of 5 steps:

1. The general update formula is introduced as an easy beginning.
2. The q-th power de-singularity subgradient is defined to characterize the subgradients and the minimum.
3. The $q$-th power de-singularity subgradient is adopted to get out of the singular point and reduce the cost simultaneously.

4. The proof of convergence is conducted.

5. The theoretical rate of convergence for the iteration sequence in a special case is deduced with the properties of the $q$-th power de-singularity subgradient.

### 3.1 General Update Formula without Coincidence

First, we consider the simplest case where the current iterate $y(p)$ does not coincide with the data points $\{x_i\}_{i=1}^m$. We first state the general iterative update formula in this case:

$$y(p+1) = T_1(y(p)) = \frac{\sum_{i=1}^m \eta_i \|y(p) - x_i\|^{q-2} x_i}{\sum_{i=1}^m \eta_i \|y(p) - x_i\|^q}. \quad (14)$$

As a more general case than [Kuhn, 1973; Cooper, 1968; Chen, 1984; Aftab et al., 2015], we present the following non-increasing theorem:

**Theorem 3.** If $y(p) \not\in \{x_i\}_{i=1}^m$, then $C_q(T_1(y(p))) \leq C_q(y(p))$ with equality only when $T_1(y(p)) = y(p)$.

The proof is provided in Supplementary B.1.

### 3.2 Characterization of Subgradients and Minimum

Second, we establish the de-singularity subgradient and characterize the minimum point $M$ of $C_q(y)$ in (13). The following corollary is a direct result of Theorem 3.

**Corollary 4.** If $y(p) \not\in \{x_i\}_{i=1}^m$, then $T_1(y(p)) = y(p) \iff y(p) = \min\{\partial C_q(y)\}$.

The proof is provided in Supplementary B.2. If $y(p) \in \{x_i\}_{i=1}^m$, then the minimum characterization relies on the subgradient(s) in $\partial C_q(y(p))$. Without loss of generality, suppose $y(p) = x_k$.

**Definition 5 (q-th Power De-singularity Subgradient).** Let $D_q(y)$ be the main component of $C_q(y)$ that excludes the term $\eta_k\|y - x_k\|^q$. Then the $q$-th power de-singularity subgradient $\nabla D_q(y)$ of $C_q(y)$ is:

$$D_q(y) = \sum_{i \neq k} \eta_i \|y - x_i\|^q, \quad (15)$$

$$\nabla D_q(y) = \sum_{i \neq k} q \eta_i \|y - x_i\|^{q-2}(y - x_i), \quad 1 \leq q < 2. \quad (16)$$

Compared with the ordinary gradient (2), the de-singularity subgradient (16) does not require more computation but checking whether $y = x_k$ for the $k$-th summand, which can be done at the same time and in the same loop with summation computing. Hence the de-singularity subgradient will not increase computational complexity.

**Theorem 6 (Characterization of Subgradients and Minimum).** Let $y(p) = x_k$. Then

$$\partial C_q(x_k) = \begin{cases} \{\nabla D_1(x_k) + \eta_k u : \|u\| \leq 1\} & \text{if } q = 1 \\ \{\nabla D_q(x_k)\} & \text{if } 1 < q < 2. \end{cases} \quad (17)$$

According to Fermat’s rule, $x_k$ is the minimum point $M$ if and only if $0 \in \partial C_q(x_k)$.

The proof is provided in Supplementary B.3. It is obvious that $\nabla D_q(x_k) \in \partial C_q(x_k)$ for all $1 \leq q < 2$. This is why $\nabla D_q(x_k)$ is termed a de-singularity subgradient: it removes the singularity and then becomes a subgradient. One can also see that (10) is a special case of (17) with $y(p) = x_k$ and $u = 0$.

$-\nabla D_q(x_k)$ can be interpreted as the resultant implemented on $x_k$ towards other data points (see Figure 2). When $q = 1$ and $\|\nabla D_1(x_k)\|$ is smaller than the intrinsic force $\eta_k$ implemented on $x_k$, the system remains still and $C_q(x_k)$ has reached its minimum. But when $1 < q < 2$, $\|\nabla D_q(x_k)\|$ should be zero to keep the system still, no matter how large the intrinsic force $\eta_k$ is.

### 3.3 Update Formula with Coincidence

The update formula with coincidence $y(p) = x_k$ should be set up separately for $q = 1$ and $1 < q < 2$. The $q = 1$ case is provided in Supplementary A.1. Now we turn to solve the $1 < q < 2$ case. The following theorem shows a way to reduce the $q$-th power cost when the iterate gets stuck in the singular point.

**Theorem 7 (De-singularity Subgradient Descent Method).** If $y(p) = x_k$ and $\|\nabla D_q(x_k)\| > 0$, then there exists a $\lambda_* > 0$ such that for any $0 < \lambda \leq \lambda_*$, $C_q(x_k - \lambda \nabla D_q(x_k)) < C_q(x_k), 1 < q < 2$.

The proof is provided in Supplementary B.4. The key point of Theorem 7 is that the negative de-singularity subgradient part (resultant) $-\lambda \|\nabla D_q(x_k)\|^2$ is of lower order infinitesimal than the positive singular part (resistance) $\eta_k^{\frac{q}{q-1}} \|\nabla D_q(x_k)\|^q$. Thus when $\lambda$ is sufficiently small, the resultant overcomes the resistance and drags the current iterate $y(p)$ out of the singular point and the $q$-th power cost is reduced simultaneously. An algorithm can be designed based on this mechanism. By omitting $\frac{q}{2} G(x_k) + o(\lambda)$ on the left side of (43), we can approximately start with the following $\lambda_0$:

$$\lambda_0 = \min \left\{ \frac{1}{\eta_k} \left( \frac{2-q}{q-1} \right)^{\frac{q}{q-1}} \right\}. \quad (18)$$
As long as \( C_q(x_k - \lambda u \nabla D_q(x_k)) \geq C_q(x_k) \), we reduce \( \lambda u \) with a factor \( \rho < 1: \lambda_{w+1} \leftarrow \rho \lambda_w, w = 0, 1, \ldots \), until we find some \( \lambda_* \) such that \( C_q(x_k - \lambda_* \nabla D_q(x_k)) < C_q(x_k) \) (see Figure 3). According to Theorem 7, this \( \lambda_* \) is sure to be found. Then the next iterate is:

\[
y_{(p+1)} = T_2(x_k) \triangleq x_k - \lambda_* \nabla D_q(x_k). \tag{19}
\]

This strategy is very efficient to get away from the singular point: it takes only 2.75 iterates in average in the worst case of our experiments (see Table 1). It can be even more efficient if we further reduce \( \lambda_0 \) and \( \rho \). Besides, the choice of \( \lambda_* \) can also be absorbed in momentum acceleration methods, such as Nesterov [Nesterov, 1983; Sutskever et al., 2013] and Adam [Kingma and Ba, 2015]. Details can be found in the code link provided in the section of acknowledgments.

3.4 A Complete Proof of Convergence

From (14) and (19), the \( q \)-th Power Weiszfeld Algorithm without Singularity (qPWAWS) can be designed as (1 < \( q \) < 2):

\[
y_{(p+1)} = T(y_{(p)}) = \left\{ \begin{array}{ll}
T_1(y_{(p)}) = \frac{\sum_{k=1}^{m} \mu_k\|y_{(p)} - x_k\|^{q-2}x_k}{\sum_{k=1}^{m} \mu_k\|y_{(p)} - x_k\|^{q-1}} & \text{if } y_{(p)} \notin \{x_k\}_{k=1}^{m}, \vspace{0.2cm} \\
T_2(y_{(p)}) = x_k - \lambda_* \nabla D_q(x_k) & \text{if } y_{(p)} = x_k
\end{array} \right.
\]

where \( T_1(y_{(p)}) \) is the update formula without coincidence (14) and \( T_2(y_{(p)}) \) is the update formula with coincidence (19). Note that the \( q = 1 \) algorithm without singularity has been proposed by [Kuhn, 1973; Vardi and Zhang, 2000], but their convergence proofs are incomplete. Thus we focus on the proof of the 1 < \( q \) < 2 case as well as fix their incomplete statements in this subsection. In calculation, the \( q = 1 \) case and the 1 < \( q \) < 2 case can be unified.

Starting with any initial point \( y_{(0)} \), the qPWAWS generates a sequence of iterates \( y_{(0)}, y_{(1)}, \ldots, y_{(p)}, \ldots \). It is necessary to prove that the sequence converges to the minimum point \( M \) unless some \( y_{(p)} \) hits \( M \).

**Lemma 8.** The sequence \( \{y_{(p)}\} \) generated by qPWAWS (20) visits each \( x_k \neq M \) at most once and will not get stuck. Except for at most a finite set of iterates, \( \{y_{(p)}\} \) and \( M \) lie in the convex hull of the data points \( \{x_i\}_{i=1}^{m} \).

The proof is provided in Supplementary B.5. It is also possible that the sequence \( \{y_{(p)}\} \) converges to the data points \( \{x_i\}_{i=1}^{m} \). In fact, \( T_1(y) \) is continuous in almost the whole \( \mathbb{R}^d \) except for the data points \( \{x_i\}_{i=1}^{m} \) where it is discontinuous. However, \( \{x_i\}_{i=1}^{m} \) are removable discontinuities. Less general versions of this characteristic are mentioned in [Kuhn, 1973; Aftab et al., 2015] but without proof. Thus we present the following lemma:

**Lemma 9.**

\[
\lim_{y \to x_k} T_1(y) = y_k, \quad \forall 1 \leq k \leq m, \quad \forall 1 \leq q < 2. \tag{21}
\]

The proof is provided in Supplementary B.6. Based on this lemma, the operator \( T_1 \) can be extended to the removable discontinuous data points:

\[
T_1(x_k) \triangleq x_k, \quad \forall 1 \leq k \leq m, \quad \forall 1 \leq q < 2. \tag{22}
\]

From (22) and Corollary 4, all the fixed points of \( T_1 \) are characterized as:

\[
T_1(y) = y \iff y \in \{x_i\}_{i=1}^{m} \cup \{M\}. \tag{23}
\]

When \( y \) gets into some neighborhood of a data point \( x_k \neq M \), the operator \( T_1 \) will eventually drive it out of the neighborhood. This property of \( T_1 \) helps to find the real minimum point. The following lemma is a nontrivial extension of a similar conclusion for the 1-st power case in [Kuhn, 1973].

**Lemma 10.** If \( x_k \neq M \), then there exists some \( \delta_0 > 0 \) such that for all \( y \in B(x_k, \delta_0) \), \( T_1^{-1}(y) \in B(x_k, \delta_0) \) and \( T_1(y) \notin B(x_k, \delta_0) \) for some \( s. \) \( B(x_k, \delta_0) \) denotes an open \( \delta_0 \)-ball centered at \( x_k \). \( T_1 \) means implementing \( T_1 \) for \( s \) times, and \( s \) depends on \( y \).

The proof is provided in Supplementary B.7.

**Theorem 11 (Convergence Theorem).** Starting from any initial point \( y_{(0)} \), if the sequence \( \{y_{(p)}\} \) generated by qPWAWS (20) does not hit \( M \), then \( y_{(p)} \to M \). If \( y_{(p)} \) hits \( M \), the characterization of minimum (Theorem 6) ensures that this hit could be recognized and the algorithm would be stopped.

The proof is provided in Supplementary B.8.

**Some notes: 1.** In [Kuhn, 1973; Vardi and Zhang, 2000], the proof of convergence assumes that \( \lim_{p \to \infty} y_{(p)} = x_k \) but \( T_1(y_{(p)}) \notin B(x_k, \delta_0) \) for all \( p \). However, Lemma 10 indicates that \( T_1 \) may be implemented for \( s > 1 \) times to drive \( y_{(p)} \) out of \( B(x_k, \delta_0) \). Besides, this \( s \) also depends on \( y_{(p)} \) and the uniform upper bound of \( s \) for all the \( y_{(p)} \) may not exist. Thus it may not be concluded that \( \lim_{p \to \infty} \|T_1(y_{(p)}) - x_k\| = \infty \) and it may fail the conclusions in [Kuhn, 1973; Vardi and Zhang, 2000]. Moreover, [Kuhn, 1973; Vardi and Zhang, 2000] have only shown that a subsequence \( \{y_{(p_n)}\} \) converges to \( M \), but not the whole sequence \( \{y_{(p)}\} \). The convergence of the whole sequence \( \{y_{(p)}\} \) should depend on the uniqueness of \( M \).

**2.** In [Aftab et al., 2015], the proof of convergence does not include Lemma 10. Thus it may not guarantee that a subsequence \( y_{(p_n)} \in B(x_k, \delta_0) \) and \( T_1(y_{(p_n)}) \notin B(x_k, \delta_0) \). In this case, (57), (58), (59) and (60) may not be deduced and the contradiction may not exist.

Therefore, the proof of convergence in this paper has also fixed some incomplete statements of some related works before.
3.5 Rate of Convergence

It is also important to analyze the rate of convergence for \(q\)PWAWS when \(y(p) \to M\). With the proposed desingularization subgradient \(\nabla D_q\) in (16), we can deduce the exact rate of convergence for \(q\)PWAWS in the special case \(M \in \{x_i\}_{i=1}^m\). To the best of our knowledge, this is a new theoretical result that \(q\)PWAWS enjoys a superlinear convergence for the iteration sequence in this case. Without loss of generality, we assume \(M = x_k\). Since the subgradients are different between the \(q = 1\) case and the \(1 < q < 2\) case (Theorem 6), we further divide the analysis into two parts. The \(q = 1\) case has been deduced by [Ostresh, 1978], while the \(1 < q < 2\) case is somewhat complicated. The key technique for the \(1 < q < 2\) case is to check the order of infinitesimal of \(\|\nabla D_q(y(p))\|\) with \(y(p) \to x_k\) at first.

**Lemma 12.** If \(1 < q < 2\), the order of infinitesimal of \(\|\nabla D_q(y(p))\|\) is no higher than that of \(\|y(p) - x_k\|\) with \(y(p) \to x_k\). Precisely, there exists some \(\zeta > 0\) such that
\[
\lim_{y(p) \to x_k} \frac{\|\nabla D_q(y(p))\|}{\|y(p) - x_k\|} \leq \zeta, \quad 1 < q < 2. \tag{24}
\]

The proof is provided in Supplementary B.9.

**Theorem 13.** If \(M = x_k\) for some \(k\), the rate of convergence for \(q\)PWAWS is:
\[
\lim_{y(p) \to x_k} \frac{\|y(p+1) - x_k\|}{\|y(p) - x_k\|} = \begin{cases} \frac{\|\nabla D_1(x_k)\|}{\eta_q} & q = 1, \\ 0 & 1 < q < 2. \end{cases} \tag{25}
\]

It is a superlinear convergence when \(1 < q < 2\) and no worse than a sublinear convergence when \(q = 1\).

The proof is provided in Supplementary B.10. As for the general case \(M \notin \{x_i\}_{i=1}^m\), the rate of convergence is unknown at present according to our knowledge. It is difficult to eliminate the infinitesimal \(\|y(p) - M\|\) and obtain a uniform upper bound of \(\lim_{y(p) \to M} \frac{\|y(p+1) - M\|}{\|y(p) - M\|}\) that is no greater than 1. When \(q = 1\), [Ostresh, 1978] shows some computational evidence that the rate of convergence for \(M \notin \{x_i\}_{i=1}^m\) is usually somewhat less than that of \(M \in \{x_i\}_{i=1}^m\). We will give some computational results for the \(1 < q < 2\) case in Section 4.3 to show that \(q\)PWAWS enjoys a reasonable rate of linear convergence. We summarize the whole \(q\)PWAWS in Supplementary A.2, which gives a complete procedure to deal with different situations that may occur in real applications.

4 Experimental Results

Since 1PWA and qPWA have already been verified to be effective in many applications [Hartley et al., 2011; Hartley et al., 2013; Fletcher et al., 2009; Yang, 2010; Huang et al., 2016], we do not intend to repeat their experiments. Instead, we focus on validating that \(q\)PWAWS successfully solves the singular problem and enjoys a reasonable rate of linear convergence.

We conduct experiments on an interesting machine learning application: online portfolio selection (OPS, [Li et al., 2016; Lai et al., 2018c; Lai et al., 2020; Lai and Yang, 2023]). By using the notations in this paper, a data point \(x_i \in \mathbb{R}^d\) indicates a price vector of \(d\) assets on the \(i\)-th day. \(\{x_i\}_{i=1}^m\) contains the asset prices on the most recent \(m\) days. Following [Huang et al., 2016], we fed the \(q\)-th power median
\[
\hat{x} = \arg\min_y \sum_{i=1}^m \|y - x_i\|^q, \quad 1 \leq q \leq 2 \tag{26}
\]
to a portfolio optimization model [Huang et al., 2016] to decide the future portfolio. We adopt the NYSE(N) data set [Li et al., 2013] and propose a new CSI300 data set.

- **NYSE(N):** it contains daily price relative sequences of \(d = 23\) stocks from New York Stock Exchange during 1/Jan/1985~30/Jun/2010 (\(T = 6431\) days).
- **CSI300:** it contains daily price relative sequences of \(d = 47\) stocks from the CSI300 constituents of Shanghai Stock Exchange and Shenzhen Stock Exchange in China during 16/Mar/2015~19/May/2017 (\(T = 534\) days).

Experiments consist of four parts: first, we verify that \(q\)PWAWS solves the singularity problem and obtains the same results as qPWA. Second, we measure the computational cost and the number of iterations for convergence of \(q\)PWAWS. Third, we give computational results on the rate of convergence for \(q\)PWAWS. Fourth, we show that the \(q\)-th power median \((1 < q < 2)\) is more effective and robust than the 1-st power median and the 2-nd power median in some cases, thus the \(q\)-th power median is useful and it is important to solve the singularity problem. As for the parameters, if not specified, the observation window size is \(m = 5\), which is consistent with previous related methods [Huang et al., 2016; Lai et al., 2018a; Lai et al., 2018b; Lai et al., 2022]. The tolerance threshold is Tol = 10\(^{-9}\). The reducing factor is \(p = 0.1\), which is a moderate value. As the observation window moves from \(t = 1\) to \(t = T - m + 1\), there are a total number of \(T - m + 1\) sets of data points \(\{x_i\}_{i=1}^m\). Thus the \(q\)-th power median computation can be conducted for \((T - m + 1)\) times to evaluate the average performance of \(q\)PWAWS.

4.1 Addressing the Singularity Problem

If the sequence of iterates does not hit the data points, then \(q\)PWAWS is equivalent to qPWA. Without loss of generality, we examine a specific case: let the starting point \(y(0) = x_1\) for \(q\)PWAWS and \(y(0) = \frac{1}{m} \sum_{i=1}^m x_i\) for qPWA. Besides, we also watch to avoid getting stuck for qPWA, since it cannot deal with singularity.

To evaluate the efficiency of \(q\)PWAWS, we measure how many iterates that \(q\)PWAWS takes to successfully get away from the singular point with a reduced cost (implementing only Step 11 in Algorithm 1). For each data set, we use two different sizes of observation windows \(m = 5\) and \(m = 10\), and let \(q = 1.1 \sim 1.9\). The mean and the standard deviation (STD) of the number of iterates are shown in Table 1. \(q\)PWAWS successfully gets away from the singular points in all the experiments. As \(q\) increases from 1.1 to 1.9, the average number of iterates decreases from about 2.54 to 2 for \(m = 5\) and from about 2.75 to 2 for \(m = 10\) on CSI300. In

\(^1\)http://www.csindex.com.cn
fact, a smaller $\rho$ and a smaller $\lambda_n$ in (18) can further reduce the number of iterates and Step 11 only need to compute the
$q$-th power cost. In general, the number of iterates to get out of the singular point is small.

\[
q \quad \text{NYSE(N)} \quad \text{CSI300}
\]

\[
\begin{array}{cccc}
1.1 & 2.01 \pm 0.49 & 2.19 \pm 0.46 & 2.54 \pm 0.50 & 2.75 \pm 0.43 \\
1.2 & 2.00 \pm 0.41 & 2.17 \pm 0.41 & 2.38 \pm 0.49 & 2.71 \pm 0.46 \\
1.3 & 1.99 \pm 0.31 & 2.15 \pm 0.37 & 2.19 \pm 0.39 & 2.64 \pm 0.48 \\
1.4 & 1.99 \pm 0.22 & 2.12 \pm 0.32 & 2.03 \pm 0.17 & 2.52 \pm 0.50 \\
1.5 & 1.99 \pm 0.10 & 2.07 \pm 0.25 & 2.00 \pm 0.00 & 2.28 \pm 0.45 \\
1.6 & 2.00 \pm 0.04 & 2.03 \pm 0.18 & 2.00 \pm 0.00 & 2.06 \pm 0.23 \\
1.7 & 2.00 \pm 0.00 & 2.00 \pm 0.05 & 2.00 \pm 0.00 & 2.00 \pm 0.00 \\
1.8 & 2.00 \pm 0.00 & 2.00 \pm 0.00 & 2.00 \pm 0.00 & 2.00 \pm 0.00 \\
1.9 & 2.00 \pm 0.00 & 2.00 \pm 0.00 & 2.00 \pm 0.00 & 2.00 \pm 0.00 \\
\end{array}
\]

Table 1: Average number of iterates for qPWAWS to get away from the singular point (mean±STD).

Next, we need to verify that qPWAWS obtains the same $q$-th power median as qPWA if the latter does not get stuck. The experimental procedure is similar to the above. In each $q$-th power median computation, denote the $q$-th power medians of qPWAWS and qPWA by $x_{WAW}$ and $x_{WA}$, respectively. Then we compute the maximum relative difference $|x_{WAW} - x_{WA}|/|x_{WA}|$ of all the $(T - m + 1)$ times of computations, shown in Table 2. Since the differences are close to zero (< 1e −07), qPWAWS obtains nearly the same $q$-th power median as qPWA.

\[
q \quad \text{NYSE(N)} \quad \text{CSI300}
\]

\[
\begin{array}{cccc}
1.1 & 5.9016e-09 & 2.379e-08 & 6.942e-09 & 2.440e-08 \\
1.2 & 6.324e-09 & 1.241e-08 & 6.893e-09 & 6.190e-09 \\
1.3 & 5.454e-09 & 6.801e-09 & 3.861e-09 & 4.015e-09 \\
1.4 & 5.329e-09 & 5.1516e-09 & 2.1728e-09 & 3.9566e-09 \\
1.5 & 4.5439e-09 & 2.6621e-08 & 2.8953e-09 & 2.423e-09 \\
1.6 & 1.889e-09 & 1.806e-09 & 2.0482e-09 & 1.402e-09 \\
1.7 & 1.6098e-09 & 1.3701e-09 & 1.2259e-09 & 1.0679e-09 \\
1.8 & 6.1063e-10 & 7.1172e-10 & 6.8329e-10 & 4.7757e-10 \\
1.9 & 2.1942e-10 & 2.8647e-10 & 3.1570e-10 & 3.0911e-10 \\
\end{array}
\]

Table 2: Maximum relative difference between $q$-th power medians of qPWAWS and qPWA: $|x_{WAW} - x_{WA}|/|x_{WA}|$.

4.2 Computational Costs and Convergence

A computer with an Intel Core i7-6700 CPU and a 4GB DDR3 memory card is used to record the computational time of qPWAWS. $(T − m + 1)$ times of computations are recorded and the mean time cost for each $q$ and $m$ is shown in Table 3. All the computations successfully converge and the average number of iterates for convergence (including the escaping iterates) is also shown in Table 3. All the time costs (< 1s −03s) and the numbers of iterates (< 27) are small, which indicates that qPWAWS runs fast.

4.3 Rate of Convergence (Computational)

In Section 3.5, we have deduced the theoretical rate of convergence for qPWAWS in the special case $M \in \{x_i\}_{i=1}^n$. Now we conduct computational experiments to show that qPWAWS enjoys a reasonable rate of linear convergence in general. The procedure is the same as the above subsection and $(T − m + 1)$ times of computations are implemented. In each computation, we take $Y_\alpha - Y_{(q)}$ as an approximation for the rate of convergence, where $Y_{(q)}$ denotes the last iterate (i.e., the minimum) in Algorithm 1. The means and the STDs of the rates of convergence for qPWAWS with different values of $q$ are shown in Table 4. When $q$ changes from 1 to 1.9, the empirical rate of convergence decreases from about 0.36 to 0.06, thus the newly-developed $1 < q < 2$ algorithm enjoys better rate of convergence than the long-used $q = 1$ algorithm. Besides, all the situations with $1 \leq q < 2$ achieve reasonable rates of linear convergence.

In order to see how the rate of convergence changes as qPWAWS runs, we plot the sequences of rates $\left\{\frac{|Y_{(p)} - Y_{(q)}|}{|Y_{(p)} - Y_{(q)}|}\right\}_{p=1}^{n}$, with $q = 1.1 \sim 1.9$ (one instance for each $q$) and $m = 5$ in Figure 4. Each plot has a slowly-decreasing period and drops sharply in the last few iterates, thus the inflection point is close to the rate of convergence for qPWAWS, which suggests a linear convergence. Besides, the plots with smaller $q$s are above the plots with larger $q$s, which accords with the results in Table 4. Other instances of rates of convergence show similar patterns and we need not plot all of them.

4.4 Investing Performance

We apply the $q$-th power median and qPWAWS to the OPS problem as a price prediction strategy and evaluate the investing performance on the two data sets NYSE(N) and CSI300. Interested readers are referred to [Huang et al., 2016] for the specific model setting. We change $q$ from 1 to 2 to see how it affects the investing performance. We set $m = 5$ to be consistent with the window size of previous works.

The first indicator of evaluation is the final cumulative wealth (CW) when a strategy goes through the whole investment. The second indicator is the daily Sharpe Ratio (SR,
Table 4: Average rate of convergence (mean±STD) for qPWAWS with different values of q.

<table>
<thead>
<tr>
<th>q</th>
<th>NYSE(N)</th>
<th>CSI300</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m = 8</td>
<td>m = 10</td>
</tr>
<tr>
<td>1</td>
<td>0.33 ± 0.03</td>
<td>0.31 ± 0.04</td>
</tr>
<tr>
<td>1.1</td>
<td>0.33 ± 0.03</td>
<td>0.31 ± 0.04</td>
</tr>
<tr>
<td>1.2</td>
<td>0.32 ± 0.03</td>
<td>0.29 ± 0.04</td>
</tr>
<tr>
<td>1.3</td>
<td>0.28 ± 0.03</td>
<td>0.27 ± 0.04</td>
</tr>
<tr>
<td>1.4</td>
<td>0.26 ± 0.03</td>
<td>0.24 ± 0.04</td>
</tr>
<tr>
<td>1.5</td>
<td>0.23 ± 0.03</td>
<td>0.21 ± 0.03</td>
</tr>
<tr>
<td>1.6</td>
<td>0.19 ± 0.03</td>
<td>0.18 ± 0.03</td>
</tr>
<tr>
<td>1.7</td>
<td>0.15 ± 0.02</td>
<td>0.14 ± 0.02</td>
</tr>
<tr>
<td>1.8</td>
<td>0.11 ± 0.02</td>
<td>0.10 ± 0.02</td>
</tr>
<tr>
<td>1.9</td>
<td>0.06 ± 0.01</td>
<td>0.05 ± 0.01</td>
</tr>
</tbody>
</table>

Table 5: Cumulative wealth (CW) and Sharpe Ratio (SR) of qPWAWS with different values of q.

<table>
<thead>
<tr>
<th>q</th>
<th>NYSE(N)</th>
<th>CSI300</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CW</td>
<td>SR</td>
</tr>
<tr>
<td>1</td>
<td>3.3183e+08</td>
<td>0.1034</td>
</tr>
<tr>
<td>1.1</td>
<td>1.0166e+09</td>
<td>0.1082</td>
</tr>
<tr>
<td>1.2</td>
<td>1.1564e+09</td>
<td>0.1086</td>
</tr>
<tr>
<td>1.3</td>
<td>1.1581e+09</td>
<td>0.1087</td>
</tr>
<tr>
<td>1.4</td>
<td>9.3994e+08</td>
<td>0.1078</td>
</tr>
<tr>
<td>1.5</td>
<td>8.4610e+08</td>
<td>0.1074</td>
</tr>
<tr>
<td>1.6</td>
<td>7.3675e+08</td>
<td>0.1068</td>
</tr>
<tr>
<td>1.7</td>
<td>6.6769e+08</td>
<td>0.1063</td>
</tr>
<tr>
<td>1.8</td>
<td>5.7382e+08</td>
<td>0.1056</td>
</tr>
<tr>
<td>1.9</td>
<td>4.7956e+08</td>
<td>0.1047</td>
</tr>
<tr>
<td>2</td>
<td>4.0764e+08</td>
<td>0.1040</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Works

This paper mainly establishes a novel de-singularity subgradient approach and a corresponding algorithm (q-th Power Weiszfeld Algorithm without Singularity, qPWAWS) for the extended Weber location problem. We characterize the subgradient(s) and the optimality of any given singular point. If this singular point is not optimal, the algorithm can escape from it and reduce the cost simultaneously. This advantage makes the sequence of qPWAWS monotonically converge to the exact minimum. A complete proof of convergence for qPWAWS is also presented, which has fixed some incomplete statements of the proofs for some previous Weiszfeld algorithms. qPWAWS enjoys a superlinear convergence for the iteration sequence in the special case where the minimum point is a singular point, which is a new theoretical result.

Experiments with real-world financial data sets indicate that qPWAWS successfully gets out of the singular point with only a small number of iterates. Besides, qPWAWS runs fast, converges with only a few iterates, obtains the same q-th power median as qPWA if the latter does not get stuck, and shows a reasonable rate of linear convergence. In some cases of the online portfolio selection, the q-th power median (1 < q < 2) outperforms the 1-st power median and the 2-nd power median in both the final cumulative wealth and the Sharpe Ratio, which suggests that the q-th power median (1 < q < 2) achieves better investing performance. Therefore, the de-singularity subgradient approach is beneficial to advancing both theory and practice for the extended Weber location problem.

Future works may fall into the following aspects: 1. Find the optimal de-singularity subgradient descent in (19). 2. Establish the de-singularity subgradient theory and algorithm for the q-th power and $L_p$-norm median problem. 3. Extend the de-singularity subgradient theory to other related optimization problems that are also challenged by the singularity problem.
Acknowledgments
This work is supported in part by the National Natural Science Foundation of China under grants 62176103, 62276114, 62206110, in part by Guangdong Basic and Applied Basic Research Foundation under grant 2023B151200064, in part by the Science and Technology Planning Project of Guangzhou under grants 2024A04J9896, 202206030007, Nansha District: 2023ZD001, in part by Guangdong-Macao Advanced Intelligent Computing Joint Laboratory under grant 2020B1212030003, in part by the Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University under grant 2022LSYS003, and in part by the Fundamental Research Funds for the Central Universities, JNU under grant 21623202.

References


[Li et al., 2013] Bin Li, Steven C. H. Hoi, Peilin Zhao, and Vivekanand Gopalkrishnan. Confidence weighted mean


