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

Zhao-Rong Lai[1] , **Xiaotian Wu**[2] , **Liangda Fang**[2] and **Ziliang Chen**[*3,4]
[1]Department of Mathematics, College of Information Science and Technology, Jinan University
[2]Department of Computer Science, College of Information Science and Technology, Jinan University
[3]Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory
[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 $1 \leqslant q < 2$, 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 $\mathbf{x}_1, \cdots, \mathbf{x}_m \in \mathbb{R}^d$ [Weber, 1909; Cooper, 1968; Chen, 1984]:

$$\mathbf{x}_* = \arg\min_{\mathbf{y}} C_q(\mathbf{y}) = \arg\min_{\mathbf{y}} \sum_{i=1}^{m} \|\mathbf{y} - \mathbf{x}_i\|^q, \qquad (1)$$

where $\|\cdot\|^q$ denotes the q-th power of the Euclidean distance, and $C_q(\mathbf{y})$ is the q-th power cost function at the point $\mathbf{y} \in \mathbb{R}^d$. The q-th power median $\mathbf{x}_*$ (especially when $1 \leqslant q < 2$) 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 $C_q(\mathbf{y})$ w.r.t. $\mathbf{y}$:

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

It is well-defined when $q \geqslant 2$ 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 $\{\mathbf{x}_i\}_{i=1}^m$ when $1 \leqslant q < 2$, since at least one of the weights $\|\mathbf{y} - \mathbf{x}_i\|^{q-2}$ becomes undefined (see Figure 1). This paper mainly addresses this case and we assume $1 \leqslant q < 2$ in the rest of this paper if not specified.

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

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

$$\partial C_q(\mathbf{y}) \triangleq \left\{ \mathbf{v} \in \mathbb{R}^d : \liminf_{\substack{\boldsymbol{z}\to\boldsymbol{y}\\ \boldsymbol{z}\neq\boldsymbol{y}}} \frac{C_q(\mathbf{z}) - C_q(\mathbf{y}) - \boldsymbol{v}^\top(\boldsymbol{z}-\boldsymbol{y})}{\|\boldsymbol{z}-\boldsymbol{y}\|_2} \geqslant 0 \right\}. \qquad (3)$$

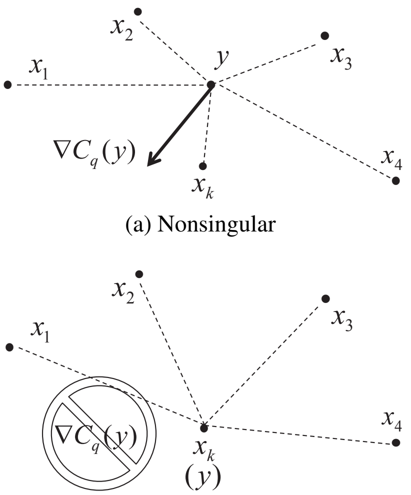
Since C_q is also proper ($\mathrm{dom}\,(C_q) \neq \emptyset$ and $C_q(\mathbf{y}) \neq -\infty$, $\forall \mathbf{y} \in \mathbb{R}^d$), its Fréchet subdifferential is equivalent to the classical subdifferential in the literature of convex analysis.

Definition 2 (Subdifferential in Convex Analysis)**.**

$$\partial C_q(\mathbf{y}) \triangleq \left\{ \mathbf{v} \in \mathbb{R}^d : \forall \mathbf{z},\ C_q(\mathbf{z}) - C_q(\mathbf{y}) - \boldsymbol{v}^\top(\boldsymbol{z}-\boldsymbol{y}) \geqslant 0 \right\}. \qquad (4)$$

[*]Corresponding author.

[**]This is a preprint version of an accepted paper for IJCAI 2024. Contents may be changed in its final version. The supplementary material and code for this paper are available at https://github.com/laizhr/qPWAWS.



(a) Nonsingular

(b) Singular

Figure 1: The singularity problem for the extended Weber location problem (1) with $1 \leqslant q < 2$: (a) The gradient $\nabla C_q(\mathbf{y})$ is well-defined when $\mathbf{y} \notin \{\mathbf{x}_i\}_{i=1}^m$. (b) The gradient $\nabla C_q(\mathbf{y})$ does not necessarily exist when $\mathbf{y}$ hits some $\mathbf{x}_k$.

If C_q is differentiable at $\mathbf{y}$, $\partial C_q(\mathbf{y})$ reduce to a gradient $\nabla C_q(\mathbf{y})$. Based on the properties of C_q ($1 \leqslant q < 2$), it is easy to verify that at least one Fréchet subgradient exists at each singular point $\mathbf{x}_k \in \{\mathbf{x}_i\}_{i=1}^m$ by satisfying the limit inferior inequality in (3). However, few existing works have been done to deal with this singularity problem, because it is usually considered a rare event in practice or just overlooked.

One of the most typical and widely-used gradient approaches is the general q-th Power Weiszfeld Algorithm (qPWA, [Cooper, 1968; Chen, 1984; Aftab *et al.*, 2015]). Since $C_q(\mathbf{y})$ is convex, $\mathbf{y}$ is the minimum of $C_q(\mathbf{y})$ if and only if $\nabla C_q(\mathbf{y}) = \mathbf{0}$. It leads to the following update formula of qPWA at the p-th iteration:

$$\mathbf{y}_{(p+1)} = \frac{\sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}\mathbf{x}_i}{\sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}}, \tag{5}$$

where $\mathbf{y}_{(p)}$ is the current iterate and $\mathbf{y}_{(p+1)}$ is the next iterate computed by a re-weighted sum of the data points $\{\mathbf{x}_i\}_{i=1}^m$. The weights of $\{\mathbf{x}_i\}_{i=1}^m$ are determined by the current distances from $\mathbf{y}_{(p)}$ to $\{\mathbf{x}_i\}_{i=1}^m$ and normalized by their sum.

(5) can be further transformed to:

$$\begin{aligned}\mathbf{y}_{(p+1)} =& \mathbf{y}_{(p)} - \frac{\sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}(\mathbf{y}_{(p)} - \mathbf{x}_i)}{\sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}} \\ \triangleq& \mathbf{y}_{(p)} - \vartheta \nabla C_q(\mathbf{y}_{(p)}), \end{aligned} \tag{6}$$

where $\vartheta = 1/(q\sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2})$ is an automatic step size for the gradient descent approach. Therefore, the Weiszfeld algorithm (5) is actually a gradient descent method (6), which shall fail if $\mathbf{y}_{(p)}$ hits one of $\{\mathbf{x}_i\}_{i=1}^m$ and $1 \leqslant q < 2$.

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 in Table 1. Let $q = 1.1$ and 6 data points be

$$\{\mathbf{x}_i\}_{i=1}^6 \triangleq \{(-2,0),(-1,0),(1,0),(2,0),(0,1),(0,-1)\}.$$

The starting point $\mathbf{y}_{(0)} \triangleq (1.68645, 0)$ is chosen distinct from $\{\mathbf{x}_i\}_{i=1}^6$. The existing algorithm qPWA gets stuck with just 1 iteration since it hits one data point $\mathbf{x}_3 = (1,0)$ and cannot proceed. The resulted cost function is $C_q = 9.4201$, which is not the true minimum. On the contrary, the proposed algorithm qPWAWS successfully escapes from the singular point $\mathbf{x}_3$ and finds the true solution $(0,0)$ with the true minimum $C_q = 8.2871$. Since C_q is a continuous function on $\{\mathbf{x}_i\}_{i=1}^6$, when some $\mathbf{x}_i$ change continuously, the "bad" starting point $\mathbf{y}_{(0)}$ also changes accordingly, possibly throughout the whole $\mathbb{R}^2$. [Chandrasekaran and Tamir, 1989; Vardi and Zhang, 2000] further point out that the bad starting point set $\{\mathbf{y}_{(0)}\}$ itself may constitute a continuum set that can be dense in an open region of $\mathbb{R}^d$ with $d \geqslant 2$, even the entire $\mathbb{R}^d$. Hence this singularity problem is quite significant, if not serious.

$q = 1.1$	Iter	$\hat{\mathbf{x}}$	$C_q(\hat{\mathbf{x}})$
qPWA	1 (get stuck)	$(1,0)$	9.4201
qPWAWS (ours)	24	$(0,0)$	8.2871

Table 1: An example of the singularity problem.

This singularity problem cannot be radically eliminated by straightforward treatments, such as perturbations and random restarts. We remind that $\mathbf{0}$ is not necessarily a subgradient for $C_q(\mathbf{x}_k)$. If $\mathbf{0} \in \partial C_q(\mathbf{x}_k)$, then $C_q(\mathbf{y}) - C_q(\mathbf{x}_k) \geqslant 0$ for all $\mathbf{y}$ based on (4). Thus $\mathbf{x}_k$ should be a minimum point of C_q. However, whether $\mathbf{x}_k$ minimizes C_q is unknown beforehand. Another treatment is to find a starting point with a smaller cost than the costs of any data points $\{\mathbf{x}_i\}_{i=1}^m$ by computing $\{C_q(\mathbf{x}_i)\}_{i=1}^m$ 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 4 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 $\mathbf{x}_k$ is exactly the minimum, we have to check whether $\mathbf{0} \in \partial C_q(\mathbf{x}_k)$. After all, once we figure out $\partial C_q(\mathbf{x}_k)$, 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 $\partial C_q(\mathbf{x}_k)$. More importantly, there are some incomplete statements in the proof of convergence, which will be fixed in Section 3.4. In the $1 < q < 2$ case, neither has the subgradient $\partial C_q(\mathbf{x}_k)$ 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) ($1 \leqslant q < 2$). 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 $\mathbf{x}_k \in \{C_q(\mathbf{x}_i)\}_{i=1}^m$. It can replace the ordinary gradient without increasing computational complexity.
2. If $\mathbf{x}_k$ 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 paper, 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 $\{\mathbf{x}_i\}_{i=1}^m$ 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 $\{\mathbf{x}_i\}_{i=1}^m$ 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-related algorithms for problem (1).

2 Related Works on the Weiszfeld Algorithms

In this section, we assume that the data points $\{\mathbf{x}_i\}_{i=1}^m$ are distinct and non-collinear (i.e., they do not lie in a hyperplane). Thus the cost function $C_q(\mathbf{y})$ 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 qPWA (5), it has been proven that $C_q(\mathbf{y}_{(p+1)}) < C_q(\mathbf{y}_{(p)})$ if $\mathbf{y}_{(p+1)} \neq \mathbf{y}_{(p)}$, 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 $\mathbf{y}_{(p)}$, a weighted 2-nd power cost function $\tilde{C}_q(\mathbf{y})$ can be created to approximate $C_q(\mathbf{y})$:

$$\tilde{C}_q(\mathbf{y}) = \sum_{i=1}^m \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}\|\mathbf{y} - \mathbf{x}_i\|^2. \tag{7}$$

By setting the gradient of $\tilde{C}_q(\mathbf{y})$ 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 L_1-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:

$$\tilde{\mathbf{T}}(\mathbf{y}) = \frac{\sum_{\mathbf{x}_i \neq \mathbf{y}} \|\mathbf{y} - \mathbf{x}_i\|^{-1}\mathbf{x}_i}{\sum_{\mathbf{x}_i \neq \mathbf{y}} \|\mathbf{y} - \mathbf{x}_i\|^{-1}}. \tag{8}$$

Then the next iterate is the combination of $\tilde{\mathbf{T}}(\mathbf{y}_{(p)})$ and the current iterate $\mathbf{y}_{(p)}$:

$$\mathbf{y}_{(p+1)} = (1 - \lambda)\tilde{\mathbf{T}}(\mathbf{y}_{(p)}) + \lambda \mathbf{y}_{(p)}, \tag{9}$$

where $0 \leqslant \lambda \leqslant 1$ 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 λ as follows:

$$\nabla D_1(\mathbf{y}_{(p)}) = \sum_{\mathbf{x}_i \neq \mathbf{y}_{(p)}} \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{-1}(\mathbf{y}_{(p)} - \mathbf{x}_i), \tag{10}$$

$$\lambda = \left\{ \begin{array}{ll} 0 & \text{if} \quad \mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m \\ \min\left\{1, \frac{1}{\|\nabla D_1(\mathbf{y}_{(p)})\|}\right\} & \text{if} \quad \mathbf{y}_{(p)} \in \{\mathbf{x}_i\}_{i=1}^m \end{array} \right. . \tag{11}$$

Compared with (2), $\nabla D_1(\mathbf{y}_{(p)})$ removes the singular term in $\nabla C_1(\mathbf{y}_{(p)})$. If $\mathbf{y}_{(p)}$ does not hit any data point, then (9) is just the same as (5). Else, λ ensures $C_1(\mathbf{y}_{(p+1)}) \leqslant C_1(\mathbf{y}_{(p)})$ 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 ($1 \leqslant q < 2$) framework for the integrity of the approach.

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

$$C_q(\mathbf{y}) = \sum_{i=1}^m \xi_i \|\mathbf{y} - \mathbf{x}_i\|^q, \tag{12}$$

where $\xi_i > 0$ is the multiplicity of the data point $\mathbf{x}_i$. It can deal with duplication and collinearity of data points. For example, if $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3$ and the other points are distinct, then we can merge $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and set $\xi_1 = 3$, $\xi_i = 1 (i \geqslant 4)$. 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 $C_q(\mathbf{y})$ is strictly convex on $\mathbf{y}$ and there is a unique minimum point $\mathbf{M}$. To be convenient for illustrations, we set $\eta_i \triangleq \xi_i^{\frac{1}{q}}$

and further change (12) to:

$$C_q(\mathbf{y}) = \sum_{i=1}^{m} \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^q. \tag{13}$$

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 $\mathbf{y}_{(p)}$ does not coincide with the data points $\{\mathbf{x}_i\}_{i=1}^m$. We first state the general iterative update formula in this case:

$$\mathbf{y}_{(p+1)} = \mathbf{T}_1(\mathbf{y}_{(p)}) \triangleq \frac{\sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}\mathbf{x}_i}{\sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}}. \tag{14}$$

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 $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$, then $C_q(\mathbf{T}_1(\mathbf{y}_{(p)})) \leqslant C_q(\mathbf{y}_{(p)})$ with equality only when $\mathbf{T}_1(\mathbf{y}_{(p)}) = \mathbf{y}_{(p)}$.*

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 $\mathbf{M}$ of $C_q(\mathbf{y})$ in (13). The following corollary is a direct result of Theorem 3.

Corollary 4. *If $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$, then $\mathbf{T}_1(\mathbf{y}_{(p)}) = \mathbf{y}_{(p)} \Leftrightarrow \mathbf{y}_{(p)}$ is the minimum point $\mathbf{M}$ of $C_q(\mathbf{y})$.*

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

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

$$D_q(\mathbf{y}) \triangleq \sum_{i \neq k} \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^q, \tag{15}$$

$$\nabla D_q(\mathbf{y}) = \sum_{i \neq k} q\eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}(\mathbf{y} - \mathbf{x}_i), \quad 1 \leqslant q < 2. \tag{16}$$

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

Theorem 6 (Characterization of Subgradients and Minimum)**.** *Let $\mathbf{y}_{(p)} = \mathbf{x}_k$. Then*

$$\partial C_q(\mathbf{x}_k) = \begin{cases} \{\nabla D_1(\mathbf{x}_k) + \eta_k \mathbf{u} : \forall \mathbf{u}, \|\mathbf{u}\| \leqslant 1\} & \textit{if} \quad q=1 \\ \{\nabla D_q(\mathbf{x}_k)\} & \textit{if} \quad 1<q<2 \end{cases}. \tag{17}$$

According to Fermat's rule, $\mathbf{x}_k$ is the minimum point $\mathbf{M}$ if and only if $\mathbf{0} \in \partial C_q(\mathbf{x}_k)$.

The proof is provided in Supplementary B.3. It is obvious that $\nabla D_q(\mathbf{x}_k) \in \partial C_q(\mathbf{x}_k)$ for all $1 \leqslant q < 2$. This is why $\nabla D_q(\mathbf{x}_k)$ 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 $\mathbf{y}_{(p)} = \mathbf{x}_k$ and $\mathbf{u} = \mathbf{0}$.

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

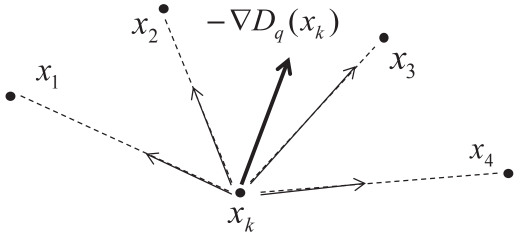


Figure 2: $-\nabla D_q(\mathbf{x}_k)$ can be interpreted as the resultant implemented on $\mathbf{x}_k$ towards other data points.

We have characterized the minimum no matter whether the current iterate $\mathbf{y}_{(p)}$ hits the data points $\{\mathbf{x}_i\}_{i=1}^m$ or not. If $\mathbf{y}_{(p)}$ has not reached the minimum point $\mathbf{M}$, then another iteration should be implemented. When $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$, the general formula (14) can be adopted. When $\mathbf{y}_{(p)} \in \{\mathbf{x}_i\}_{i=1}^m$, we will show later that $\mathbf{y}_{(p)} - \lambda\nabla D_q(\mathbf{y}_{(p)})$ can get away from the singular point and reduce the q-th power cost simultaneously.

3.3 Update Formula with Coincidence

The update formula with coincidence $\mathbf{y}_{(p)} = \mathbf{x}_k$ 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 $\mathbf{y}_{(p)} = \mathbf{x}_k$ and $\|\nabla D_q(\mathbf{x}_k)\| > 0$, then there exists a $\lambda_* > 0$ such that for any $0 < \lambda \leqslant \lambda_*$, $C_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k)) < C_q(\mathbf{x}_k)$, $1 < q < 2$.*

The proof is provided in Supplementary B.4. The key point of Theorem 7 is that the negative de-singularity subgradient part (resultant) $-\lambda\|\nabla D_q(\mathbf{x}_k)\|^2$ is of lower order infinitesimal than the positive singular part (resistance)

$\eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q$. Thus when λ is sufficiently small, the resultant overcomes the resistance and drags the current iterate $\mathbf{y}_{(p)}$ 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 $\frac{\lambda}{2}G(\mathbf{x}_k) + o(\lambda)$ on the left side of (43), we can approximately start with the following λ_0:

$$\lambda_0 = \min\left\{\frac{1}{q}\eta_k^{-\frac{q}{q-1}}\|\nabla D_q(\mathbf{x}_k)\|^{\frac{2-q}{q-1}}, 1\right\}. \tag{18}$$

As long as $C_q(\mathbf{x}_k - \lambda_w \nabla D_q(\mathbf{x}_k)) \geqslant C_q(\mathbf{x}_k)$, we reduce λ_w with a factor $\rho < 1$: $\lambda_{w+1} \leftarrow \rho\lambda_w$, $w = 0, 1, \cdots$, until we find some λ_* such that $C_q(\mathbf{x}_k - \lambda_* \nabla D_q(\mathbf{x}_k)) < C_q(\mathbf{x}_k)$ (see Figure 3). According to Theorem 7, this λ_* is sure to be found. Then the next iterate is:

$$\mathbf{y}_{(p+1)} = \mathbf{T}_2(\mathbf{x}_k) \triangleq \mathbf{x}_k - \lambda_* \nabla D_q(\mathbf{x}_k). \tag{19}$$

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 2). It can be even more efficient if we further reduce λ_0 and ρ. Besides, the choice of λ_* 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.

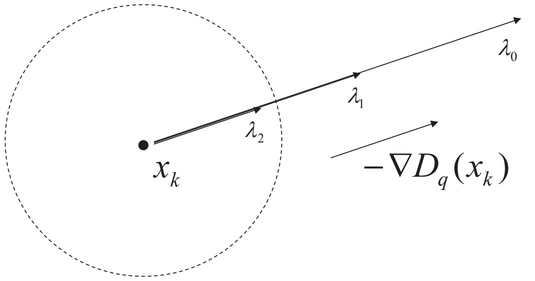


Figure 3: Theorem 7 indicates that a small displacement from $\mathbf{x}_k$ towards $-\nabla D_q(\mathbf{x}_k)$ can reduce the q-th power cost. Thus we can start from a λ_0 and reduce it with a factor $\rho < 1$ at each time, until $\mathbf{x}_k - \lambda_* \nabla D_q(\mathbf{x}_k)$ reduces the cost.

3.4 A Complete Proof of Convergence

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

$$\mathbf{y}_{(p+1)} = \mathbf{T}(\mathbf{y}_{(p)}) \triangleq \begin{cases} \mathbf{T}_1(\mathbf{y}_{(p)}) = \frac{\sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}\mathbf{x}_i}{\sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}} \\ \qquad\qquad \text{if } \mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m \\ \mathbf{T}_2(\mathbf{y}_{(p)}) = \mathbf{x}_k - \lambda_* \nabla D_q(\mathbf{x}_k) \\ \qquad\qquad \text{if } \mathbf{y}_{(p)} = \mathbf{x}_k \end{cases}, \tag{20}$$

where $\mathbf{T}_1(\mathbf{y}_{(p)})$ is the update formula without coincidence (14) and $\mathbf{T}_2(\mathbf{y}_{(p)})$ 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 $\mathbf{y}_{(0)}$, the qPWAWS generates a sequence of iterates $\mathbf{y}_{(0)}, \mathbf{y}_{(1)}, \cdots, \mathbf{y}_{(p)} \cdots$. It is necessary to prove that the sequence converges to the minimum point $\mathbf{M}$ unless some $\mathbf{y}_{(p)}$ hits $\mathbf{M}$.

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

The proof is provided in Supplementary B.5. It is also possible that the sequence $\{\mathbf{y}_{(p)}\}$ converges to the data points $\{\mathbf{x}_i\}_{i=1}^m$. In fact, $\mathbf{T}_1(\mathbf{y})$ is continuous in almost the whole $\mathbb{R}^d$ except for the data points $\{\mathbf{x}_i\}_{i=1}^m$ where it is discontinuous. However, $\{\mathbf{x}_i\}_{i=1}^m$ 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.

$$\lim_{\mathbf{y} \to \mathbf{x}_k} \mathbf{T}_1(\mathbf{y}) = \mathbf{x}_k, \quad \forall 1 \leqslant k \leqslant m, \quad \forall 1 \leqslant q < 2. \tag{21}$$

The proof is provided in Supplementary B.6. Based on this lemma, the operator $\mathbf{T}_1$ can be extended to the removable discontinuous data points:

$$\mathbf{T}_1(\mathbf{x}_k) \triangleq \mathbf{x}_k, \quad \forall 1 \leqslant k \leqslant m, \quad \forall 1 \leqslant q < 2. \tag{22}$$

From (22) and Corollary 4, all the fixed points of $\mathbf{T}_1$ are characterized as:

$$\mathbf{T}_1(\mathbf{y}) = \mathbf{y} \iff \mathbf{y} \in \{\mathbf{x}_i\}_{i=1}^m \bigcup \{\mathbf{M}\}. \tag{23}$$

When $\mathbf{y}$ gets into some neighborhood of a data point $\mathbf{x}_k \neq \mathbf{M}$, the operator $\mathbf{T}_1$ will eventually drive it out of the neighborhood. This property of $\mathbf{T}_1$ 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 $\mathbf{x}_k \neq \mathbf{M}$, then there exists some $\delta_0 > 0$ such that for all $\mathbf{y} \in B(\mathbf{x}_k, \delta_0)$, $\mathbf{T}_1^{s-1}(\mathbf{y}) \in B(\mathbf{x}_k, \delta_0)$ and $\mathbf{T}_1^s(\mathbf{y}) \notin B(\mathbf{x}_k, \delta_0)$ for some s. $B(\mathbf{x}_k, \delta_0)$ denotes an open δ_0-ball centered at $\mathbf{x}_k$. $\mathbf{T}_1^s$ means implementing $\mathbf{T}_1$ for s times, and s depends on $\mathbf{y}$.*

The proof is provided in Supplementary B.7.

Theorem 11 (Convergence Theorem)**.** *Starting from **any initial point** $\mathbf{y}_{(0)}$, if the sequence $\{\mathbf{y}_{(p)}\}$ generated by qPWAWS (20) does not hit $\mathbf{M}$, then $\mathbf{y}_{(p)} \to \mathbf{M}$. If $\mathbf{y}_{(p)}$ hits $\mathbf{M}$, 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 $\lim_{v \to \infty} \mathbf{y}_{(p_v)} = \mathbf{x}_k$ but $\mathbf{T}_1(\mathbf{y}_{(p_v)}) \notin B(\mathbf{x}_k, \delta_0)$ for all v. However, Lemma 10 indicates that $\mathbf{T}_1$ may be implemented for $s > 1$ times to drive $\mathbf{y}_{(p_v)}$ out of $B(\mathbf{x}_k, \delta_0)$. Besides, this s also depends on $\mathbf{y}_{(p_v)}$ and the uniform upper bound of s for all the $\mathbf{y}_{(p_v)}$ may not exist. Thus it may not be concluded that $\lim_{v\to\infty}\frac{\|\mathbf{T}_1(\mathbf{y}_{(p_v)})-\mathbf{x}_k\|}{\|\mathbf{y}_{(p_v)}-\mathbf{x}_k\|} = \infty$ 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 $\{\mathbf{y}_{(p_v)}\}$ converges to $\mathbf{M}$, but not the whole sequence $\{\mathbf{y}_{(p)}\}$. The convergence of the whole sequence $\{\mathbf{y}_{(p)}\}$ should depend on the uniqueness of $\mathbf{M}$.

2. In [Aftab *et al.*, 2015], the proof of convergence does not include Lemma 10. Thus it may not guarantee that a subsequence $\mathbf{y}_{(p_u)} \in B(\mathbf{x}_k, \delta_0)$ and $\mathbf{T}_1(\mathbf{y}_{(p_u)}) \notin B(\mathbf{x}_k, \delta_0)$. 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 qPWAWS when $\mathbf{y}_{(p)} \to \mathbf{M}$. With the proposed desingularity subgradient $\tilde{\nabla} D_q$ in (16), we can deduce the exact rate of convergence for qPWAWS in the special case $\mathbf{M} \in \{\mathbf{x}_i\}_{i=1}^m$. To the best of our knowledge, this is a new theoretical result that qPWAWS enjoys a superlinear convergence for the iteration sequence in this case. Without loss of generality, we assume $\mathbf{M} = \mathbf{x}_k$. 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 $\|\tilde{\nabla} D_q(\mathbf{y}_{(p)})\|$ with $\mathbf{y}_{(p)} \to \mathbf{x}_k$ at first.

Lemma 12. *If $1 < q < 2$, the order of infinitesimal of $\|\tilde{\nabla} D_q(\mathbf{y}_{(p)})\|$ is no higher than that of $\|\mathbf{y}_{(p)} - \mathbf{x}_k\|$ with $\mathbf{y}_{(p)} \to \mathbf{x}_k$. Precisely, there exists some $\zeta \geqslant 0$ such that*

$$\lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k} \frac{\|\tilde{\nabla} D_q(\mathbf{y}_{(p)})\|}{\|\mathbf{y}_{(p)} - \mathbf{x}_k\|} \leqslant \zeta, \quad 1 < q < 2. \tag{24}$$

The proof is provided in Supplementary B.9.

Theorem 13. *If $\mathbf{M} = \mathbf{x}_k$ for some k, the rate of convergence for qPWAWS is:*

$$\lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k} \frac{\|\mathbf{y}_{(p+1)} - \mathbf{x}_k\|}{\|\mathbf{y}_{(p)} - \mathbf{x}_k\|} = \begin{cases} \frac{\|\tilde{\nabla} D_1(\mathbf{x}_k)\|}{\eta_k} & q = 1, \\ 0 & 1 < q < 2. \end{cases} \tag{25}$$

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 $\mathbf{M} \notin \{\mathbf{x}_i\}_{i=1}^m$, the rate of convergence is unknown at present according to our knowledge. It is difficult to eliminate the infinitesimal $\|\mathbf{y}_{(p)} - \mathbf{M}\|$ and obtain a uniform upper bound of $\lim_{\mathbf{y}_{(p)}\to\mathbf{M}} \frac{\|\mathbf{y}_{(p+1)}-\mathbf{M}\|}{\|\mathbf{y}_{(p)}-\mathbf{M}\|}$ that is no greater than 1. When $q = 1$, [Ostresh, 1978] shows some computational evidence that the rate of convergence for $\mathbf{M} \notin \{\mathbf{x}_i\}_{i=1}^m$ is usually somewhat less than that of $\mathbf{M} \in \{\mathbf{x}_i\}_{i=1}^m$. We will give some computational results for the $1 < q < 2, \mathbf{M} \notin \{\mathbf{x}_i\}_{i=1}^m$ case in Section 4.3 to show that qPWAWS enjoys a reasonable rate of linear convergence. We summarize the whole qPWAWS 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 qPWAWS 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 $\mathbf{x}_i \in \mathbb{R}^d$ indicates a price vector of d assets on the i-th day. $\{\mathbf{x}_i\}_{i=1}^m$ contains the asset prices on the most recent m days. Following [Huang *et al.*, 2016], we fed the q-th power median

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{y}} \sum_{i=1}^m \|\mathbf{y} - \mathbf{x}_i\|^q, \quad 1 \leqslant q \leqslant 2 \tag{26}$$

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[1] 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 qPWAWS 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 qPWAWS. Third, we give computational results on the rate of convergence for qPWAWS. 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 $\rho = 0.1$, 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 $\{\mathbf{x}_i\}_{i=1}^m$. Thus the q-th power median computation can be conducted for $(T - m + 1)$ times to evaluate the average performance of qPWAWS.

4.1 Addressing the Singularity Problem

If the sequence of iterates does not hit the data points, then qPWAWS is equivalent to qPWA. Without loss of generality, we examine a specific case: let the starting point $\mathbf{y}_{(0)} = \mathbf{x}_1$

[1] http://www.csindex.com.cn

for qPWAWS and $\mathbf{y}_{(0)} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$ for qPWA. Besides, we also watch to avoid getting stuck for qPWA, since it cannot deal with singularity.

To evaluate the efficiency of qPWAWS, we measure how many iterates that qPWAWS 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 2. qPWAWS 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 fact, a smaller ρ and a smaller λ_0 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	NYSE(N)		CSI300	
	$m=5$	$m=10$	$m=5$	$m=10$
1.1	2.01 ± 0.49	2.19 ± 0.46	2.54 ± 0.50	2.75 ± 0.43
1.2	2.00 ± 0.41	2.17 ± 0.41	2.38 ± 0.49	2.71 ± 0.46
1.3	1.99 ± 0.31	2.15 ± 0.37	2.19 ± 0.39	2.64 ± 0.48
1.4	1.99 ± 0.22	2.12 ± 0.32	2.03 ± 0.17	2.52 ± 0.50
1.5	1.99 ± 0.10	2.07 ± 0.25	2.00 ± 0.00	2.28 ± 0.45
1.6	2.00 ± 0.04	2.03 ± 0.18	2.00 ± 0.00	2.06 ± 0.23
1.7	2.00 ± 0.00	2.00 ± 0.05	2.00 ± 0.00	2.00 ± 0.00
1.8	2.00 ± 0.00	2.00 ± 0.00	2.00 ± 0.00	2.00 ± 0.00
1.9	2.00 ± 0.00	2.00 ± 0.00	2.00 ± 0.00	2.00 ± 0.00

Table 2: 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 $\hat{\mathbf{x}}_{WAWS}$ and $\hat{\mathbf{x}}_{WA}$, respectively. Then we compute the maximum relative difference $\|\hat{\mathbf{x}}_{WAWS} - \hat{\mathbf{x}}_{WA}\| / \|\hat{\mathbf{x}}_{WA}\|$ of all the $(T - m + 1)$ times of computations, shown in Table 3. Since the differences are close to zero ($< 1e - 07$), qPWAWS obtains nearly the same q-th power median as qPWA.

q	NYSE(N)		CSI300	
	$m=5$	$m=10$	$m=5$	$m=10$
1.1	$5.9010e-09$	$2.3797e-08$	$6.9425e-09$	$2.4404e-08$
1.2	$6.3249e-09$	$1.2411e-08$	$6.8934e-09$	$6.1940e-09$
1.3	$5.4543e-09$	$6.8011e-09$	$3.8610e-09$	$4.0452e-09$
1.4	$5.2329e-09$	$5.1516e-09$	$2.1728e-09$	$3.9566e-09$
1.5	$4.5439e-09$	$2.6621e-09$	$2.8953e-09$	$2.4237e-09$
1.6	$1.8894e-09$	$1.8003e-09$	$2.0482e-09$	$1.4021e-09$
1.7	$1.4698e-09$	$1.3701e-09$	$1.2259e-09$	$1.0679e-09$
1.8	$6.1063e-10$	$7.1172e-10$	$6.8329e-10$	$4.7757e-10$
1.9	$2.1942e-10$	$2.8647e-10$	$3.1570e-10$	$3.0911e-10$

Table 3: Maximum relative difference between q-th power medians of qPWAWS and qPWA: $\|\hat{\mathbf{x}}_{WAWS} - \hat{\mathbf{x}}_{WA}\| / \|\hat{\mathbf{x}}_{WA}\|$.

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 4. All the computations successfully converge and the average number of iterates for convergence (including the escaping iterates) is also shown in Table 4. All the time costs ($< 1e - 03s$) and the numbers of iterates (< 27) are small, which indicates that qPWAWS runs fast.

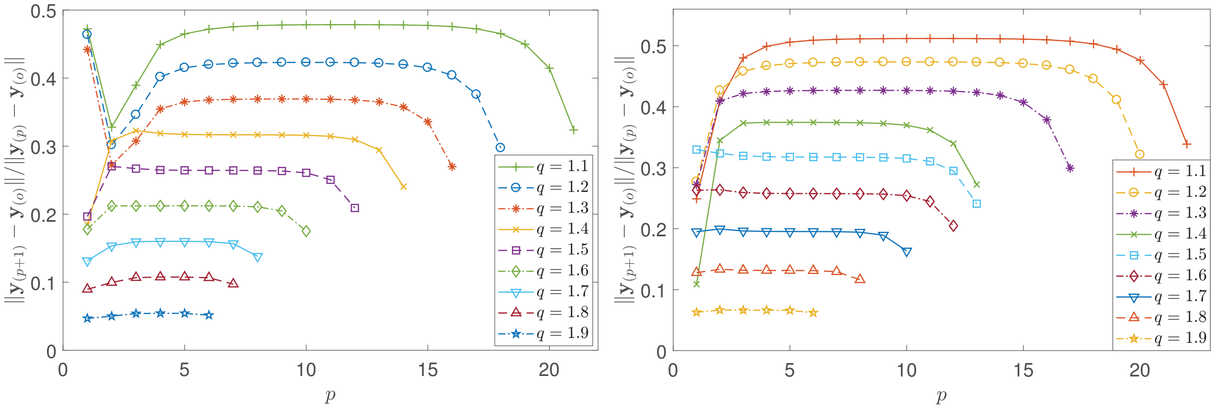
q	$m=5$		$m=10$	
	Time	Iters	Time	Iters
	NYSE(N)			
1.1	$3.6059e-04$	25.31 ± 4.16	$6.0150e-04$	25.03 ± 6.50
1.2	$3.3310e-04$	22.78 ± 3.48	$5.5750e-04$	22.38 ± 4.64
1.3	$3.0204e-04$	20.58 ± 3.06	$4.8230e-04$	20.08 ± 3.51
1.4	$2.7411e-04$	18.51 ± 2.58	$4.4243e-04$	18.02 ± 2.70
1.5	$2.5000e-04$	16.61 ± 2.12	$3.9745e-04$	16.12 ± 2.09
1.6	$2.3567e-04$	14.85 ± 1.70	$3.5268e-04$	14.36 ± 1.60
1.7	$2.0549e-04$	13.30 ± 1.20	$3.1274e-04$	12.71 ± 1.21
1.8	$2.0305e-04$	11.71 ± 0.84	$2.7530e-04$	11.07 ± 0.91
1.9	$1.5853e-04$	9.98 ± 0.55	$2.2369e-04$	9.31 ± 0.63
	CSI300			
1.1	$5.0074e-04$	26.86 ± 4.85	$7.4105e-04$	26.89 ± 7.87
1.2	$3.9539e-04$	23.77 ± 3.70	$7.2515e-04$	23.97 ± 5.30
1.3	$4.2771e-04$	21.09 ± 3.19	$6.4174e-04$	21.49 ± 3.88
1.4	$3.8925e-04$	18.49 ± 2.74	$5.9630e-04$	19.18 ± 2.87
1.5	$3.2568e-04$	16.21 ± 2.31	$5.1977e-04$	16.85 ± 2.18
1.6	$3.2078e-04$	14.32 ± 2.09	$4.5260e-04$	14.82 ± 1.64
1.7	$3.2034e-04$	13.25 ± 1.39	$4.2355e-04$	13.06 ± 1.21
1.8	$2.8558e-04$	11.78 ± 0.96	$3.8196e-04$	11.33 ± 0.85
1.9	$2.0650e-04$	10.06 ± 0.61	$3.2261e-04$	9.48 ± 0.56

Table 4: Average computational time (in seconds) and average number of iterates (mean±STD) for convergence of qPWAWS.

4.3 Rate of Convergence (Computational)

In Section 3.5, we have deduced the theoretical rate of convergence for qPWAWS in the special case $\mathbf{M} \in \{\mathbf{x}_i\}_{i=1}^{m}$. 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 subsections and $(T - m + 1)$ times of computations are implemented. In each computation, we take $\frac{\|\mathbf{y}_{(o-1)} - \mathbf{y}_{(o)}\|}{\|\mathbf{y}_{(o-2)} - \mathbf{y}_{(o)}\|}$ as an approximation for the rate of convergence, where $\mathbf{y}_{(o)}$ 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 5. 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 $1 \leqslant q < 2$ 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 $\left\{ \frac{\|\mathbf{y}_{(p+1)} - \mathbf{y}_{(o)}\|}{\|\mathbf{y}_{(p)} - \mathbf{y}_{(o)}\|} \right\}_{p=1}^{o-2}$ 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 qs are above the plots with larger qs, which accords with the results in Table 5. Other instances of rates of convergence show similar patterns and we need not plot all of them.



(a) NYSE(N)

(b) CSI300

Figure 4: Sequences of rates of convergence $\left\{ \frac{\|\mathbf{y}_{(p+1)} - \mathbf{y}_{(o)}\|}{\|\mathbf{y}_{(p)} - \mathbf{y}_{(o)}\|} \right\}_{p=1}^{o-2}$ for qPWAWS with $q = 1.1 \sim 1.9$ and $m = 5$.

q	NYSE(N)		CSI300	
	$m = 5$	$m = 10$	$m = 5$	$m = 10$
1	0.35 ± 0.03	0.33 ± 0.04	0.36 ± 0.03	0.34 ± 0.04
1.1	0.33 ± 0.03	0.31 ± 0.04	0.33 ± 0.03	0.32 ± 0.04
1.2	0.31 ± 0.03	0.29 ± 0.04	0.31 ± 0.03	0.30 ± 0.04
1.3	0.28 ± 0.03	0.27 ± 0.04	0.29 ± 0.03	0.27 ± 0.04
1.4	0.26 ± 0.03	0.24 ± 0.04	0.26 ± 0.03	0.25 ± 0.04
1.5	0.23 ± 0.03	0.21 ± 0.03	0.23 ± 0.03	0.22 ± 0.03
1.6	0.19 ± 0.03	0.18 ± 0.03	0.19 ± 0.03	0.18 ± 0.03
1.7	0.15 ± 0.02	0.14 ± 0.02	0.15 ± 0.03	0.15 ± 0.03
1.8	0.11 ± 0.02	0.10 ± 0.02	0.11 ± 0.02	0.10 ± 0.02
1.9	0.06 ± 0.01	0.05 ± 0.01	0.06 ± 0.01	0.05 ± 0.01

Table 5: Average rate of convergence (mean±STD) for qPWAWS with different values of q.

q	NYSE(N)		CSI300	
	CW	SR	CW	SR
1	$3.3183e + 08$	0.1034	1.7750	0.0479
1.1	$1.0166e + 09$	0.1082	1.7803	0.0481
1.2	$1.1564e + 09$	0.1086	1.7544	0.0473
1.3	$\mathbf{1.1581e + 09}$	**0.1087**	1.7447	0.0470
1.4	$9.3994e + 08$	0.1078	1.7469	0.0471
1.5	$8.4610e + 08$	0.1074	1.8024	0.0489
1.6	$7.3675e + 08$	0.1068	**1.8434**	**0.0502**
1.7	$6.6769e + 08$	0.1063	1.8120	0.0492
1.8	$5.7382e + 08$	0.1056	1.7892	0.0485
1.9	$4.7956e + 08$	0.1047	1.7591	0.0475
2	$4.0764e + 08$	0.1040	1.7311	0.0466

Table 6: Cumulative wealth (CW) and Sharpe Ratio (SR) of qPWAWS with different values of q.

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, [Sharpe, 1966]), which is a risk-adjusted average return that considers both return and risk in the investment. The results in Table 6 indicate that $q = 1.3$ and $q = 1.6$ achieve the best CW and SR on NYSE(N) and CSI300, respectively. They improve on the trivial $q = 1$ and $q = 2$ to some extent, which suggests 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. Hence the q-th power median and qPWAWS are useful in this machine learning scenario.

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 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

[Afsari *et al.*, 2013] Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the riemannian center of mass. *SIAM journal on control and optimization*, 51(3):2230–2260, 2013.

[Aftab *et al.*, 2015] Khurrum Aftab, Richard Hartley, and Jochen Trumpf. Generalized weiszfeld algorithms for l_q optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):728–745, Apr. 2015.

[Beck and Sabach, 2015] Amir Beck and Shoham Sabach. Weiszfeld's method: Old and new results. *Journal of Optimization Theory and Applications*, 164(1):1–40, Jan. 2015.

[Brimberg and Love, 1993] Jack Brimberg and Robert F. Love. Global convergence of a generalized iterative procedure for the minisum location problem with lp distances. *Operations Research*, 41(6):1153–1163, 1993.

[Chandrasekaran and Tamir, 1989] R. Chandrasekaran and A. Tamir. Open questions concerning weiszfeld's algorithm for the fermat-weber location problem. *Mathematical Programming*, 44:293–295, 1989.

[Chartrand and Yin, 2008] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872, 2008.

[Chen, 1984] Reuven Chen. Solution of location problems with radial cost functions. *Computers & Mathematics with Applications*, 10(1):87–94, 1984.

[Cooper, 1968] Leon Cooper. An extension of the generalized weber problem. *Journal of Regional Science*, 8(2):181–197, Dec. 1968.

[Daubechies *et al.*, 2010] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, Jan. 2010.

[Eldar and Mishali, 2009] Yonina C. Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[Fletcher *et al.*, 2009] P. Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, Mar 2009.

[Hartley *et al.*, 2011] Richard Hartley, Khurrum Aftab, and Jochen Trumpf. L1 rotation averaging using the weiszfeld algorithm. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2011.

[Hartley *et al.*, 2013] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International Journal of Computer Vision*, 103(3):267–305, 2013.

[Huang *et al.*, 2016] Dingjiang Huang, Junlong Zhou, Bin Li, Steven C. H. Hoi, and Shuigeng Zhou. Robust median reversion strategy for online portfolio selection. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2480–2493, Sep. 2016.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[Kuhn, 1973] Harold W. Kuhn. A note on fermat's problem. *Mathematical Programming*, 4:98–107, 1973.

[Lai and Yang, 2023] Zhao-Rong Lai and Haisheng Yang. A survey on gaps between mean-variance approach and exponential growth rate approach for portfolio optimization. *ACM Computing Surveys*, 55(2):1–36, Mar. 2023. Article No. 25.

[Lai *et al.*, 2018a] Zhao-Rong Lai, Dao-Qing Dai, Chuan-Xian Ren, and Ke-Kun Huang. A peak price tracking based learning system for portfolio selection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):2823–2832, Jul. 2018.

[Lai *et al.*, 2018b] Zhao-Rong Lai, Dao-Qing Dai, Chuan-Xian Ren, and Ke-Kun Huang. Radial basis functions with adaptive input and composite trend representation for portfolio selection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6214–6226, Dec. 2018.

[Lai *et al.*, 2018c] Zhao-Rong Lai, Pei-Yi Yang, Liangda Fang, and Xiaotian Wu. Short-term sparse portfolio optimization based on alternating direction method of multipliers. *Journal of Machine Learning Research*, 19(63):1–28, 2018.

[Lai *et al.*, 2020] Zhao-Rong Lai, Liming Tan, Xiaotian Wu, and Liangda Fang. Loss control with rank-one covariance estimate for short-term portfolio optimization. *Journal of Machine Learning Research*, 21(97):1–37, Jun. 2020.

[Lai *et al.*, 2022] Zhao-Rong Lai, Cheng Li, Xiaotian Wu, Quanlong Guan, and Liangda Fang. Multitrend conditional value at risk for portfolio optimization. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

[Li *et al.*, 2013] Bin Li, Steven C. H. Hoi, Peilin Zhao, and Vivekanand Gopalkrishnan. Confidence weighted mean reversion strategy for online portfolio selection. *ACM Transactions on Knowledge Discovery from Data*, 7(1), Mar. 2013. Article 4.

[Li *et al.*, 2016] Bin Li, Doyen Sahoo, and Steven C.H. Hoi. OLPS: a toolbox for on-line portfolio selection. *Journal of Machine Learning Research*, 17(1):1242–1246, 2016.

[Nesterov, 1983] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Soviet Mathematics Doklady*, 269:543–547, 1983.

[Ostresh, 1978] Lawrence M. Ostresh. On the convergence of a class of iterative methods for solving the weber location problem. *Operations Research*, 26(4):597–609, Jul.-Aug. 1978.

[Rockafellar and Wets, 2009] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[Sharpe, 1966] William F. Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, Jan. 1966.

[Sutskever *et al.*, 2013] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, Jun 2013.

[Vardi and Zhang, 2000] Yehuda Vardi and Cun-Hui Zhang. The multivariate ℓ^1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97(4):1423–1426, Feb. 2000.

[Weber, 1909] Alfred Weber. *Uber den Standort der Industrien*. Tubingen : Mohr, 1909.

[Weiszfeld, 1937] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.

[Yang, 2010] Le Yang. Riemannian median and its estimation. *LMS Journal of Computation and Mathematics*, 13:461–479, 2010.

Supplementary Material

A Additional Content

A.1 Update Formula with Coincidence for $q = 1$

The $q = 1$ case is solved by [Vardi and Zhang, 2000] and interested readers are referred to it for a detailed proof. We directly give the $q = 1$ formula as follows:

$$\tilde{\mathbf{T}}(\mathbf{y}_{(p)}) = \frac{\sum_{i \neq k} \eta_i \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{-1} \mathbf{x}_i}{\sum_{i \neq k} \eta_i \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{-1}}, \tag{27}$$

$$\mathbf{y}_{(p+1)} = (1 - \lambda)\tilde{\mathbf{T}}(\mathbf{y}_{(p)}) + \lambda \mathbf{y}_{(p)},$$

$$\lambda = \min \left\{ 1, \frac{\eta_k}{\|\nabla D_1(\mathbf{y}_{(p)})\|} \right\}. \tag{28}$$

This strategy ensures $\mathbf{y}_{(p+1)} \neq \mathbf{y}_{(p)} \Leftrightarrow C_1(\mathbf{y}_{(p+1)}) < C_1(\mathbf{y}_{(p)})$ and $\mathbf{y}_{(p+1)} = \mathbf{y}_{(p)} \Leftrightarrow \mathbf{y}_{(p)} = \mathbf{M}$.

A.2 Solving Algorithm

Note that the multiplicities are changed from $\{\eta_i\}_{i=1}^m$ back to $\{\xi_i\}_{i=1}^m$ to simplify the expressions.

Algorithm 1 q-th power Weiszfeld algorithm without singularity (qPWAWS)

```
Require: Given $m$ distinct data points $\{\mathbf{x}_i\}_{i=1}^m$, the corresponding multiplicities $\{\xi_i\}_{i=1}^m$, the order of power $q$, the reducing factor $\rho$ and the tolerance threshold $Tol$.
  1. Initialize with a starting point $\mathbf{y}_{(0)}$.
  while 1 do
    if $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$ then
      2. Compute $\mathbf{y}_{(p+1)} = \frac{\sum_{i=1}^m \xi_i \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2} \mathbf{x}_i}{\sum_{i=1}^m \xi_i \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2}}$.
      if $\mathbf{y}_{(p+1)} = \mathbf{y}_{(p)}$ then
        3. $\mathbf{M} = \mathbf{y}_{(p)}$. Break.
      end if
      4. $p \leftarrow p + 1$.
    else
      5. Suppose $\mathbf{y}_{(p)} = \mathbf{x}_k$,
        compute $\nabla D_q(\mathbf{x}_k) = \sum_{i \neq k} q \xi_i \|\mathbf{x}_k - \mathbf{x}_i\|^{q-2} (\mathbf{x}_k - \mathbf{x}_i)$.
      if $q = 1$ then
        if $\|\nabla D_1(\mathbf{x}_k)\| \leqslant \xi_k$ then
          6. $\mathbf{M} = \mathbf{x}_k$. Break.
        else
          7. Compute $\tilde{\mathbf{T}}(\mathbf{x}_k) = \frac{\sum_{i \neq k} \xi_i \|\mathbf{x}_k - \mathbf{x}_i\|^{-1} \mathbf{x}_i}{\sum_{i \neq k} \xi_i \|\mathbf{x}_k - \mathbf{x}_i\|^{-1}}$,
          $\lambda = \frac{\xi_k}{\|\nabla D_1(\mathbf{x}_k)\|}$, $\mathbf{y}_{(p+1)} = (1 - \lambda)\tilde{\mathbf{T}}(\mathbf{x}_k) + \lambda \mathbf{x}_k$.
          8. $p \leftarrow p + 1$.
        end if
      else
        if $\|\nabla D_q(\mathbf{x}_k)\| = 0$ then
          9. $\mathbf{M} = \mathbf{x}_k$. Break.
        else
          10. Set $w = 0$,
          $\lambda_w = \min \left\{ \frac{1}{q} \xi_k^{-\frac{1}{q-1}} \|\nabla D_q(\mathbf{x}_k)\|^{\frac{2-q}{q-1}}, 1 \right\}$.
          while $C_q(\mathbf{x}_k - \lambda_w \nabla D_q(\mathbf{x}_k)) \geqslant C_q(\mathbf{x}_k)$ do
            11. $\lambda_{w+1} = \rho \lambda_w$. $w \leftarrow w + 1$.
          end while
          12. $\mathbf{y}_{(p+1)} = \mathbf{x}_k - \lambda_w \nabla D_q(\mathbf{x}_k)$. $p \leftarrow p + 1$.
        end if
      end if
    end if
    if $\|\mathbf{y}_{(p+1)} - \mathbf{y}_{(p)}\| / \|\mathbf{y}_{(p)}\| \leqslant Tol$ then
      13. $\mathbf{M} = \mathbf{y}_{(p+1)}$. Break.
    end if
  end while
Ensure: The minimum point $\mathbf{M}$.
```

B Proofs

B.1 Proof of Theorem 3

To prove this theorem, we need the following lemma:

Lemma 14 ([Weiszfeld, 1937; Cooper, 1968; Chen, 1984; Aftab *et al.*, 2015])**.** *If $a_i > 0$ and $b_i > 0$, $0 < q < n$ and $\sum_{i=1}^m a_i^{q-n} b_i^n < \sum_{i=1}^m a_i^q$, then $\sum_{i=1}^m b_i^q \leqslant \sum_{i=1}^m a_i^q$ and the equality holds only when $a_i = b_i, \forall i$.*

Proof. Consider the following function $g(t)$:

$$g(t) = \sum_{i=1}^m a_i^{q-t} b_i^t, 0 \leqslant t \leqslant n. \tag{29}$$

The second derivative of g with respect to t is:

$$g''(t) = \sum_{i=1}^m a_i^{q-t} b_i^t (\log a_i - \log b_i)^2. \tag{30}$$

Since all the $a_i, b_i > 0$, then $g''(t) > 0$ and $g(t)$ is a strictly convex function unless $a_i = b_i, \forall i$. If $g(t)$ is a strictly convex function, then $g(n) < g(0)$ implies $g(q) < g(0)$. Thus the lemma is proven. □

Lemma 14 reveals the relation between the q-th power ($1 \leqslant q < 2$) cost in (13) and the following weighted 2-nd power cost:

$$\tilde{C}_q(\mathbf{y}) = \sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2} \|\mathbf{y} - \mathbf{x}_i\|^2. \tag{31}$$

Proof. $\tilde{C}_q(\mathbf{y})$ in (31) is a strictly convex function on $\mathbf{y}$. By taking the gradient of $\tilde{C}_q(\mathbf{y})$ and setting it to zero, it yields:

$$\nabla \tilde{C}_q(\mathbf{y}) = \sum_{i=1}^m 2\eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2} (\mathbf{y} - \mathbf{x}_i) = \mathbf{0}. \tag{32}$$

Hence $\mathbf{T}_1(\mathbf{y}_{(p)})$ is the minimizer of $\tilde{C}_q(\mathbf{y})$. It yields $\tilde{C}_q(\mathbf{T}_1(\mathbf{y}_{(p)})) \leqslant \tilde{C}_q(\mathbf{y}_{(p)}) = C_q(\mathbf{y}_{(p)})$ with equality holds only when $\mathbf{T}_1(\mathbf{y}_{(p)}) = \mathbf{y}_{(p)}$.

If $\tilde{C}_q(\mathbf{T}_1(\mathbf{y}_{(p)})) < C_q(\mathbf{y}_{(p)})$, it means:

$$\begin{aligned} \sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2} \|\mathbf{T}_1(\mathbf{y}_{(p)}) - \mathbf{x}_i\|^2 &< \sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^q, \\ \sum_{i=1}^m \|\eta_i(\mathbf{y}_{(p)} - \mathbf{x}_i)\|^{q-2} \|\eta_i(\mathbf{T}_1(\mathbf{y}_{(p)}) - \mathbf{x}_i)\|^2 &< \sum_{i=1}^m \|\eta_i(\mathbf{y}_{(p)} - \mathbf{x}_i)\|^q. \end{aligned} \tag{33}$$

By setting $a_i = \|\eta_i(\mathbf{y}_{(p)} - \mathbf{x}_i)\|$, $b_i = \|\eta_i(\mathbf{T}_1(\mathbf{y}_{(p)}) - \mathbf{x}_i)\|$ for all i and using Lemma 14 ($n = 2$), it leads to:

$$C_q(\mathbf{T}_1(\mathbf{y}_{(p)})) = \sum_{i=1}^m \|\eta_i(\mathbf{T}_1(\mathbf{y}_{(p)}) - \mathbf{x}_i)\|^q < \sum_{i=1}^m \|\eta_i(\mathbf{y}_{(p)} - \mathbf{x}_i)\|^q = C_q(\mathbf{y}_{(p)}). \tag{34}$$

It proves Theorem 3. □

B.2 Proof of Corollary 4

Proof. With $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$ and (14), the following equivalence holds:

$$\mathbf{T}_1(\mathbf{y}_{(p)}) = \mathbf{y}_{(p)} \quad \Longleftrightarrow \quad \mathbf{0} = \sum_{i=1}^m \eta_i^q \|\mathbf{y}_{(p)} - \mathbf{x}_i\|^{q-2} (\mathbf{y}_{(p)} - \mathbf{x}_i) = \frac{1}{q} \nabla C_q(\mathbf{y}_{(p)}). \tag{35}$$

Since $C_q(\mathbf{y})$ is strictly convex, $\nabla C_q(\mathbf{y}_{(p)}) = \mathbf{0} \Leftrightarrow \mathbf{y}_{(p)} = \mathbf{M}$. □

B.3 Proof of Theorem 6

Proof. Let $\mathbf{x}_k + \lambda\mathbf{z}$ $(\lambda > 0, \|\mathbf{z}\| = 1)$ be a point displaced from $\mathbf{x}_k$ towards an arbitrary direction. Then the gradient of $C_q(\mathbf{x}_k + \lambda\mathbf{z})$ with respect to λ is:

$$\frac{\mathrm{d}C_q(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda} = \sum_{i \neq k} q\eta_i^q \|\mathbf{x}_k + \lambda\mathbf{z} - \mathbf{x}_i\|^{q-2}(\mathbf{x}_k + \lambda\mathbf{z} - \mathbf{x}_i)^\top \mathbf{z} + q\eta_k^q \lambda^{q-1}. \tag{36}$$

The limit of $\frac{\mathrm{d}}{\mathrm{d}\lambda}C_q(\mathbf{x}_k + \lambda\mathbf{z})$ when $\lambda \to 0$ is:

$$\frac{\mathrm{d}C_1(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = \sum_{i \neq k} \eta_i \|\mathbf{x}_k - \mathbf{x}_i\|^{-1}(\mathbf{x}_k - \mathbf{x}_i)^\top \mathbf{z} + \eta_k, \quad q = 1. \tag{37a}$$

$$\frac{\mathrm{d}C_q(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = \sum_{i \neq k} q\eta_i^q \|\mathbf{x}_k - \mathbf{x}_i\|^{q-2}(\mathbf{x}_k - \mathbf{x}_i)^\top \mathbf{z}, \quad 1 < q < 2. \tag{37b}$$

From Definition 5, (37a) and (37b) can be formulated as:

$$\frac{\mathrm{d}C_1(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = \nabla D_1(\mathbf{x}_k)^\top \mathbf{z} + \eta_k, \quad q = 1. \tag{38a}$$

$$\frac{\mathrm{d}C_q(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = \nabla D_q(\mathbf{x}_k)^\top \mathbf{z}, \quad 1 < q < 2. \tag{38b}$$

Thus the multiplicity η_k affects the gradient only when $q = 1$. By setting
$\mathbf{z} = -\frac{\nabla D_q(\mathbf{x}_k)}{\|\nabla D_q(\mathbf{x}_k)\|}$ in (38a) and (38b), we have:

$$\min_{\mathbf{z}} \frac{\mathrm{d}C_1(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = -\|\nabla D_1(\mathbf{x}_k)\| + \eta_k, \quad q = 1. \tag{39a}$$

$$\min_{\mathbf{z}} \frac{\mathrm{d}C_q(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} = -\|\nabla D_q(\mathbf{x}_k)\|, \quad 1 < q < 2. \tag{39b}$$

$$C_q(\mathbf{x}_k) \text{ is the minimum} \quad \Longleftrightarrow \quad \min_{\mathbf{z}} \frac{\mathrm{d}C_q(\mathbf{x}_k + \lambda\mathbf{z})}{\mathrm{d}\lambda}|_{\lambda=0} \geqslant 0, \ 1 \leqslant q < 2. \tag{39c}$$

Combining (39a), (39b) and (39c), one can find that the subgradient sets in (17) are equivalent to Definition 1. Thus Theorem 6 is proven. □

B.4 Proof of Theorem 7

Proof. This theorem indicates that a sufficiently small displacement towards the negative de-singularity subgradient $-\nabla D_q(\mathbf{x}_k)$ can reduce the q-th power cost. $C_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k))$ is continuous on $\lambda > 0$. It consists of two parts: the nonsingular part $D_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k))$ and the singular part $\eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q$. From the Taylor series expansion of $D_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k))$,

$$\begin{aligned}
&C_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k)) \\
=&D_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k)) + \eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q \\
=&D_q(\mathbf{x}_k) - \lambda\|\nabla D_q(\mathbf{x}_k)\|^2 + \frac{\lambda^2}{2}\nabla D_q(\mathbf{x}_k)^\top H(\mathbf{x}_k)\nabla D_q(\mathbf{x}_k) + o(\lambda^2) + \eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q,
\end{aligned} \tag{40}$$

where $H(\mathbf{x}_k)$ is the Hessian of $D_q(\mathbf{y})$ at $\mathbf{x}_k$. Besides, it is easy to find that $D_q(\mathbf{x}_k) = C_q(\mathbf{x}_k)$. Then (40) can be rearranged to:

$$C_q(\mathbf{x}_k - \lambda\nabla D_q(\mathbf{x}_k)) - C_q(\mathbf{x}_k) = -\lambda\|\nabla D_q(\mathbf{x}_k)\|^2 + \frac{\lambda^2}{2}G(\mathbf{x}_k) + o(\lambda^2) + \eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q, \tag{41}$$

where $G(\mathbf{x}_k) \triangleq \nabla D_q(\mathbf{x}_k)^\top H(\mathbf{x}_k)\nabla D_q(\mathbf{x}_k)$. Therefore, we need to find a λ such that the right side of (41) is negative.

When $\lambda \to 0$, the negative term $-\lambda\|\nabla D_q(\mathbf{x}_k)\|^2$ dominates the other terms on the right side, thus it is possible to make the right side negative. To specify, a λ should be found to satisfy the following inequality:

$$-\lambda\|\nabla D_q(\mathbf{x}_k)\|^2 + \frac{\lambda^2}{2}G(\mathbf{x}_k) + o(\lambda^2) + \eta_k^q \lambda^q \|\nabla D_q(\mathbf{x}_k)\|^q < 0. \tag{42}$$

Dividing both sides of (42) by λ yields:

$$\begin{aligned}
&-\|\nabla D_q(\mathbf{x}_k)\|^2 + \frac{\lambda}{2}G(\mathbf{x}_k) + o(\lambda) + \eta_k^q \lambda^{q-1} \|\nabla D_q(\mathbf{x}_k)\|^q < 0, \\
&\frac{\lambda}{2}G(\mathbf{x}_k) + o(\lambda) + \eta_k^q \lambda^{q-1} \|\nabla D_q(\mathbf{x}_k)\|^q < \|\nabla D_q(\mathbf{x}_k)\|^2.
\end{aligned} \tag{43}$$

Since $1 < q < 2$, the left side of (43) approaches zero when $\lambda \to 0$, while $\|\nabla D_q(\mathbf{x}_k)\|^2 > 0$. Therefore, there exists a $\lambda_* > 0$ such that for any $0 < \lambda \leqslant \lambda_*$, (43) holds. From (43) back to (41), the theorem is proven. □

B.5 Proof of Lemma 8

Proof. From Corollary 4, Theorem 3, Theorem 6 and Theorem 7, qPWAWS has the following decreasing property:

$$C_q(\mathbf{y}_{(0)}) > C_q(\mathbf{y}_{(1)}) > \cdots > C_q(\mathbf{y}_{(p)}) > \cdots > C_q(\mathbf{M}), \tag{44}$$

unless some $\mathbf{y}_{(p)}$ hits $\mathbf{M}$. In particular, if $\mathbf{y}_{(p)} = \mathbf{x}_k$ but $\mathbf{x}_k \neq \mathbf{M}$, then $C_q(\mathbf{y}_{(p)}) > C_q(\mathbf{y}_{(p+1)})$ and the subsequent iterates will never get back to $\mathbf{x}_k$, otherwise the decreasing property will be violated. Hence the sequence of iterates visits each $\mathbf{x}_k \neq \mathbf{M}$ at most once and will not get stuck.

From (20), if $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$, $\mathbf{T}_1(\mathbf{y}_{(p)})$ is a weighted sum of the data points $\{\mathbf{x}_i\}_{i=1}^m$ with positive weights that sum to one. Hence $\mathbf{y}_{(p+1)}{=}\mathbf{T}_1(\mathbf{y}_{(p)})$ lies in the convex hull of $\{\mathbf{x}_i\}_{i=1}^m$. Since $\{\mathbf{y}_{(p)}\}$ visits each $\mathbf{x}_k \neq \mathbf{M}$ at most once, $\mathbf{T}_2(\mathbf{y}_{(p)})$ is invoked at most finite times. Because $\mathbf{T}_2(\mathbf{y}_{(p)})$ cannot ensure that $\mathbf{y}_{(p+1)}$ lies in the convex hull, there are at most a finite set of iterates that do not lie in the convex hull.

Last, if $\mathbf{M} \in \{\mathbf{x}_i\}_{i=1}^m$, then $\mathbf{M}$ is trivially in the convex hull. If $\mathbf{M} \notin \{\mathbf{x}_i\}_{i=1}^m$, Corollary 4 and (35) indicate that $\mathbf{M}$ lies in the convex hull. □

B.6 Proof of Lemma 9

Proof. First, taking a difference between both sides of (21) leads to

$$\mathbf{T}_1(\mathbf{y}) - \mathbf{x}_k = \frac{\sum_{i\neq k} \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}(\mathbf{x}_i - \mathbf{x}_k)}{\sum_{i=1}^m \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}} \tag{45}$$

$$= \frac{\|\mathbf{y} - \mathbf{x}_k\|^{2-q} \cdot (\sum_{i\neq k} \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}(\mathbf{x}_i - \mathbf{x}_k))}{\eta_k^q + \|\mathbf{y} - \mathbf{x}_k\|^{2-q} \cdot (\sum_{i\neq k} \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2})}. \tag{46}$$

Then the limit of its L_2-norm is

$$\lim_{\mathbf{y}\to\mathbf{x}_k} \|\mathbf{T}_1(\mathbf{y}){-}\mathbf{x}_k\| {=} \frac{0 \cdot \| \sum_{i\neq k} \eta_i^q \|\mathbf{x}_k - \mathbf{x}_i\|^{q-2}(\mathbf{x}_i - \mathbf{x}_k)\|}{\eta_k^q + 0 \cdot (\sum_{i\neq k} \eta_i^q \|\mathbf{x}_k - \mathbf{x}_i\|^{q-2})} = 0. \tag{47}$$

Since $\eta_k^q \neq 0$, the above limit is well-defined and equals 0. It indicates that $\mathbf{T}_1(\mathbf{y}) \to \mathbf{x}_k$ when $\mathbf{y} \to \mathbf{x}_k$. □

B.7 Proof of Lemma 10

Proof. From (6), if $\mathbf{x}_k \neq \mathbf{M}$, then $\|\nabla D_q(\mathbf{x}_k)\| > 0$. This is the key condition for $\mathbf{T}_1$ to drive $\mathbf{y}$ away. For any sufficiently small $0 < \epsilon < 1$, since $\|\nabla D_q(\mathbf{y})\|$ is continuous around $\mathbf{x}_k$, there exists $\delta_1 > 0$ such that

$$\mathbf{y} \in B(\mathbf{x}_k, \delta_1) \Longrightarrow \|\nabla D_q(\mathbf{y})\| > \|\nabla D_q(\mathbf{x}_k)\| - \epsilon > 0. \tag{48}$$

Second, when $\mathbf{y} \to \mathbf{x}_k$, the weight of $\mathbf{x}_k$ in (14) will approach 1. In other words, there exists $\delta_2 > 0$ such that

$$\mathbf{y} \in B(\mathbf{x}_k, \delta_2) \Longrightarrow 1 - \epsilon < \frac{\eta_k^q \|\mathbf{y} - \mathbf{x}_k\|^{q-2}}{\sum_{i=1}^m \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}} < 1. \tag{49}$$

Third, to handle some remainders, define $\delta_3 > 0$ as follows:

$$\delta_3 = \left(\frac{(1-\epsilon)(\|\nabla D_q(\mathbf{x}_k)\| - \epsilon)}{q\eta_k^q(1+2\epsilon)} \right)^{1/(q-1)}, \quad 1 < q < 2. \tag{50}$$

$$\mathbf{y} \in B(\mathbf{x}_k, \delta_3) \Longrightarrow \frac{(1-\epsilon)(\|\nabla D_q(\mathbf{x}_k)\| - \epsilon)}{q\eta_k^q \|\mathbf{y} - \mathbf{x}_k\|^{q-1}} > 1 + 2\epsilon. \tag{51}$$

When $0 < \epsilon < 1$ is sufficiently small, $\delta_3 > 0$ is well defined.

Let $\delta_0 = \min\{\delta_1, \delta_2, \delta_3\}$ and $\mathbf{y} \in B(\mathbf{x}_k, \delta_0)$, then:

$$\mathbf{T}_1(\mathbf{y}) - \mathbf{x}_k = \frac{\sum_{i=1}^m \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}(\mathbf{x}_i - \mathbf{y})}{\sum_{i=1}^m \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}} + \mathbf{y} - \mathbf{x}_k$$

$$= \frac{-\nabla D_q(\mathbf{y})/q}{\sum_{i=1}^m \eta_i^q \|\mathbf{y}{-}\mathbf{x}_i\|^{q-2}} + \left(\frac{\eta_k^q \|\mathbf{y}{-}\mathbf{x}_k\|^{q-2}}{\sum_{i=1}^m \eta_i^q \|\mathbf{y}{-}\mathbf{x}_i\|^{q-2}} - 1 \right) (\mathbf{x}_k{-}\mathbf{y}). \tag{52}$$

Therefore

$$\|\mathbf{T}_1(\mathbf{y}) - \mathbf{x}_k\| > \frac{\|\nabla D_q(\mathbf{y})/q\|}{\sum_{i=1}^m \eta_i^q \|\mathbf{y} - \mathbf{x}_i\|^{q-2}} - \epsilon \|\mathbf{x}_k - \mathbf{y}\|$$

$$
\begin{aligned}
&> \frac{(1-\epsilon)\|\nabla D_q(\mathbf{y})/q\|}{\eta_k^q\|\mathbf{y}-\mathbf{x}_k\|^{q-2}} - \epsilon\|\mathbf{x}_k - \mathbf{y}\| \\
&> \frac{(1-\epsilon)(\|\nabla D_q(\mathbf{x}_k)\| - \epsilon)}{q\eta_k^q\|\mathbf{y}-\mathbf{x}_k\|^{q-2}} - \epsilon\|\mathbf{x}_k - \mathbf{y}\| \\
&> (1+2\epsilon)\|\mathbf{x}_k - \mathbf{y}\| - \epsilon\|\mathbf{x}_k - \mathbf{y}\| \\
&= (1+\epsilon)\|\mathbf{x}_k - \mathbf{y}\|,
\end{aligned} \tag{53}
$$

where the first inequality is based on the triangle inequality and (49); The second inequality is based on the left inequality of (49); The third and the fourth inequalities are based on (48) and (51), respectively.

Therefore, $\|\mathbf{T}_1(\mathbf{y}) - \mathbf{x}_k\| > (1+\epsilon)\|\mathbf{y} - \mathbf{x}_k\|$. If $\mathbf{T}_1(\mathbf{y}) \in B(\mathbf{x}_k, \delta_0)$, then $\|\mathbf{T}_1^2(\mathbf{y}) - \mathbf{x}_k\| > (1+\epsilon)\|\mathbf{T}_1(\mathbf{y}) - \mathbf{x}_k\| > (1+\epsilon)^2\|\mathbf{y} - \mathbf{x}_k\|$. As long as the current iterate lies in $B(\mathbf{x}_k, \delta_0)$, $\mathbf{T}_1$ will keep on driving it out of $B(\mathbf{x}_k, \delta_0)$. Thus there exists some s such that $\mathbf{T}_1^{s-1}(\mathbf{y}) \in B(\mathbf{x}_k, \delta_0)$ and $\mathbf{T}_1^{s}(\mathbf{y}) \notin B(\mathbf{x}_k, \delta_0)$ (see Figure 5). □

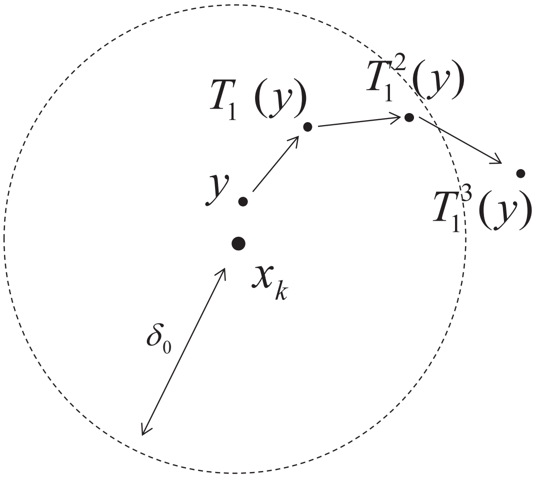


Figure 5: Once $\mathbf{y}$ gets into $B(\mathbf{x}_k, \delta_0)$ where $\mathbf{x}_k \neq \mathbf{M}$, $\mathbf{T}_1$ will eventually drive it out of $B(\mathbf{x}_k, \delta_0)$.

B.8 Proof of Theorem 11

Proof. We can assume that $\mathbf{y}_{(p)}$ differs from $\mathbf{M}$ for all p. From Lemma 8, since at most a finite set of iterates do not lie in the convex hull of $\{\mathbf{x}_i\}_{i=1}^m$, the whole sequence $\{\mathbf{y}_{(p)}\}$ is a compact set in $\mathbb{R}^d$. Moreover, by omitting at most a finite number of iterates, we can assume that $\{\mathbf{y}_{(p)}\} \bigcap \{\mathbf{x}_i\}_{i=1}^m = \varnothing$. By the Bolzano-Weierstrass Theorem, there exists a subsequence $\{\mathbf{y}_{(p_v)}\}$ such that $\lim_{v\to\infty} \mathbf{y}_{(p_v)} = \mathbf{y}_*$ for some $\mathbf{y}_* \in \mathbb{R}^d$. Since the extended operator $\mathbf{T}_1$ is continuous,

$$\lim_{v\to\infty} \mathbf{T}_1(\mathbf{y}_{(p_v)}) = \mathbf{T}_1(\mathbf{y}_*). \tag{54}$$

According to the decreasing property of qPWAWS (44), the sequence $C_q(\mathbf{y}_{(p)})$ is bounded below and decreasing, thus it has a limit and any subsequence of $C_q(\mathbf{y}_{(p)})$ should have the same limit. In particular, $C_q(\mathbf{y}_{(p_v)})$ and $C_q(\mathbf{T}_1(\mathbf{y}_{(p_v)}))$ are two subsequences of $C_q(\mathbf{y}_{(p)})$. Hence

$$\lim_{v\to\infty} C_q(\mathbf{T}_1(\mathbf{y}_{(p_v)})) = \lim_{v\to\infty} C_q(\mathbf{y}_{(p_v)}). \tag{55}$$

Since C_q is continuous, (54) and (55) indicate

$$C_q(\mathbf{T}_1(\mathbf{y}_*)) = C_q(\mathbf{y}_*). \tag{56}$$

If $\mathbf{y}_* \notin \{\mathbf{x}_i\}_{i=1}^m$, then Theorem 3 and (56) indicate $\mathbf{y}_* = \mathbf{T}_1(\mathbf{y}_*)$. By Corollary 4, $\mathbf{y}_* = \mathbf{M}$. If $\mathbf{y}_* \in \{\mathbf{x}_i\}_{i=1}^m$, then (56) trivially holds from (22). To summarize, $\mathbf{y}_* \in \{\mathbf{x}_i\}_{i=1}^m \bigcup \{\mathbf{M}\}$ and only the points in the finite set $\{\mathbf{x}_i\}_{i=1}^m \bigcup \{\mathbf{M}\}$ satisfy (56) and constitute the fixed points of $\mathbf{T}_1$.

The next step is to prove that if $\mathbf{y}_* = \mathbf{x}_k$ for some k, then $\mathbf{x}_k = \mathbf{M}$. If not, we invoke Lemma 10 to induce a contradiction. Since $\lim_{v\to\infty} \mathbf{y}_{(p_v)} = \mathbf{x}_k$, once $\mathbf{y}_{(p_v)}$ gets into $B(\mathbf{x}_k, \delta_0)$, it will be driven out by $\mathbf{T}_1$. Thus for each $\mathbf{y}_{(p_v)} \in B(\mathbf{x}_k, \delta_0)$, there exists a $\mathbf{y}_{(p_u)} \in B(\mathbf{x}_k, \delta_0)$ and a $\mathbf{T}_1(\mathbf{y}_{(p_u)}) \notin B(\mathbf{x}_k, \delta_0)$. In other words, $\mathbf{y}_{(p_u)}$ is the iterate that is going to be driven out of $B(\mathbf{x}_k, \delta_0)$. Since $\mathbf{y}_{(p_v)}$ is an infinite sequence converging to $\mathbf{x}_k$, $\mathbf{y}_{(p_u)}$ and $\mathbf{T}_1(\mathbf{y}_{(p_u)})$ are also infinite sequences. By the

Bolzano-Weierstrass Theorem, $\mathbf{y}_{(p_u)}$ has a subsequence that converges to some $\mathbf{y}_{*1}$. We still denote this subsequence by $\mathbf{y}_{(p_u)}$. Then $\mathbf{T}_1(\mathbf{y}_{(p_u)})$ also has a subsequence that converges to some $\mathbf{y}_{*2}$ and the subsequence can still be denoted by $\mathbf{T}_1(\mathbf{y}_{(p_u)})$. Note that $\mathbf{y}_{(p_u)}$ and $\mathbf{T}_1(\mathbf{y}_{(p_u)})$ are not necessarily subsequences of $\mathbf{y}_{(p_v)}$. Then

$$\lim_{u\to\infty}\mathbf{y}_{(p_u)} = \mathbf{y}_{*1}, \quad \lim_{u\to\infty}\mathbf{T}_1(\mathbf{y}_{(p_u)}) = \mathbf{y}_{*2}. \tag{57}$$

$$\|\mathbf{y}_{*1} - \mathbf{x}_k\| \leqslant \delta_0, \quad \|\mathbf{y}_{*2} - \mathbf{x}_k\| \geqslant \delta_0. \tag{58}$$

Similar to the demonstrations of (54),(55) and (56), the accumulation point $\mathbf{y}_{*1}$ is also a fixed point of $\mathbf{T}_1$:

$$\lim_{u\to\infty}\mathbf{T}_1(\mathbf{y}_{(p_u)}) = \mathbf{T}_1(\lim_{u\to\infty}\mathbf{y}_{(p_u)}) = \mathbf{T}_1(\mathbf{y}_{*1}) = \mathbf{y}_{*1}. \tag{59}$$

From (57), (58) and (59),

$$\mathbf{y}_{*1} = \mathbf{y}_{*2}, \quad \|\mathbf{y}_{*1} - \mathbf{x}_k\| = \delta_0. \tag{60}$$

The process of inducing $\mathbf{y}_{*1} = \mathbf{y}_{*2}$ can be shown as Figure 6.

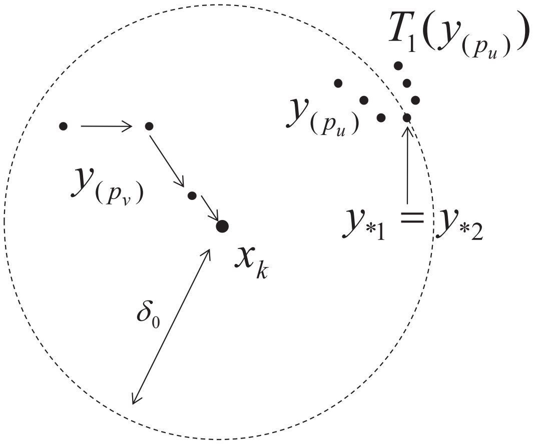


Figure 6: The process of inducing $\mathbf{y}_{*1} = \mathbf{y}_{*2}$. $\mathbf{y}_{(p_v)}$ is a subsequence that converges to $\mathbf{x}_k \neq \mathbf{M}$. By Lemma 10, each $\mathbf{y}_{(p_v)}$ induces a $\mathbf{y}_{(p_u)} \in B(\mathbf{x}_k, \delta_0)$ and a $\mathbf{T}_1(\mathbf{y}_{(p_u)}) \notin B(\mathbf{x}_k, \delta_0)$. Then the subsequences $\mathbf{y}_{(p_u)}$ and $\mathbf{T}_1(\mathbf{y}_{(p_u)})$ have the same accumulation point $\mathbf{y}_{*1} = \mathbf{y}_{*2}$ at the boundary of $B(\mathbf{x}_k, \delta_0)$.

For any $0 < \delta < \delta_0$, we can apply the same method to obtain a distinct fixed point $\mathbf{y}_{*\delta}$ such that $\mathbf{T}_1(\mathbf{y}_{*\delta}) = \mathbf{y}_{*\delta}$ and $\|\mathbf{y}_{*\delta} - \mathbf{x}_k\| = \delta$. Then there are infinite fixed points $\{\mathbf{y}_{*\delta}\}$, which is contradictory to the finite set of fixed points $\{\mathbf{x}_i\}_{i=1}^m \bigcup\{\mathbf{M}\}$. Therefore, $\mathbf{y}_* = \mathbf{x}_k = \mathbf{M}$.

The last step is to prove that the whole sequence $\{\mathbf{y}_{(p)}\}$ converges to $\mathbf{y}_*$. From the above illustrations, the accumulation point $\mathbf{y}_* = \mathbf{M}$. If there is another accumulation point $\tilde{\mathbf{y}} \neq \mathbf{y}_*$, then $\tilde{\mathbf{y}} \in \{\mathbf{x}_i\}_{i=1}^m$. Without loss of generality, suppose $\tilde{\mathbf{y}} = \mathbf{x}_k \neq \mathbf{M}$, then the above method can be repeated to induce an infinite set of fixed points $\{\mathbf{y}_{*\delta}\}$, which leads to a contradiction. Hence, there is only one accumulation point $\mathbf{y}_* = \mathbf{M}$ for the whole sequence $\{\mathbf{y}_{(p)}\}$. Thus $\{\mathbf{y}_{(p)}\}$ and any subsequences converge to $\mathbf{y}_* = \mathbf{M}$. □

B.9 Proof of Lemma 12

Proof. It is straightforward to check from (16) that $\nabla D_q(\mathbf{y})$ is analytic in some neighborhood $B(\mathbf{x}_k, \delta)$ of $\mathbf{x}_k$ such that $B(\mathbf{x}_k, \delta)\bigcap\{\mathbf{x}_i\}_{i\neq k} = \varnothing$, since the singular component has been excluded from $\nabla D_q(\mathbf{y})$. Furthermore, $\|\nabla D_q(\mathbf{y})\|^2 = \nabla D_q(\mathbf{y})^\top \nabla D_q(\mathbf{y})$ is also analytic in this neighborhood $B(\mathbf{x}_k, \delta)$. Thus we can adopt the second-order Taylor series expansion of $\|\nabla D_q(\mathbf{y})\|^2$ for $\mathbf{y}_{(p)} \in B(\mathbf{x}_k, \delta)$ at $\mathbf{x}_k$:

$$\|\nabla D_q(\mathbf{y}_{(p)})\|^2 =\|\nabla D_q(\mathbf{x}_k)\|^2 + 2\nabla D_q(\mathbf{x}_k)^\top H(\mathbf{x}_k)(\mathbf{y}_{(p)} - \mathbf{x}_k)+\frac{1}{2}(\mathbf{y}_{(p)}-\mathbf{x}_k)^\top J(\mathbf{x}_k)(\mathbf{y}_{(p)}-\mathbf{x}_k)+o(\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2), \tag{61}$$

where $H(\mathbf{x}_k)$ and $J(\mathbf{x}_k)$ are the Hessians of $D_q(\mathbf{y})$ and $\|\nabla D_q(\mathbf{y})\|^2$ at $\mathbf{x}_k$, respectively.

Theorem 6 indicates that $\nabla D_q(\mathbf{x}_k) = \mathbf{0}$ when $1 < q < 2$, thus (61) can be further simplified as

$$\|\nabla D_q(\mathbf{y}_{(p)})\|^2 =\frac{1}{2}(\mathbf{y}_{(p)}-\mathbf{x}_k)^\top J(\mathbf{x}_k)(\mathbf{y}_{(p)}-\mathbf{x}_k)+o(\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2) \leqslant \frac{\vartheta_J}{2}\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2+o(\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2), \tag{62}$$

where ϑ_J denotes the largest eigenvalue of $J(\mathbf{x}_k)$. Since $\|\nabla D_q(\mathbf{y}_{(p)})\|^2 > 0$ when $\mathbf{y}_{(p)} \neq \mathbf{x}_k$, (62) implies that

$$\frac{\vartheta_J}{2}\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2+o(\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2) \geqslant \|\nabla D_q(\mathbf{y}_{(p)})\|^2 > 0 \quad \Longrightarrow \quad \frac{\vartheta_J}{2}+o(1)>0 \quad \Longrightarrow \quad \vartheta_J \geqslant 0. \tag{63}$$

It means that the inequality in (62) really holds without contradiction.

Next, dividing the leftmost side and the rightmost side of (62) by $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2$ leads to

$$\frac{\|\nabla D_q(\mathbf{y}_{(p)})\|^2}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2} \leqslant \frac{\vartheta_J}{2}+o(1) \quad \Longrightarrow \quad \lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k} \frac{\|\nabla D_q(\mathbf{y}_{(p)})\|^2}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2} \leqslant \frac{\vartheta_J}{2}. \tag{64}$$

Since the square root operator $\sqrt{\cdot}$ is continuous in the interval $(0,+\infty)$ and right continuous at 0, we can take $\sqrt{\cdot}$ inside the limit of (64) and get

$$\lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k} \frac{\|\nabla D_q(\mathbf{y}_{(p)})\|}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|} = \sqrt{\lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k} \frac{\|\nabla D_q(\mathbf{y}_{(p)})\|^2}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^2}} \leqslant \sqrt{\frac{\vartheta_J}{2}}. \tag{65}$$

Let $\zeta \triangleq \sqrt{\frac{\vartheta_J}{2}}$ and the proof is finished. □

B.10 Proof of Theorem 13

Proof. Since $\mathbf{y}_{(p)} \to \mathbf{x}_k$ and the data points are distinct, we can assume that $\mathbf{y}_{(p)} \notin \{\mathbf{x}_i\}_{i=1}^m$, $\forall p \geqslant P$ for some sufficiently large P. Therefore, $\mathbf{y}_{(p+1)} = \mathbf{T}_1(\mathbf{y}_{(p)})$, $\forall p \geqslant P$. We begin with an important equation:

$$\mathbf{y}_{(p+1)}-\mathbf{x}_k = \frac{\sum_{i\neq k} \eta_i^q \|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2}(\mathbf{x}_i-\mathbf{x}_k)}{\eta_k^q\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{q-2}+\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2}}. \tag{66}$$

The key technique is to eliminate the singular term $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{q-2}$ in the denominator of the right side of (66) and construct the rate of convergence simultaneously.

In the $q=1$ case, we divide both sides of (66) by the nonzero scalar $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|$:

$$\frac{\mathbf{y}_{(p+1)}-\mathbf{x}_k}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|} = \frac{\sum_{i\neq k}\eta_i\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{-1}(\mathbf{x}_i-\mathbf{x}_k)}{\eta_k+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|\cdot(\sum_{i\neq k}\eta_i\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{-1})}. \tag{67}$$

Taking L_2-norm $\|\cdot\|$ on both sides of (67) and letting $\mathbf{y}_{(p)}\to\mathbf{x}_k$ lead to

$$\begin{aligned} \lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k}\frac{\|\mathbf{y}_{(p+1)}-\mathbf{x}_k\|}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|} &= \lim_{\mathbf{y}_{(p)}\to\mathbf{x}_k}\frac{\|\sum_{i\neq k}\eta_i\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{-1}(\mathbf{x}_i-\mathbf{x}_k)\|}{\eta_k+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|\cdot(\sum_{i\neq k}\eta_i\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{-1})} \\ &= \frac{\|-\nabla D_1(\mathbf{x}_k)\|}{\eta_k+0\cdot(\sum_{i\neq k}\eta_i\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{-1})} = \frac{\|\nabla D_1(\mathbf{x}_k)\|}{\eta_k}. \end{aligned} \tag{68}$$

Since $\eta_k > 0$, the above limit is well-defined. Based on Theorem 6, the convergence is sublinear, linear or superlinear when $\|\nabla D_1(\mathbf{x}_k)\| = \eta_k$, $0 < \|\nabla D_1(\mathbf{x}_k)\| < \eta_k$ or $\|\nabla D_1(\mathbf{x}_k)\| = 0$, respectively.

In the $1 < q < 2$ case, it is a little subtle and we take two steps to eliminate the singular term $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{q-2}$ in the denominator of (66). First, we divide both sides of (66) by the nonzero scalar $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|$:

$$\frac{\mathbf{y}_{(p+1)}-\mathbf{x}_k}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|} = \frac{\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2}(\mathbf{x}_i-\mathbf{x}_k)}{\eta_k^q\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{q-1}+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}. \tag{69}$$

Second, we multiply both the numerator and the denominator of the right side of (69) by $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{1-q}$:

$$\begin{aligned} &\frac{\mathbf{y}_{(p+1)}-\mathbf{x}_k}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|} \\ =&\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{1-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2}(\mathbf{x}_i-\mathbf{x}_k))}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})} \\ =&\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{1-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})(\mathbf{y}_{(p)}-\mathbf{x}_k)}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})} - \frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{1-q}\cdot\nabla D_q(\mathbf{y}_{(p)})/q}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}. \end{aligned} \tag{70}$$

Taking L_2-norm $\|\cdot\|$ on the leftmost side and the rightmost side of (70) leads to

$$\begin{aligned}
&\frac{\|\mathbf{y}_{(p+1)}-\mathbf{x}_k\|}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|}\\
\leqslant&\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}+\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{1-q}\cdot\|\nabla D_q(\mathbf{y}_{(p)})\|/q}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}\\
=&\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}+\frac{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot\left(\frac{\|\nabla D_q(\mathbf{y}_{(p)})\|}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|}\right)/q}{\eta_k^q+\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{y}_{(p)}-\mathbf{x}_i\|^{q-2})}.
\end{aligned}\tag{71}$$

Because $1<q<2$, $0<2-q<1$ and $\|\mathbf{y}_{(p)}-\mathbf{x}_k\|^{2-q}$ is nonsingular when $\mathbf{y}_{(p)}\rightarrow\mathbf{x}_k$. Then we can adopt Lemma 12 to dominate the rightmost side of (71):

$$\lim_{\mathbf{y}_{(p)}\rightarrow\mathbf{x}_k}\frac{\|\mathbf{y}_{(p+1)}-\mathbf{x}_k\|}{\|\mathbf{y}_{(p)}-\mathbf{x}_k\|}\leqslant\frac{0\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{x}_k-\mathbf{x}_i\|^{q-2})+0\cdot\zeta/q}{\eta_k^q+0\cdot(\sum_{i\neq k}\eta_i^q\|\mathbf{x}_k-\mathbf{x}_i\|^{q-2})}=0,\tag{72}$$

which shows a superlinear convergence for $1<q<2$. □