Microsoft Research
微软亚洲研究院

Optimized Product Quantization

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun

MSRA

Introduction

- Compact Coding for ANN Search
  - Memory
    - 128-d float: 512 bytes → 16 bytes
    - 1 billion items: 512 GB → 16 GB
  - Time
    - Computation: x10-x100 faster
    - Transmission (disk/web): x30 faster

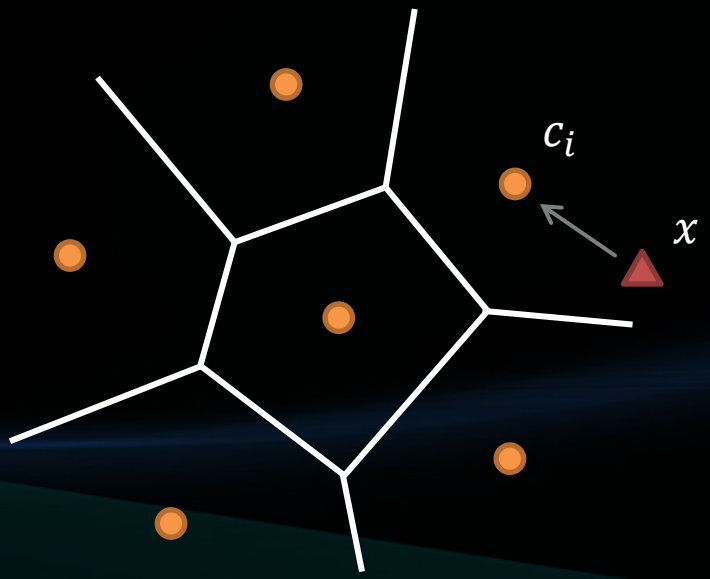
Background

- Vector Quantization (VQ)

x → c_i → $i(x)$

nearest codeword

code stored



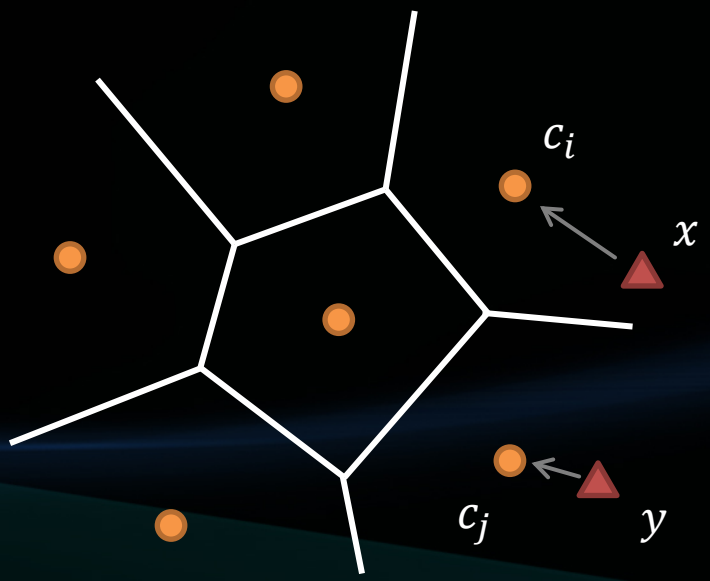
codewords: K

code length: $B = log_2 K$

Background

- VQ for ANN Search

$$d(x, y) \approx d(c_i, c_j) \triangleq lut(i, j)$$



K-by-K look-up table

Background

- K-means

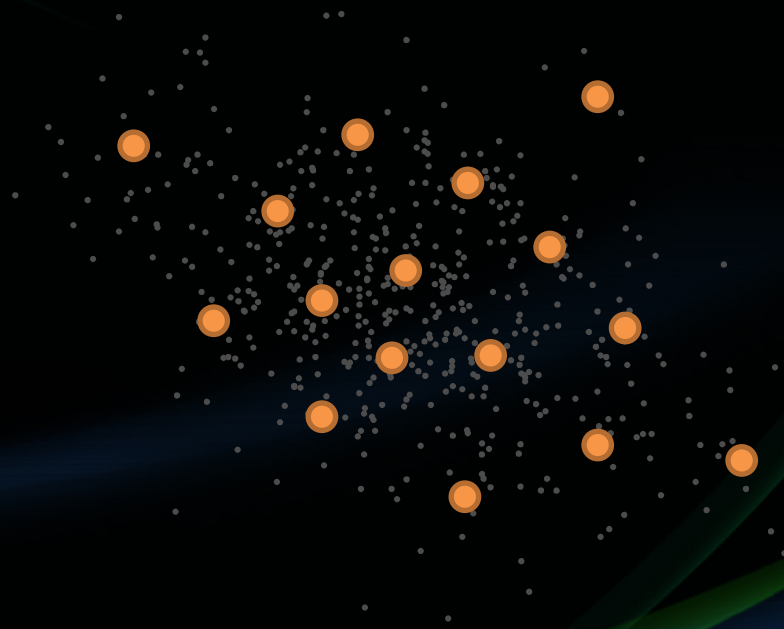
$$\min_{c} \sum_{x} \left\| x - c_{i(x)} \right\|^2 \quad \text{(distortion)}$$

☺

- Minimal distortion

☹

- Intractable look-up: $K = 2^B$



Background

- Product Quantization (PQ) [PAMI 2011]

$$\min_{C^1,\dots,C^M} \sum_x \|x - c_{i(x)}\|^2$$

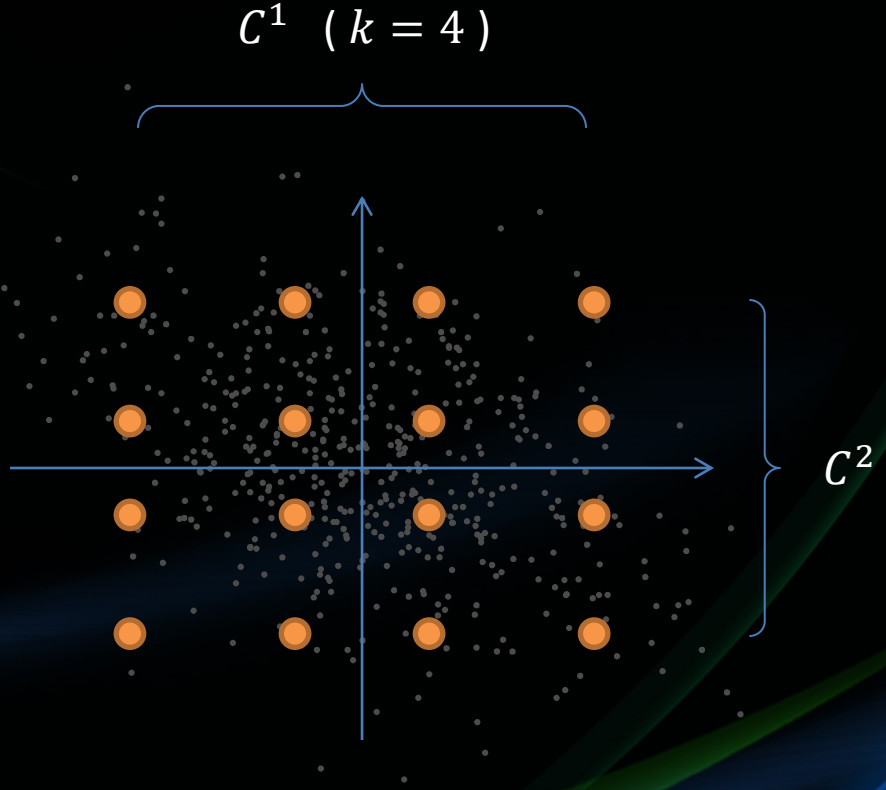
s.t. $c \in C^1 \times C^2 \dots \times C^M$

☺

- Huge codebook: $K = k^M$
- Tractable: M k-by-k tables

☹

- Sensitive to projection



Background

- Iterative Quantization (ITQ) [CVPR 2011]

$$\min_{R} \sum_{x} \left\| x - c_{i(x)} \right\|^2$$

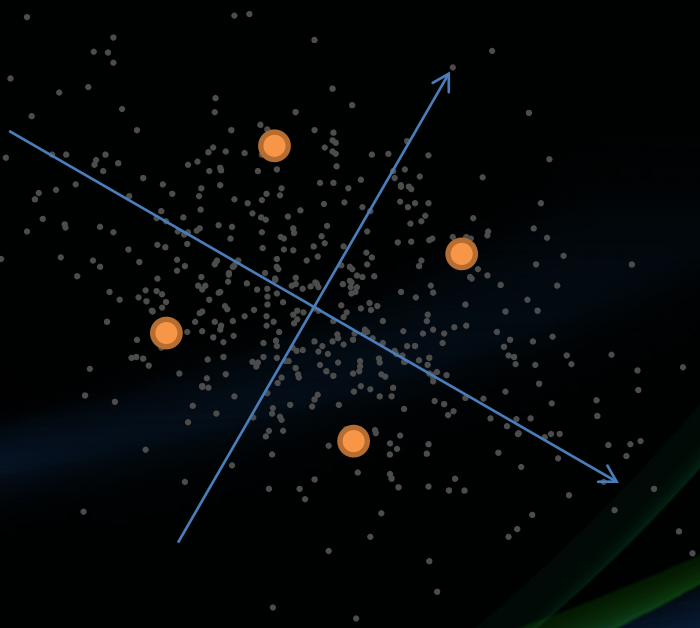
s.t. $Rc \in \{-1,1\}^D$, $R^T R = I$

☺

- Optimized wrt R

☹

- 1-d subspace
- $k = 2$ only
- no look-up



Our method

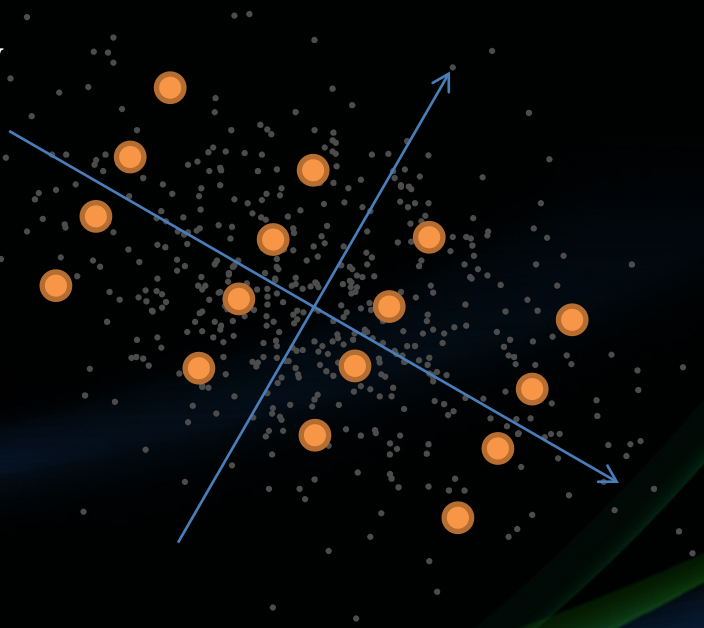
- Optimized Product Quantization (OPQ) [CVPR 2013]

$$\min_{R,C^1,\dots,C^M} \sum_x \|x - c_{i(x)}\|^2$$

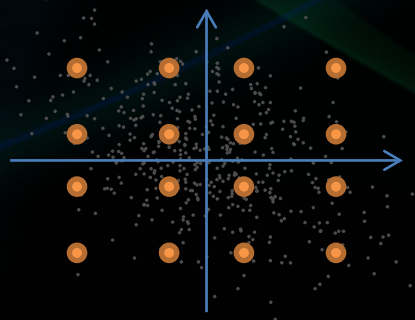
s.t. $Rc \in C^1 \times C^2 \dots \times C^M, R^T R = I$

☺

- Huge codebook: $K = k^M$
- Tractable: M k-by-k tables
- High-dim subspace
- $k \geq 2$
- Optimize wrt R



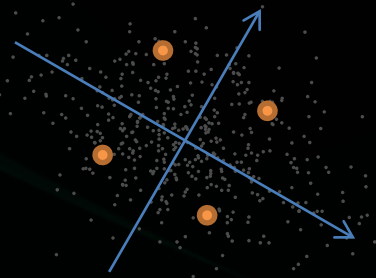
Relations



PQ

$$\min_{C^1,\dots,C^M} \sum_x \|x - c_{i(x)}\|^2$$

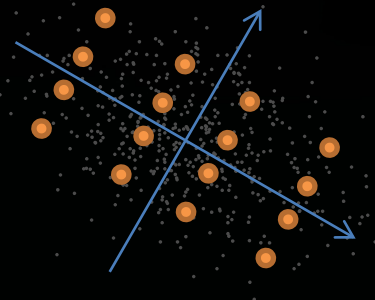
s.t. $c \in C^1 \times C^2 \dots \times C^M$



ITQ

$$\min_{R} \sum_x \|x - c_{i(x)}\|^2$$

s.t. $Rc \in \{-1,1\}^D$



OPQ

$$\min_{R,C^1,\dots,C^M} \sum_x \|x - c_{i(x)}\|^2$$

s.t. $Rc \in C^1 \times C^2 \dots \times C^M$

OPQ vs. PO	optimized R
OPQ vs. ITQ	optimized $C^1, \dots, C^M$
Challenges	coupled R and $C^1, \dots, C^M$

Solutions

- Challenges - coupled R and $C^1, \dots, C^M$
- Solution I
  - decoupling
  - fix R solve $C^1, \dots, C^M$
  - fix $C^1, \dots, C^M$ solve R
- Solution II
  - lower bound: involves R only
  - minimize lower bound w.r.t R

Solution I

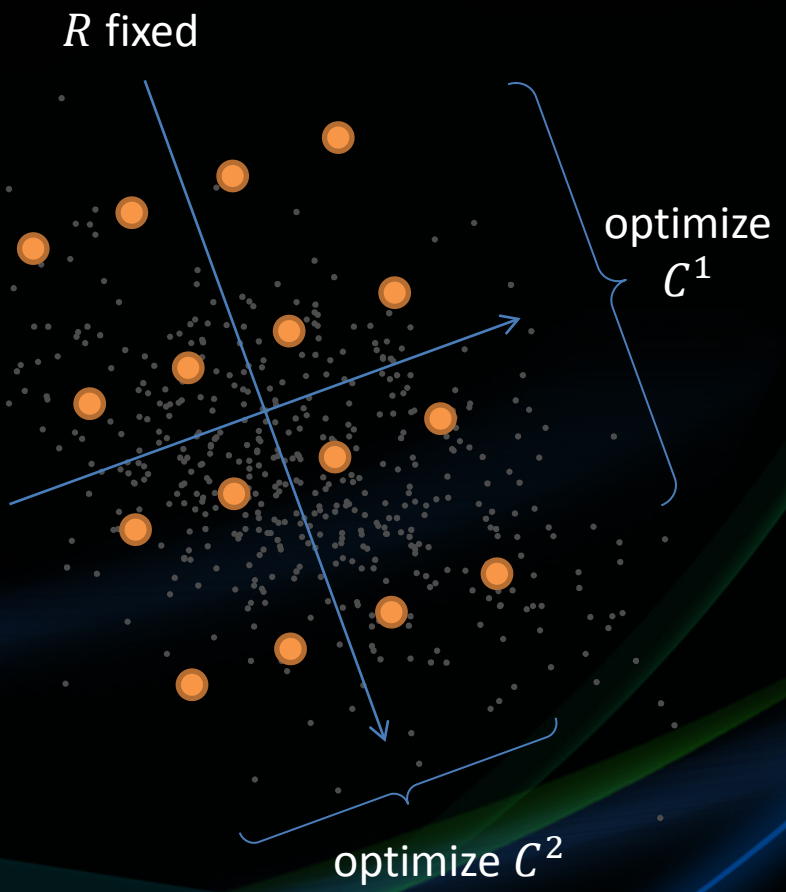
- Step 1: fix R, solve for $C^1, \dots, C^M$

$$\min_{C^1,\dots,C^M} \sum_x \|\hat{x} - \hat{c}_{i(x)}\|^2$$

s.t. $\hat{c} \in C^1 \times C^2 \dots \times C^M$

with $\hat{x} = Rx$, and $\hat{c} = Rc$

Standard PQ
in a projected space



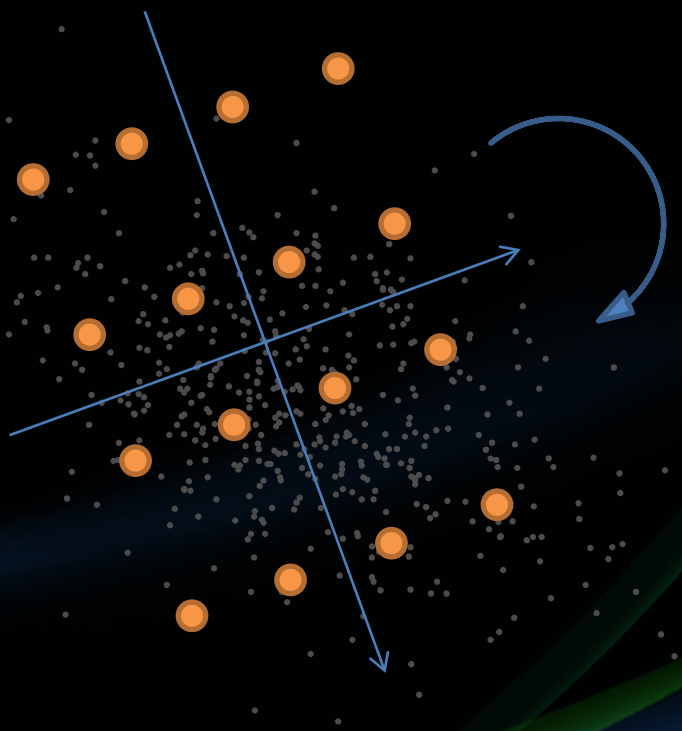
Solution I

- Step 2: fix $C^1, \ldots, C^M$, solve for R

$$\min_R \sum_x \|Rx - \hat{c}_{i(x)}\|^2$$

s.t. $\quad \hat{c} \in C^1 \times C^2 \ldots \times C^M, R^T R = I$

Rotate codewords without changing their relative positions



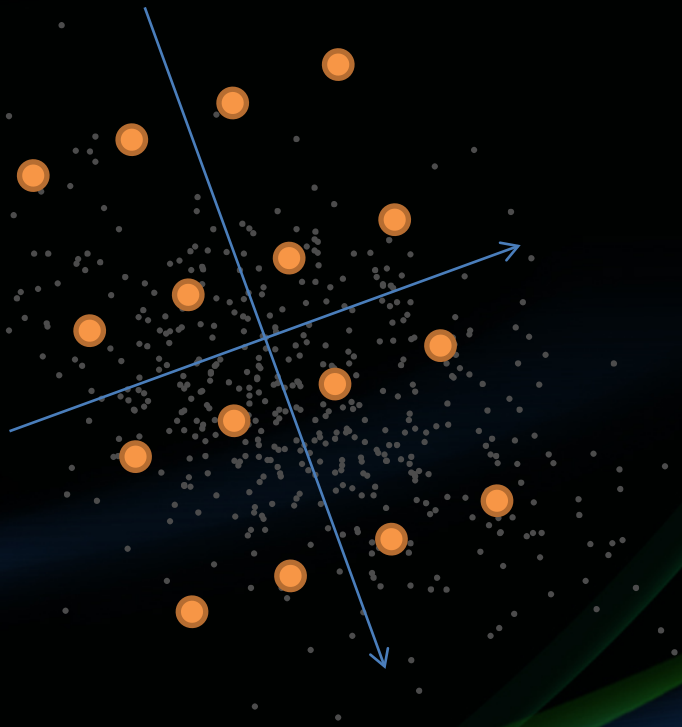
Solution I

- Step 2: fix $C^1, \ldots, C^M$, solve for R

$$\min_R \|RX - Y\|_F^2$$

$$\text{s.t.} \quad R^T R = I$$

  - Let $X = \{x\}, Y = \{y\}, y = \hat{c}_{i(x)}$
  - $R = VU^T,\ [U, V] = \text{svd}(XY^T)$



Solution I

- Initialize
- Repeat
  - Fix R, solve:

    $$\min_{C^1,\dots,C^M} \sum_x \left\| \hat{x} - \hat{c}_{i(x)} \right\|^2$$ (classical PQ)

  - Fix $C^1, \dots, C^M$, solve:

    $$\min_R \|RX - Y\|_F^2$$ (classical ITQ)
- Until convergence

Solution II

- Decoupling
  - lower bound: involves R only
  - minimize lower bound w.r.t R
- Assumes Gaussian distribution
  - analytical forms
  - theoretical guarantees
  - simple, non-iterative

Solution II

- Assumes $x \sim N(0, \Sigma)$
- $\hat{x} = Rx \sim N(0, \hat{\Sigma})$, with $\hat{\Sigma} = R\Sigma R^T$
- Decompose $\hat{x} = (\hat{x}^1, \hat{x}^2, \ldots, \hat{x}^M)$ into M subspaces

$$\hat{x}^m \sim N(0, \hat{\Sigma}_{mm})$$

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \cdots & \hat{\Sigma}_{1M} \\ \vdots & \ddots & \vdots \\ \hat{\Sigma}_{M1} & \cdots & \hat{\Sigma}_{MM} \end{pmatrix}$$

Solution II

- Rate distortion theory for $\hat{x}^m \sim N(0, \hat{\Sigma}_{mm})$

$$E^m \geq k^{-2\frac{M}{D}} \frac{D}{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}}$$

  - Nearly achieved: k-means

- PQ distortion for $\hat{x} \sim N(0, \hat{\Sigma})$

$$E \geq k^{-2\frac{M}{D}} \frac{D}{M} \sum_{m=1}^{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}}$$

- Optimize lower bound

$$\min_{R} \sum_{m=1}^{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}}$$
$$\text{s.t.} \quad R^T R = I$$

Solution II

- Optimization - achieve the lower bound of the lower bound

$$\min_{R} \sum_{m=1}^{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} \quad \text{s.t.} \quad R^T R = I$$

$$\sum \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} \geq M \prod \left|\hat{\Sigma}_{mm}\right|^{\frac{1}{D}} \geq M\left|\hat{\Sigma}\right|^{\frac{1}{D}} \equiv M|\Sigma|^{\frac{1}{D}}$$

AM-GM inequality

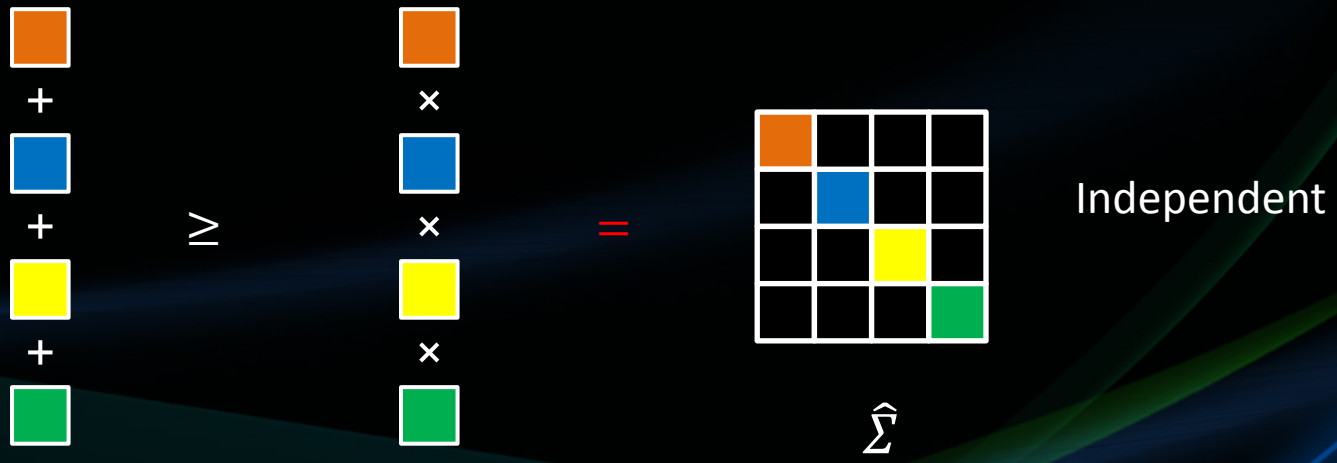
Fischer inequality

Solution II

- Optimization - achieve the lower bound of the lower bound

$$\min_{R} \sum_{m=1}^{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} \quad \text{s.t.} \quad R^T R = I$$

$$\sum \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} \geq M \prod \left|\hat{\Sigma}_{mm}\right|^{\frac{1}{D}} = M\left|\hat{\Sigma}\right|^{\frac{1}{D}} \equiv M|\Sigma|^{\frac{1}{D}}$$

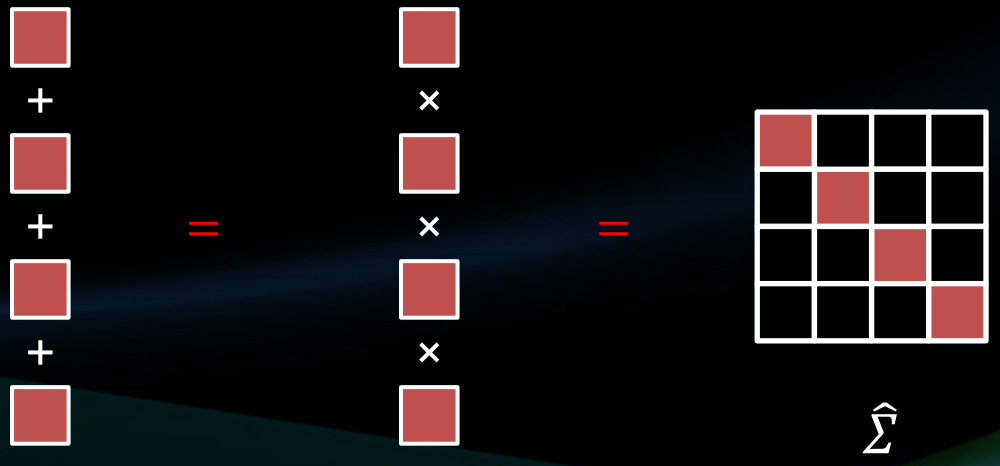


Solution II

- Optimization - achieve the lower bound of the lower bound

$$\min_{R} \sum_{m=1}^{M} \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} \quad \text{s.t.} \quad R^T R = I$$

$$\sum \left|\hat{\Sigma}_{mm}\right|^{\frac{M}{D}} = M \prod \left|\hat{\Sigma}_{mm}\right|^{\frac{1}{D}} = M\left|\hat{\Sigma}\right|^{\frac{1}{D}} \equiv M|\Sigma|^{\frac{1}{D}}$$



Solution II

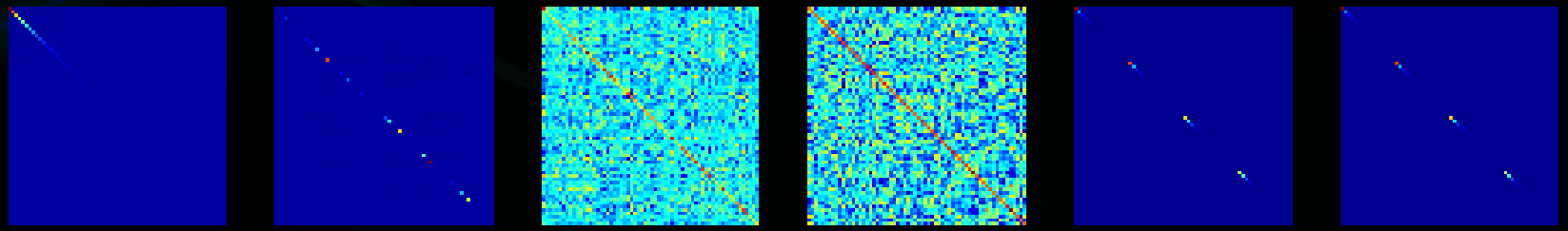
- Algorithm
  - *independent:* PCA
  - *balanced*:

$$\left|\hat{\Sigma}_{11}\right| = \cdots = \left|\hat{\Sigma}_{mm}\right| = \cdots = \left|\hat{\Sigma}_{MM}\right|$$

    - $\left|\hat{\Sigma}_{mm}\right| = \prod \{$ eigenvalues of $\hat{\Sigma}_{mm} \}$
    - Greedy allocation:
      - Sort the eigenvalues of Σ
      - Prepare M buckets
      - Allocate the largest eigenvalue to the bucket having smallest product

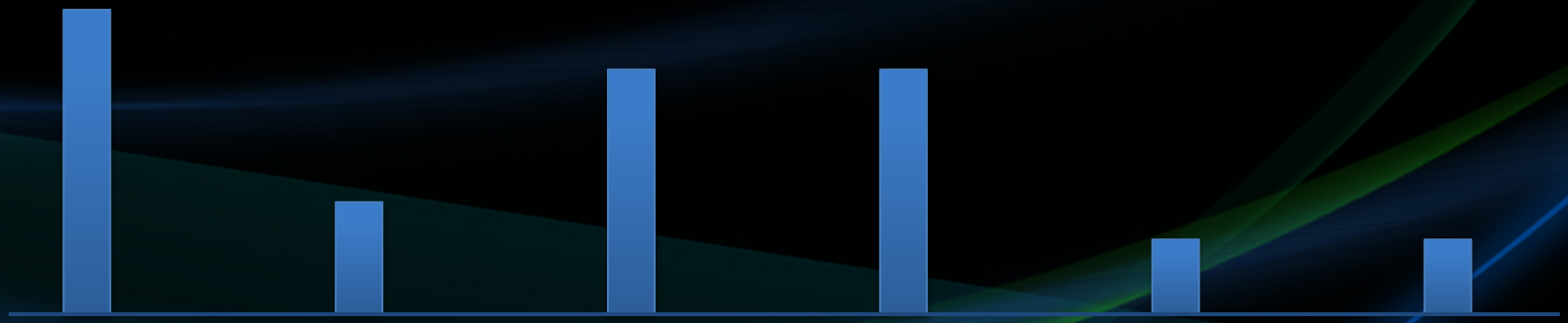
Verification

- 64-d Gaussian, descending eigenvalues, 4 subspaces



R	not rotated	random order	random rotation	forced balance	solution I	solution II
independent	Yes	Yes	No	No	Yes	Yes
balanced	No	almost	almost	Yes	Yes	Yes

distortion

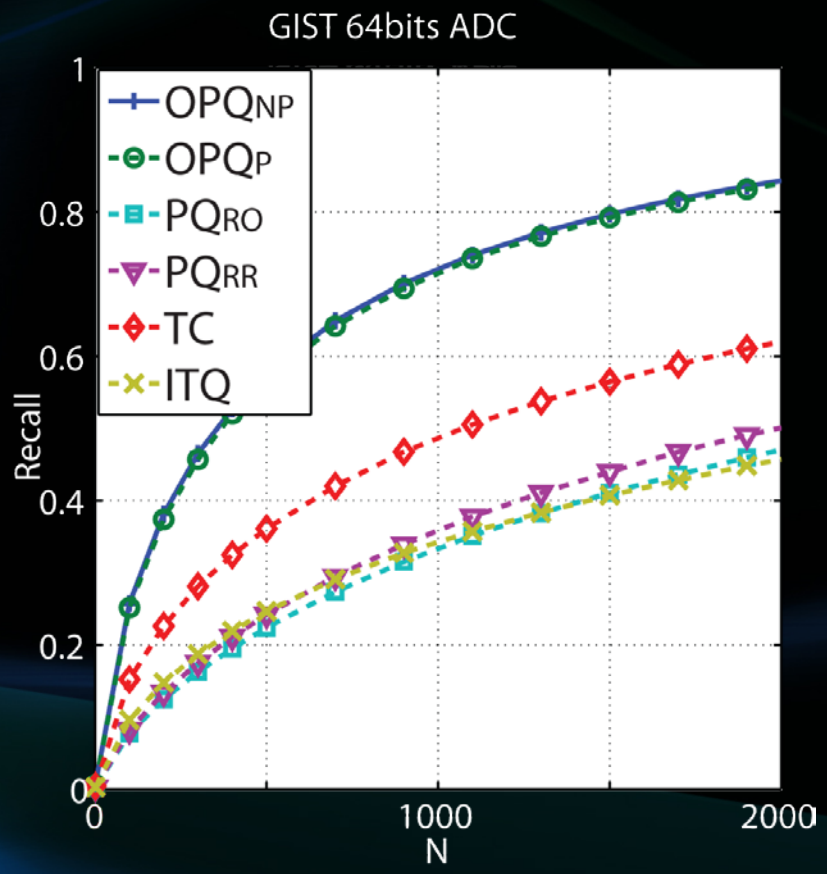


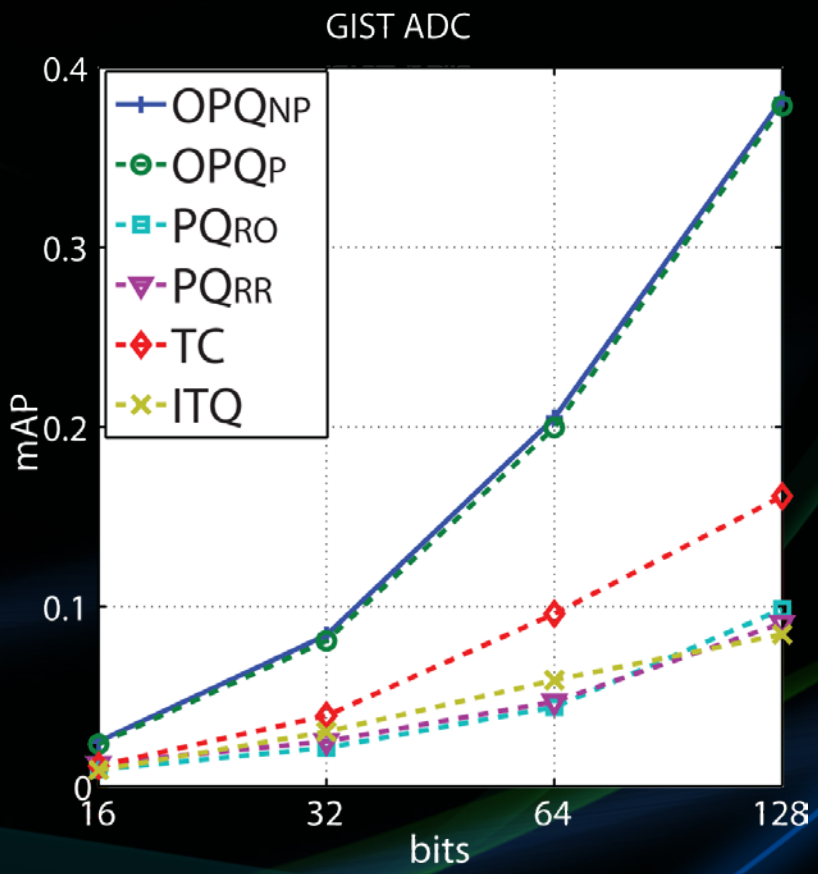
Solution I vs. II

- I – non-parametric
  - Better fits non-Gaussian
  - Iterative (offline)
  - Needs init (e.g. by II)
- II – parametric
  - Guarantees for Gaussian
  - Solid theories
  - Non-iterative
  - Less well for non-Gaussian
- Best practice – solution I + solution II (initialize)

Experiments

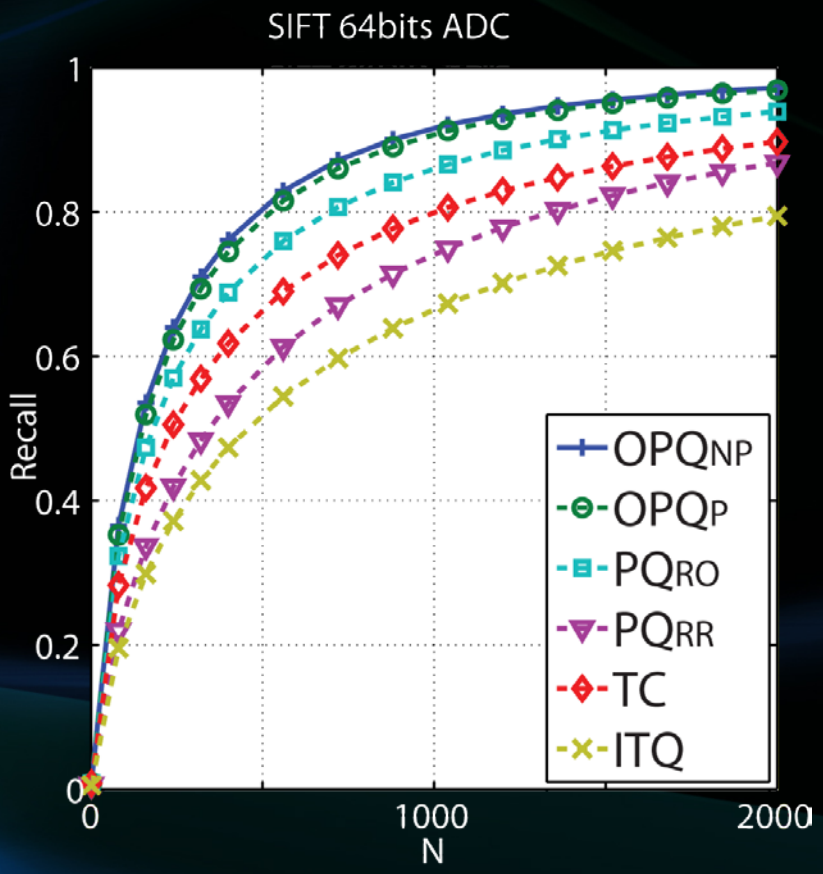
- 1 million GIST, 100 NNs, exhaustive ranking

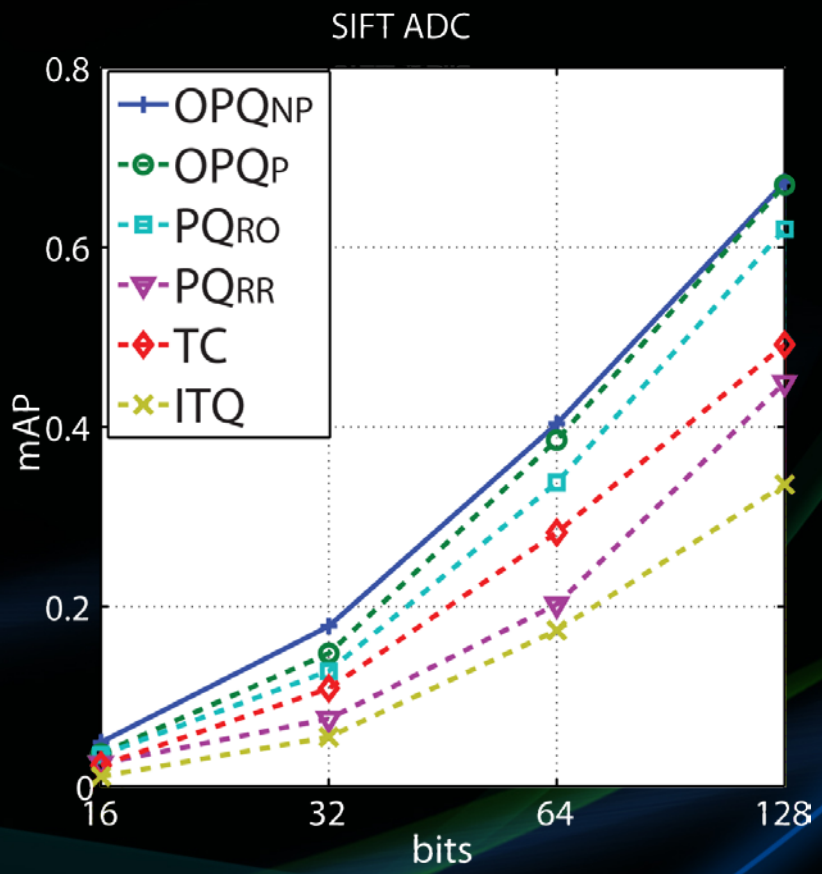




Experiments

- 1 million SIFT, 100 NNs, exhaustive ranking





Experiments

- 1 billion SIFT, 1 NNs, inverted indexing + re-ranking
  - Build inverted indexing via PQ [Babenko, CVPR 2012]
  - Re-rank short lists via PQ [Jegou, PAMI 2011]
  - We optimize both

	short list length	Recall@100	time
[CVPR 2012]	10,000	74.8	7ms
Ours	10,000	79.4	7ms
[CVPR 2012]	100,000	96.0	49ms
Ours	100,000	97.3	49ms

Experiments

- Image retrieval
  - feature: VLAD [Jegou, PAMI 2011]
  - dataset: Holiday [Jegou, PAMI 2011]
  - ground truth: semantic

memory / image	8 bytes	16 bytes	32 bytes
mAP (PQ_{RR})	38.1	47.9	53.0
mAP (OPQ)	47.7	52.2	54.3

Conclusion

- Excellent performance for ANN
- Solid theories
- Widely applicable