Robust Iterative Quantization for Efficient ℓ_p-norm Similarity Search*

Yuchen Guo†, Guiguang Ding†, Jungong Han‡, Xiaoming Jin†
†School of Software, Tsinghua University, Beijing 100084, China
‡Northumbria University, Newcastle, NE1 8ST, UK
yuchen.w.guo@gmail.com, {dinggg,xmjin}@tsinghua.edu.cn,jungong.han@northumbria.ac.uk

Abstract

Iterative Quantization (ITQ) is one of the most successful hashing based nearest-neighbor search methods for large-scale information retrieval in the past a few years due to its simplicity and superior performance. However, the performance of this algorithm degrades significantly when dealing with noisy data. Additionally, it can barely facilitate a wide range of applications as the distortion measurement only limits to ℓ_2 *norm*. In this paper, we propose an ITQ+ algorithm, aiming to enhance both robustness and generalization of the original ITQ algorithm. Specifically, a $\ell_{p,q}$-norm loss function is proposed to conduct the ℓ_p-norm similarity search, rather than a ℓ_2 *norm* search. Despite the fact that changing the loss function to $\ell_{p,q}$-norm makes our algorithm more robust and generic, it brings us a challenge that minimizes the obtained *orthogonality constrained* $\ell_{p,q}$*-norm function*, which is non-smooth and non-convex. To solve this problem, we propose a novel and efficient optimization scheme. Extensive experiments on benchmark datasets demonstrate that ITQ+ is overwhelmingly better than the original ITQ algorithm, especially when searching similarity in noisy data.

1 Introduction

Similarity search is of great importance to applications in various areas, such as data mining [Altman, 1992], machine learning [Cheng *et al.*, 2015], information retrieval [Furnas *et al.*, 1988], and etc. Formally, given a database $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, similarity search is to find those instances that most closely resemble a query $\mathbf{x}_q$ based on a similarity or distance measure $d(\mathbf{x}_q, \mathbf{x}_i)$, e.g., Euclidean distance. The small-database case is well solved, however, the cost of computing the distance between the query and all database instances becomes prohibitively high in the case that the reference database is huge. To address this problem, hashing based methods [Gionis *et al.*, 1999; Andoni and Indyk, 2006] are proposed, which transform the data from a real-value representation into a sequence of binary bits. The binary representation and those bit-wise operations make computing the Hamming distance between binary codes extremely efficient in a modern CPU architecture [He *et al.*, 2013], therefore enabling a fast nearest neighbor search. Such a procedure can be considered as a means for transforming high-dimensional feature vectors to a low-dimensional Hamming space, while retaining the original similarity structure of data as much as possible. In this way, the original distance $d(\mathbf{x}_q, \mathbf{x}_i)$, thanks to the similarity preservation, can be effectively approximated by the Hamming distance $d_h(\mathbf{b}_q, \mathbf{b}_i)$ between binary codes.

Locality Sensitive Hashing (LSH) [Gionis *et al.*, 1999], as the seminal work, adopts random split to generate binary codes. Although enjoying asymptotic theoretical benefits, LSH needs long codes for a good performance because of its *data-independent* [Zhang *et al.*, 2010] property. To obtain compact binary codes, many machine learning techniques are exploited [Wang *et al.*, 2014]. Among all, Iterative Quantization (ITQ) [Gong *et al.*, 2013] that aims to minimize the distortion between binary codes and the original features, has shown the state-of-the-art performance for learning ℓ_2-norm similarity-preserving binary codes, and thus ITQ has been utilized in information retrieval, image classification, and etc.

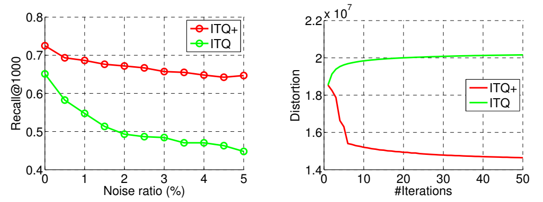
1.1 Problem Statement

The extraordinary performance and a large number of the follow-up works [Ge *et al.*, 2014; Zhang *et al.*, 2014; Kong and Li, 2012; Xu *et al.*, 2013] motivate us to closely investigate the ITQ algorithm. Specifically, the objective of ITQ is to learn the binary representation and an orthogonal rotation matrix $\mathbf{R} \in \mathbb{R}^{k \times k}$ which minimizes the distortion as follows,

$$\min_{\mathbf{b}_i, \mathbf{R}} \mathcal{O}_{\text{ITQ}} = \sum\nolimits_{i=1}^{n} \|\mathbf{b}_i - \mathbf{x}_i\mathbf{R}\|_2^2, \text{ s.t. } \mathbf{R}\mathbf{R}' = \mathbf{I}, \qquad (1)$$

where $\mathbf{b}_i \in \{-1, 1\}^k$ is the binary representation of $\mathbf{x}_i$, k is the length of binary codes and $\mathbf{I}$ is the identity matrix. Without loss of generality, hereafter we assume the data is zero-centered, i.e., $\sum_i \mathbf{x}_i = \mathbf{0}$. A close look at the objective function reveals that a squared ℓ_2 loss is applied to measure the distortion. But unfortunately, this sort of distance measurement comes with certain vulnerabilities. For instance, there are noise and outliers in real-world datasets but the squared

*Corresponding author: Guiguang Ding. This research was supported by the National Natural Science Foundation of China (Grant No.61271394 and 61571269), the National Basic Research Project of China (Grant No. 2015CB352300), and the Royal Society Newton Mobility Grant (IE150997).



(a) SIFT1M, 64 bits, $q = 1$ (b) SIFT1M, 64 bits, $p = 1$

Figure 1: ITQ does not perform well with the noisy data (left) and can not deal with some other similarity measures (right).

loss is sensitive to them because their large distortion may dominate the sum of the squared loss [Huang *et al.*, 2013; Pan *et al.*, 2014; Jiang *et al.*, 2015], which may markedly degrade the quality of binary codes. Solving this problem becomes important when we need to search nearest neighbors for data in the wild, such as Flickr images and Youtube videos, as the noises are commonly existed. To verify our observation, we carried out an experiment based on SIFT1M dataset, in which the noise is manually added into the data and we plot the similarity search performance of ITQ w.r.t. the noise ratio. As can be seen from Figure 1(a), the performance of ITQ degrades significantly even with only 1% of noise. Additionally, ITQ works pretty well for ℓ_2-norm similarity search, i.e., $d(\mathbf{x}_q, \mathbf{x}_i) = \|\mathbf{x}_q - \mathbf{x}_i\|_2$ but not for the other measurements like a Manhattan distance $d(\mathbf{x}_q, \mathbf{x}_i) = \|\mathbf{x}_q - \mathbf{x}_i\|_1$ [Singh and Jain, 2010]. In practice, the preferred measure means may need to be defined by users depending on the application. This indicates that a good similarity search algorithm should be generic enough to deal with different distance measurements. Again, to demonstrate this, we plot the ℓ_1-norm distortion w.r.t. the number of iterations of ITQ, as shown in Figure 1(b). It can be observed that the distortion keeps increasing with more iterations. This phenomenon reveals that the distance approximation gets worse such that more iterations might result in worse similarity search result.

1.2 Our Contributions

Aiming to address the two problems mentioned above, we intend to develop an improved ITQ algorithm, in which both robustness and generalization are enhanced by using a ℓ_p-norm distance. Recent studies have shown that the q-th order ($q < 2$, especially $q \leq 1$) of ℓ_2 loss, i.e., $\|\mathbf{b}_i - \mathbf{x}_i\mathbf{R}\|_2^q$, is more robust to the noise and outliers in data than the squared loss [Wright *et al.*, 2009; Wang *et al.*, 2013]. Based on the triangle inequality, preserving ℓ_p-norm distance can be achieved by minimizing the ℓ_p-norm distortion, i.e., $\|\mathbf{b}_i - \mathbf{x}_i\mathbf{R}\|_p$. Therefore, in this paper, we propose a $\ell_{p,q}$-norm loss function for learning robust ℓ_p-norm similarity-preserving binary codes, termed as ITQ+. In Figure 1, we exhibit the effectiveness of ITQ+, in contrast to the original ITQ algorithm. In summary, the major contributions of this paper are two folds.

- We propose a new $\ell_{p,q}$-norm loss for binary-code learning. It is robust to noise by using a small q and supports ℓ_p-norm ($p \leq 2$) similarity search with the ℓ_p-norm loss.
- To minimize the orthogonality constrained $\ell_{p,q}$-norm function, a novel and efficient iterative optimization algorithm is proposed and its convergence property is theoretically investigated. To the best of our knowledge, it is the first work that provides the theoretical solution to this challenging non-smooth and non-convex problem.

In addition, it is worthwhile to highlight two important properties of the ITQ+ algorithm from the application perspective.

- ITQ+ is resistant to the noise, enabling us to search similarity in wild data. Such an algorithm is favorably demanded by the applications like Internet image retrieval.
- Our algorithm is more generic in the sense that multiple distortion measurements are implemented in one framework, allowing us to facilitate a wide range of applications in which different measurements may be requested.

2 The Proposed Method

2.1 Objective Function

As we mentioned before, the algorithm can be enhanced if we make the squared loss less sensitive to the noise. In this paper, we adopt a widely used method that replaces the squared loss by the the q-th order loss. It has been shown in the literatures [Wright *et al.*, 2009; Huang *et al.*, 2013; Wang *et al.*, 2013] that the loss function is more robust to noise and outliers in data in case of $q < 2$, especially $q \leq 1$. Inspired by the idea, we reformulate the objective function of ITQ in Eq. (1) from the squared loss to the q-th order loss as

$$\min_{\mathbf{b}_i,\mathbf{R}} \mathcal{O}_{\text{ITQ}} = \sum\nolimits_{i=1}^{n} \|\mathbf{b}_i - \mathbf{x}_i\mathbf{R}\|_2^q, \text{ s.t. } \mathbf{R}\mathbf{R}' = \mathbf{I}. \qquad (2)$$

Denote $Q(\mathbf{x})$ as the quantization result of $\mathbf{x}$. Based on the vector norm property, we obtain two inequalities as follows,

$$\begin{aligned}|\|\mathbf{x}-\mathbf{y}\|_p - \|Q(\mathbf{x}) - Q(\mathbf{y})\|_p| &\leq K_1\|\mathbf{x}-\mathbf{y}-Q(\mathbf{x})+Q(\mathbf{y})\|_p \\ &\leq K_2(\|\mathbf{x}-Q(\mathbf{x})\|_p + \|\mathbf{y}-Q(\mathbf{y})\|_p),\end{aligned}$$

where $\|\mathbf{x}\|_p = (\sum_j |x_j|^p)^{\frac{1}{p}}$ denotes the ℓ_p norm of a vector. The above inequalities are based on the triangle inequality. From these inequalities, we can observe that with smaller distortion ($\|\mathbf{x} - Q(\mathbf{x})\|_p$), the distance between $Q(\mathbf{x})$ and $Q(\mathbf{y})$ can accurately approximate the distance between $\mathbf{x}$ and $\mathbf{y}$ [Ge *et al.*, 2014; Zhang *et al.*, 2014]. Furthermore, in the extreme situation where the distortion is 0, we have $\|Q(\mathbf{x}) - Q(\mathbf{y})\|_p = \|\mathbf{x} - \mathbf{y}\|_p$. Fortunately, the binary codes used in the hashing algorithm are exactly the quantization result of original features. Therefore, to learn ℓ_p-norm similarity-preserving binary codes, we need to minimize the ℓ_p-norm distortion. Here, ITQ can be considered as a special case ($p = 2$) of our scheme . To clarify it, we can rewrite the objective function in Eq. (2) from the ℓ_2-norm loss to the ℓ_p-norm loss, which leads to the objective function of ITQ+:

$$\min_{\mathbf{b}_i,\mathbf{R}} \mathcal{O}_{\text{ITQ+}} = \sum\nolimits_{i=1}^{n} \|\mathbf{b}_i - \mathbf{x}_i\mathbf{R}\|_p^q, \text{ s.t. } \mathbf{R}\mathbf{R}' = \mathbf{I}. \qquad (3)$$

2.2 Learning Algorithm

Changing to a $\ell_{p,q}$-norm loss is not difficult, but minimizing obtained orthogonality constrained $\ell_{p,q}$-norm is non-trivial since it becomes a non-smooth and non-convex optimization problem when $p \leq 1$ or $q \leq 1$. Solving this problem is much more difficult than minimizing the $\ell_{2,2}$-norm in ITQ, for which many solutions are available [Gong *et al.*, 2013]. To solve our problem, we propose an efficient iterative optimization algorithm, which can be decomposed into two parts:

Fix $\mathbf{R}$ and update $\mathbf{b}_i$. Similar to ITQ, the problem that arises in Eq. (3) can be optimized w.r.t. every element in $\mathbf{b}_i$ individually, which reduces the optimization problem below

$$\min_{b_{ij}} \mathcal{O}_{ij} = |b_{ij} - \mathbf{x}_i \mathbf{R}_{*j}|, \text{ s.t. } b_{ij} \in \{-1, 1\}. \tag{4}$$

The solution for the above problem can be written as follows:

$$b_{ij} \leftarrow \text{sign}(\mathbf{x}_i \mathbf{R}_{*j}). \tag{5}$$

Here, $\text{sign}(x) = 1$ if $x \geq 0$ or -1 otherwise. It is an explicit hashing function for the out-of-sample data[1]. Hence, given a new data $\mathbf{x}$, we can adopt Eq. (5) to generate its binary codes.

Fix $\mathbf{b}_i$ and update $\mathbf{R}$. This is the most difficult part in the entire solution, because the $\ell_{p,q}$ norm is neither smooth nor convex, and meanwhile, the orthogonality constraint limits the feasible set. To solve it, we rewrite the complicated $\ell_{p,q}$-norm problem into a weighted $\ell_{2,2}$-norm problem as below,

$$\min_{\mathbf{R}\mathbf{R}'=\mathbf{I}} \mathcal{O} = \sum_{i=1}^{n} \|\mathbf{w}_i \circ (\mathbf{b}_i - \mathbf{x}_i \mathbf{R})\|_2^2 = \|\mathbf{W} \circ (\mathbf{B} - \mathbf{X}\mathbf{R})\|_F^2, \tag{6}$$

where $\mathbf{X} = [\mathbf{x}_1; ...; \mathbf{x}_n]$ are the original training features, $\mathbf{B} = [\mathbf{b}_1; ...; \mathbf{b}_n]$ are the binary codes. $\mathbf{W} = [\mathbf{w}_1; ...; \mathbf{w}_n]$ is the weighting matrix, $\|\cdot\|_F$ is the Frobenius norm of matrix, and "$\circ$" represents the element-wise multiplication operation. In our algorithm, the weighting matrix is defined as follows

$$\begin{gathered} f_i = \|\mathbf{b}_i - \mathbf{x}_i \mathbf{R}\|_p^{q-p}, \; g_{ij} = |b_{ij} - \mathbf{x}_i \mathbf{R}_{*j}|^{p-2} \\ w_{ij} = (f_i g_{ij})^{0.5}. \end{gathered} \tag{7}$$

Based on the above definitions, it is easy to verify that Eq. (6) is equivalent to Eq. (3). Now, if we fix $\mathbf{W}$, solving the weighted $\ell_{2,2}$-norm problem is much easier than the original problem since it is a smooth and convex function. The only challenge left in this problem is to solve the orthogonality constraint. To address this issue, in this paper, we adopt the optimization algorithm proposed in [Wen and Yin, 2013], which starts by computing the gradient of $\mathcal{O}$ w.r.t. $\mathbf{R}$ as below

$$\mathbf{G} = \frac{\partial \mathcal{O}}{\partial \mathbf{R}} = \mathbf{X}'(\mathbf{W} \circ \mathbf{W} \circ (\mathbf{X}\mathbf{R} - \mathbf{B})). \tag{8}$$

Next, we construct a skew-symmetric matrix based on $\mathbf{G}$ as

$$\mathbf{A} = \mathbf{G}\mathbf{R}' - \mathbf{R}\mathbf{G}'. \tag{9}$$

Having obtained $\mathbf{G}$, the next step is to search the sequential point using the Crank-Nicolson-like scheme [Goldfarb *et al.*, 2009; Vese and Osher, 2002], which is described as follows

$$\mathbf{R}_{t+1} = \mathbf{R}_t - \tau \mathbf{A}(\frac{\mathbf{R}_{t+1} + \mathbf{R}_t}{2}), \tag{10}$$

where τ is the step size. The solution to Eq. (10) is given by

$$\mathbf{R}_{t+1} = (\mathbf{I} + \frac{\tau}{2}\mathbf{A})^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{A})\mathbf{R}_t. \tag{11}$$

Algorithm 1 Learning ITQ+

```
Input: Centered training data $\mathbf{X}$;
    Parameters $p \leq 2$ and $q \leq p$;
Output: Orthogonal matrix $\mathbf{R}$;
Initialize $\mathbf{R} = \mathbf{I}$, and $\mathbf{B} = \text{sign}(\mathbf{X}\mathbf{R})$;
repeat
    Update $b_{ij}$ with Eq. (5);
    Construct weighting matrix $\mathbf{W}$ by Eq. (7);
    Update $\mathbf{R}$ with Eq. (8)(9)(11);
until Convergence.
Return $\mathbf{R}$;
```

The objective function value in Eq. (6) will keep decreasing w.r.t. the updating rule in Eq. (11) until the stationary point is achieved. Note that $\mathbf{R}_{t+1}$ satisfies the orthogonal constraint (detailed proof can be found in [Wen and Yin, 2013]). We update $\mathbf{R}$ by fixing $\mathbf{W}$ as we can see that $\mathbf{W}$ depends on $\mathbf{R}$. Therefore, the updates of $\mathbf{R}$ and $\mathbf{W}$ can be implemented by an iterative strategy, which can be described in Algorithm 1.

2.3 Convergence Analysis

In this subsection, we will prove that the objective function value in Eq. (3) is non-increasing at each iteration of Algorithm 1, and is guaranteed to converge at the local optimum.

First, let us make it clear that $\mathcal{O}_{\text{ITQ+}}$ is obviously non-increasing w.r.t. the updating rule for b_{ij} in Eq. (5) because it is the global optimum given $\mathbf{R}$. Now, we need to prove that $\mathcal{O}_{\text{ITQ+}}$ is non-increasing under the updating rule in Eq. (11).

Lemma 1 *1 Given any $a > 0$ and $0 < b \leq a$, for $\forall x \geq 0$, we have the inequality: $ax^b - bx^a + b - a \leq 0$.*

Proof 1 Denote $c = b/a$ and $f(x) = x^c - cx + c - 1$. Apparently, $f(1) = 0$. Then, we have $f'(x) = cx^{c-1} - c$,leading to $f'(1) = 0$. In addition, $f''(x) = c(c-1)x^{c-2} \leq 0$ when $x \geq 0$ because $0 < c \leq 1$. This implies $f'(x) \geq 0 \, \forall x \in [0, 1]$ and $f'(x) \leq 0$ when $x > 1$. Therefore, $f(x) \leq f(1) = 0$. Finally, we can obtain $af(x^a) = ax^b - bx^a + b - a \leq 0$.□

Theorem 1 *The objective function $\mathcal{O}_{\text{ITQ+}}$ is non-increasing under the updating rule for $\mathbf{R}$ in Eq. (11).*

Proof 2 Let $\mathbf{Y} = \mathbf{B} - \mathbf{X}\mathbf{R}_t$, $\mathbf{Z} = \mathbf{B} - \mathbf{X}\mathbf{R}_{t+1}$, we have

$$\mathcal{O}_{\text{ITQ+}}^t = \sum_{i=1}^{n}(\sum_{j=1}^{k}|y_{ij}|^p)^{\frac{q}{p}}, \mathcal{O}_{\text{ITQ+}}^{t+1} = \sum_{i=1}^{n}(\sum_{j=1}^{k}|z_{ij}|^p)^{\frac{q}{p}}. \tag{12}$$

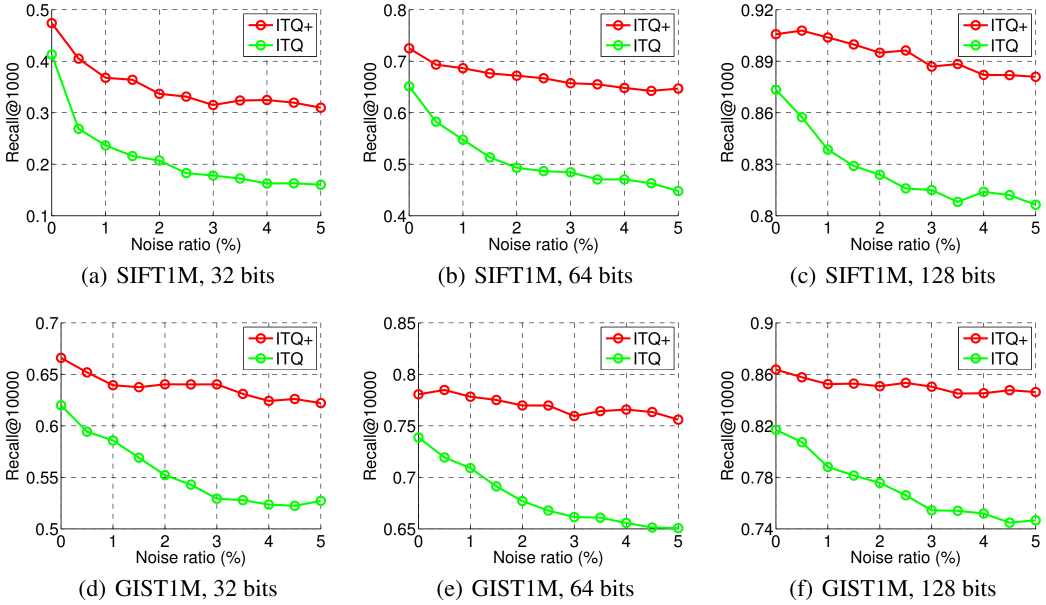
Based on the proof in [Wen and Yin, 2013], we know that Eq. (11) can decrease the value of $\mathcal{O}$ in Eq. (6), i.e., we have

$$\sum_{ij} f_i g_{ij} z_{ij}^2 \leq \sum_{ij} f_i g_{ij} y_{ij}^2. \tag{13}$$

Now if we set $a = 2$, $b = p$, x will be $|z_{ij}|/|y_{ij}|$. Based on the Lemma 1 above, we can obtain the following inequalities

$$\begin{aligned} & 2(\frac{|z_{ij}|}{|y_{ij}|})^p - p(\frac{|z_{ij}|}{|y_{ij}|})^2 + p - 2 \leq 0 \\ \Rightarrow & |z_{ij}|^p - \frac{p}{2}|y_{ij}|^{p-2}|z_{ij}|^2 \leq |y_{ij}|^p - \frac{p}{2}|y_{ij}|^{p-2}|y_{ij}|^2 \\ \Rightarrow & \sum_{ij} f_i(|z_{ij}|^p - \frac{p}{2}g_{ij}z_{ij}^2) \leq \sum_{ij} f_i(|y_{ij}|^p - \frac{p}{2}g_{ij}y_{ij}^2). \end{aligned} \tag{14}$$

[1] The out-of-sample data is the one that is not in the training set.



(a) SIFT1M, 32 bits (b) SIFT1M, 64 bits (c) SIFT1M, 128 bits

(d) GIST1M, 32 bits (e) GIST1M, 64 bits (f) GIST1M, 128 bits

Figure 2: Performance w.r.t. the noise ratio. We set $q = 1$ for ITQ+.

Combining inequalities (13) with (14) will bring us

$$\sum_i f_i\|\mathbf{z}_i\|_p^p = \sum_{ij} f_i|z_{ij}|^p \le \sum_{ij} f_i|y_{ij}|^p = \sum_i f_i\|\mathbf{y}_i\|_p^p. \qquad (15)$$

Denote $a = p$, $b = q$, and $x = \|\mathbf{z}_i\|_p/\|\mathbf{y}_i\|_p$, then we get

$$\begin{aligned} & p(\frac{\|\mathbf{z}_i\|_p}{\|\mathbf{y}_i\|_p})^q - q(\frac{\|\mathbf{z}_i\|_p}{\|\mathbf{y}_i\|_p})^p + q - p \le 0 \\ \Rightarrow & \|\mathbf{z}_i\|_p^q - \frac{q}{p}\|\mathbf{y}_i\|_p^{q-p}\|\mathbf{z}_i\|_p^p \le \|\mathbf{y}_i\|_p^q - \frac{q}{p}\|\mathbf{y}_i\|_p^{q-p}\|\mathbf{y}_i\|_p^p \qquad (16) \\ \Rightarrow & \sum_i(\|\mathbf{z}_i\|_p^q - \frac{q}{p}f_i\|\mathbf{z}_i\|_p^p) \le \sum(\|\mathbf{y}_i\|_p^q - \frac{q}{p}f_i\|\mathbf{y}_i\|_p^p). \end{aligned}$$

Again, if we combine inequalities (15) with (16), we obtain

$$\mathcal{O}_{\text{ITQ+}}^{t+1} = \sum_i \|\mathbf{z}_i\|_p^q \le \sum_i \|\mathbf{y}_i\|_p^q = \mathcal{O}_{\text{ITQ+}}^t, \qquad (17)$$

which ends the proof to Theorem 1. □

We have the following inequalities with the above proofs:

$$\mathcal{O}_{\text{ITQ+}}(\mathbf{B}_t, \mathbf{R}_t) \ge \mathcal{O}_{\text{ITQ+}}(\mathbf{B}_{t+1}, \mathbf{R}_t) \ge \mathcal{O}_{\text{ITQ+}}(\mathbf{B}_{t+1}, \mathbf{R}_{t+1})$$

which states that $\mathcal{O}_{\text{ITQ+}}$ is non-increasing with Algorithm 1.

3 Experiment

3.1 Datasets and Metrics

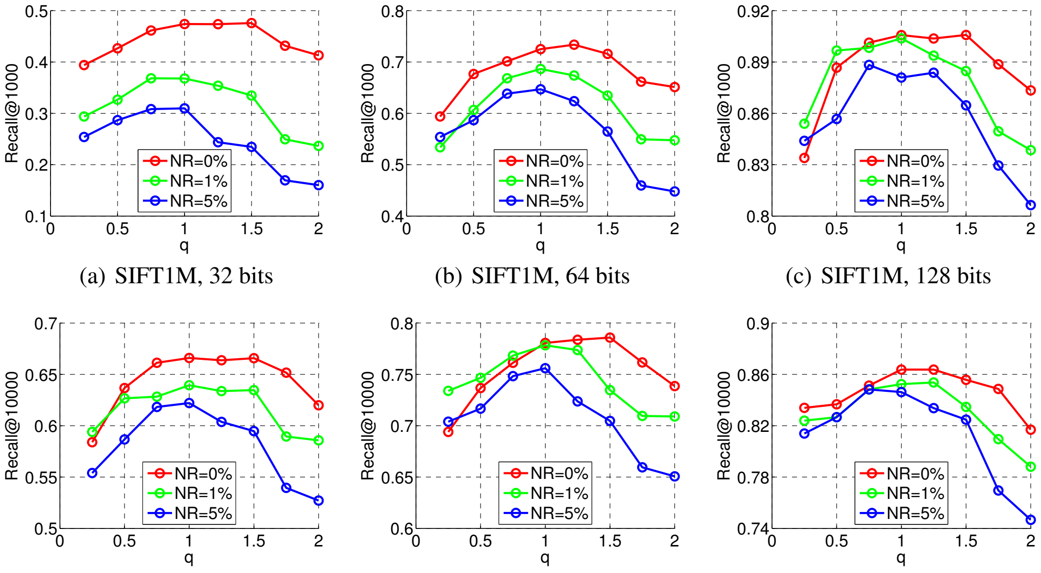
To demonstrate the effectiveness of ITQ+, we carried out comprehensive experiments for similarity search. In this paper, we adopt two widely used benchmark datasets. The first one is SIFT1M [Jégou *et al.*, 2011] which consists of 128-dimensional SIFT [Lowe, 2004] descriptors. It is made up of 1 million base vectors, $10,000$ query vectors and $100,000$ vectors for training. The second dataset is GIST1M [Jégou *et al.*, 2011] which contains 960-dimensional GIST [Oliva and Torralba, 2001] descriptors. This dataset contains 1 million base vectors, $1,000$ query vectors and $500,000$ for learning.

Following the settings in [Gong *et al.*, 2013; He *et al.*, 2013], we use Recall@R as the metric to evaluate the similarity search performance, which reflects the ratio between the number of the true positives in the first R retrieved points based on Hamming ranking and the total number of true positives. More precisely, the true positives for each query are the 10 nearest neighbors of the query in the base by running a brute-force linear scan measured by the ℓ_p-norm distance.

In addition, following the setting in [Jegou *et al.*, 2010; Gong *et al.*, 2013], we first centralize the data and perform a PCA to reduce the feature dimensionality to the length of binary codes. Afterwards, the rotation matrix $\mathbf{R}$ is learned from the reduced data. The binary codes are generated by a sign function after rotation. In addition, we repeat each experiment for 50 times and the average performance is reported.

3.2 Robustness Study

We firstly investigate the robustness of ITQ+ against the noise and outliers. To better investigate this property, we have manually added some noise to the data, where each dimension of each noisy point is sampled randomly from $100 \times \mathcal{N}(0, 1)$. This also implies that the distribution of noise is not as same as that of the original data. To understand the boundary of the algorithm, we continuously change the noise ratio (NR: the ratio between noisy points and original points), and evaluate the ℓ_2-norm similarity search performance of different methods. For ITQ+, we consistently set $q = 1$ when comparing to ITQ. The results of ITQ+ and ITQ on two datasets with different binary code lengths are shown in Figure 2. It can be observed that our ITQ+ is overwhelmingly better than ITQ at all the situations in terms of the Recall. On average, we



(a) SIFT1M, 32 bits

(b) SIFT1M, 64 bits

(c) SIFT1M, 128 bits

(d) GIST1M, 32 bits

(e) GIST1M, 64 bits

(f) GIST1M, 128 bits

Figure 3: Performance w.r.t. q.

have improved the recall of the original ITQ by $\mathbf{12.2}\%$ when NR $= 5\%$. It is worthwhile to point out that the results actually demonstrate the following properties of our algorithm.

ITQ+ performs observably better than ITQ even when applying to the original data (no manual noise; NR $= 0$). The major reason is that the data are from the real-world dataset, on which the noises and outliers have existed. Therefore, it turns out that noisy data and outliers in the real-world dataset are indeed influential in the performance of ITQ because their large errors may dominate the total distortion due to the squared loss. In contrast, in ITQ+, we adopt the q-th ($q < 2$) order loss function that can effectively suppress the effect of noisy data and outliers as the learned parameters can better capture the intrinsic information in the dataset. In other words, our ITQ+ is better suited to deal with data in the wild.

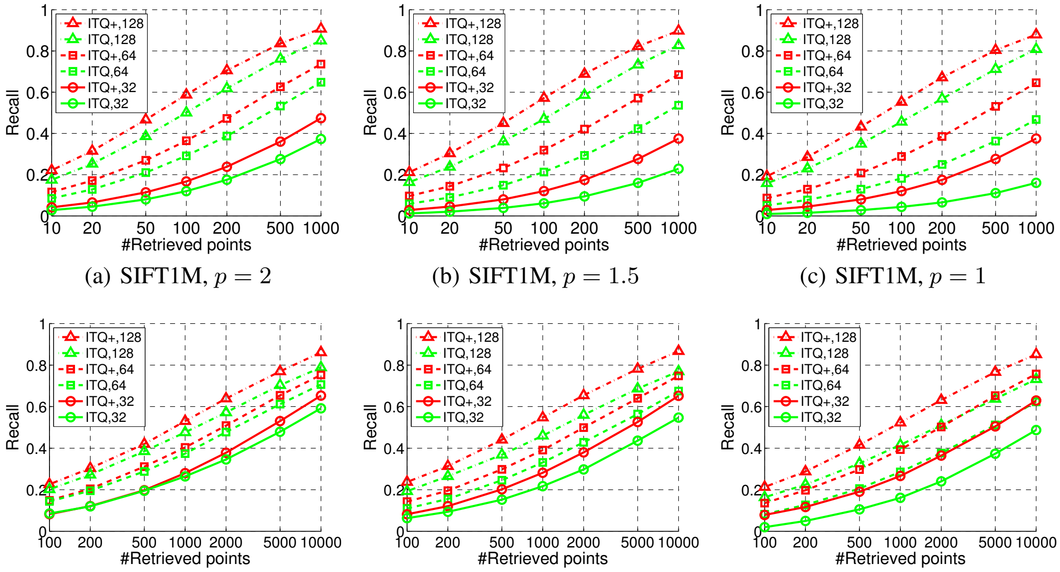
When NR gets increased, the similarity search performance of ITQ degrades rapidly. This phenomenon once again demonstrates that ITQ is sensitive to noise and outliers in data because of the squared loss, as we have mentioned before. On the contrary, ITQ+ shows very stable performance in most cases when we increase NR. More importantly, it can be seen that the performance gap between ITQ+ and ITQ becomes even larger when increasing NR. This again demonstrates the superior robustness of the proposed ITQ+ against the noise.

3.3 Effect of Parameter q

There is one important parameter q in ITQ+. Here, we investigate how the algorithm will behave when varying q. To do so, we change the value of q and plot the corresponding performance of ITQ+ on both datasets with different binary code lengths and noise ratios. The results are illustrated in Figure 3. It is noticed that ITQ is a special case of ITQ+ when $q = 2$. We can get the following observations based on the results.

Firstly, in all settings, we can find a Bell-shape curve for ITQ+. Basically, the model is affected by both noise and normal data. With a large q (say, $q > 1.5$), ITQ+ will increase the weight of those large-distortion entries such that the model will be biased by them. Unfortunately, due to the existence of noisy entries and their large distortions, the learned model will deviate significantly to fit the outliers from the one which best suits to the normal data. Therefore, the performance of ITQ+ degrades significantly when we increases q from 1.5 to 2, especially in more noisy settings, e.g., NR $= 5\%$. On the other hand, if q is too small (say,$q < 0.5$), we cannot obtain good results either. According to the principle, the difference between normal and noisy data becomes smaller in this case, though the effect of outliers is suppressed. In the extreme case where $q = 0$, every entry has the same distortion 1 such that any model is the solution for this case. Thus, it is almost impossible to find the optimal model for normal data. This interprets why ITQ+ performs worse when we decrease q from 0.5 to 0.25, especially when there is less noise, e.g., NR $= 0$. In Figure 3, we can see that ITQ+ performs stably good when $q \in [0.75, 1.25]$ where the outliers affect ITQ+ much less and that a model which can well fit to the normal data is learned.

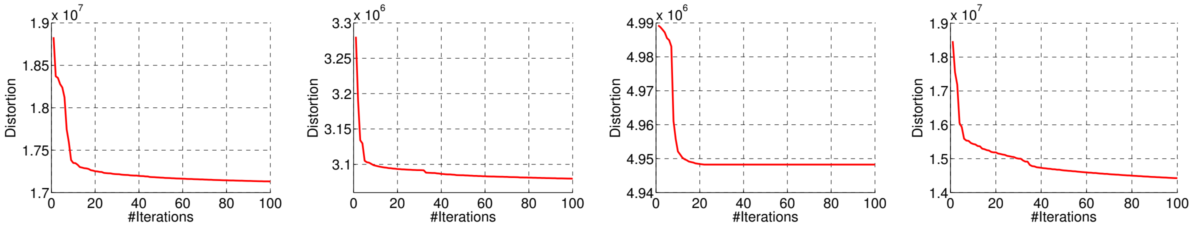
Secondly, we can observe that the performance-vs-q curve behaves differently at different noise levels. Specifically, given a small NR, e.g., NR $= 0$, ITQ+ seems more sensitive to q when $q < 1$, because the the performance changes dramatically when varying q in this range. On the other hand, given a large NR, e.g., NR $= 5\%$, ITQ+ becomes more sensitive when $q > 1$. The reason is analogous to our analysis in the last paragraph. When there is little noise, the primary target of ITQ+ is to fit the normal data. In this case, the performance may degrade rapidly if q is too small. On the other hand, as a result of the increasing noise, the primary target of ITQ+ becomes to suppress the influence of noise. Thus, increasing the value of q when $q > 1$ leads to much worse performance.



(a) SIFT1M, $p = 2$ (b) SIFT1M, $p = 1.5$ (c) SIFT1M, $p = 1$

(d) GIST1M, $p = 2$ (e) GIST1M, $p = 1.5$ (f) GIST1M, $p = 1$

Figure 4: Performance for ℓ_p-norm similarity search.



(a) $p = 2, q = 1.5$ (b) $p = 2, q = 1$ (c) $p = 1.5, q = 1$ (d) $p = q = 1$

Figure 5: Convergence study, SIFT1M, 64 bits.

3.4 ℓ_p-norm Similarity Search

In this subsection, we will demonstrate the effectiveness of ITQ+ for ℓ_p-norm similarity search. For ITQ+, we can set the parameter p depending on the specific task and we set $q = 1$.

We consider the ℓ_2-norm (Euclidean distance), $\ell_{1.5}$-norm and ℓ_1-norm (Manhattan distance) similarity search because of the space limitation. The recall curves of ITQ+ and ITQ on both datasets with different code lengths for three tasks are plotted in Figure 4, respectively. Here, we use ℓ_2-norm as the reference as ITQ is designed for this task. We can observe that ITQ+ has stable performance on different tasks whereas ITQ performs much worse on other two tasks than on ℓ_2-norm task. For example, the Recall@1000 of ITQ drops from 0.651 for ℓ_2-norm to 0.474 for ℓ_1-norm on SIFT1M with 64 bits. Consequently, the performance gap between ITQ+ and ITQ becomes much larger when we change p from 2 to 1.5 and 1. This result demonstrates that ITQ+ can well support the ℓ_p-norm similarity search but ITQ cannot handle the tasks excluding ℓ_2-norm search. In fact, as we have shown in Figure 1(b), the learning algorithm of ITQ may unavoidably lead to larger distortion with more iterations since it adopts ℓ_2 loss.

3.5 Convergence Study

We have theoretically proved that Algorithm 1 leads to non-increasing objective value. Now, we empirically investigate its convergence property by conducting the experiment on SIFT1M with 64 bits. Because Algorithm 1 is designed for the general $\ell_{p,q}$-norm loss, we assign different values to p and q. The objective function value in Eq. (3) w.r.t. the number of iterations with different settings are plotted in Figure 5. As can be seen, the objective value decreases steadily with more iterations and can achieve a nearly stable value within less than 50 iterations, which validates the effectiveness of Algorithm 1. For a fair comparison, we terminate the algorithm after 50 iterations in all experiments as suggested by ITQ.

4 Conclusion

In this paper, we have presented an enhanced ITQ algorithm, termed ITQ+, which changes the $\ell_{2,2}$-norm loss to a more general $\ell_{p,q}$-norm loss. The benefits are twofold. On the one hand, the algorithm becomes more robust to the noise, which potentially makes ITQ+ better suited to search similarity in the real-world data. On the other hand, promoting to $\ell_{p,q}$-norm loss allows ITQ+ to handle various applications, where different distance measurements are requested. The major technical challenge comes from minimizing the new $\ell_{p,q}$- loss function, which is a non-smooth and non-convex optimization problem. To solve this orthogonality constrained $\ell_{p,q}$-norm minimization problem, we propose an efficient algorithm and rigorously prove its convergence. Comprehensive experiments on two benchmarks show that ITQ+ performs significantly better than ITQ, and demonstrate that ITQ+ is robust to noise and works well for ℓ_p-norm similarity search.

References

[Altman, 1992] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):91–110, 1992.

[Andoni and Indyk, 2006] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS'06*, 2006.

[Cheng *et al.*, 2015] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE TGRS*, 2015.

[Furnas *et al.*, 1988] George W. Furnas, Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR*, 1988.

[Ge *et al.*, 2014] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. *TPAMI*, 2014.

[Gionis *et al.*, 1999] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *VLDB*, 1999.

[Goldfarb *et al.*, 2009] D. Goldfarb, Z. Wen, and W Yin. A curvilinear search method for the p-harmonic flow on spheres. *SIAM J. Imaging Sci*, 2(1):84–109, 2009.

[Gong *et al.*, 2013] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, 2013.

[He *et al.*, 2013] Kaiming He, Fang Wen, and Jian Sun. K-means hashing: an affinity-preserving quantization method for learning binary compact codes. In *CVPR*, 2013.

[Huang *et al.*, 2013] Jin Huang, Feiping Nie, Heng Huang, and Chris H. Q. Ding. Robust manifold nonnegative matrix factorization. *TKDD*, 8(3):11, 2013.

[Jegou *et al.*, 2010] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[Jégou *et al.*, 2011] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *TPAMI*, 33(1):117–128, 2011.

[Jiang *et al.*, 2015] Wenhao Jiang, Feiping Nie, and Heng Huang. Robust dictionary learning with capped l1-norm. In *IJCAI*, pages 3590–3596, 2015.

[Kong and Li, 2012] Weihao Kong and Wu-Jun Li. Isotropic hashing. In *NIPS*, 2012.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Oliva and Torralba, 2001] A. Oliva and T. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[Pan *et al.*, 2014] Qihe Pan, Deguang Kong, Chris H. Q. Ding, and Bin Luo. Robust non-negative dictionary learning. In *AAAI*, pages 2027–2033, 2014.

[Singh and Jain, 2010] Uday Pratap Singh and Sanjeev Jain. Content based image retrieval using euclidean and manhattan metrics. *Journal of Math. Sciences: Advances and Appl.*, 4(1):217–226, 2010.

[Vese and Osher, 2002] L.A. Vese and S.J. Osher. Numerical methods for p-harmonic flows and applications to image processing. *SIAM J. Numer. Anal*, 2002.

[Wang *et al.*, 2013] H. Wang, F. Nie, W. Cai, and H. Huang. Semi-supervised robust dictionary learning via efficient l-norms minimization. In *ICCV*, 2013.

[Wang *et al.*, 2014] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. *CoRR*, abs/1408.2927, 2014.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2):397–434, 2013.

[Wright *et al.*, 2009] John Wright, Arvind Ganesh, Shankar Rao, YiGang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.

[Xu *et al.*, 2013] Bin Xu, Jiajun Bu, Yue Lin, Chun Chen, Xiaofei He, and Deng Cai. Harmonious hashing. In *IJCAI*, 2013.

[Zhang *et al.*, 2010] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. Self-taught hashing for fast similarity search. In *SIGIR*, 2010.

[Zhang *et al.*, 2014] Ting Zhang, Chao Du, and Jingdong Wang. Composite quantization for approximate nearest neighbor search. In *ICML*, pages 838–846, 2014.