UNIVERSITÉ DE GENÈVE

SIP stochastic information processing

Vector compression for similarity search using Multi-layer Sparse Ternary Codes

Sohrab Ferdowsi, Slava Voloshynovskiy, Dimche Kostadinov

Department of Computer Science, University of Geneva, Switzerland

{Sohrab.Ferdowsi, svolos, Dimche.Kostadinov}@unige.ch

http://sip.unige.ch

Overview

Applications:

- Large-scale retrieval systems
- Learned compression of feature vectors
- Compressed representation useful for fast similarity search

Contributions:

- Rate-distortion (R-D) study of ternary and binary encoding
- Designing R-D efficient multi-layer Sparse Ternary Codes (STC)

Background

- Ability to **search for similarity** within a database is crucial for modern retrieval systems.
- A wide-spread solution is binary hashing.
- We proposed ternary hashing [WIFS'16] as an alternative to binary hashing.
- We showed that ternary encoding has higher **coding gain** than binary encoding [ISIT'17].
- Here we extend ternary encoding for the task of compression, so that we can have list-refinement.
- Our design challenge: To have good R-D performance within STC limitations.

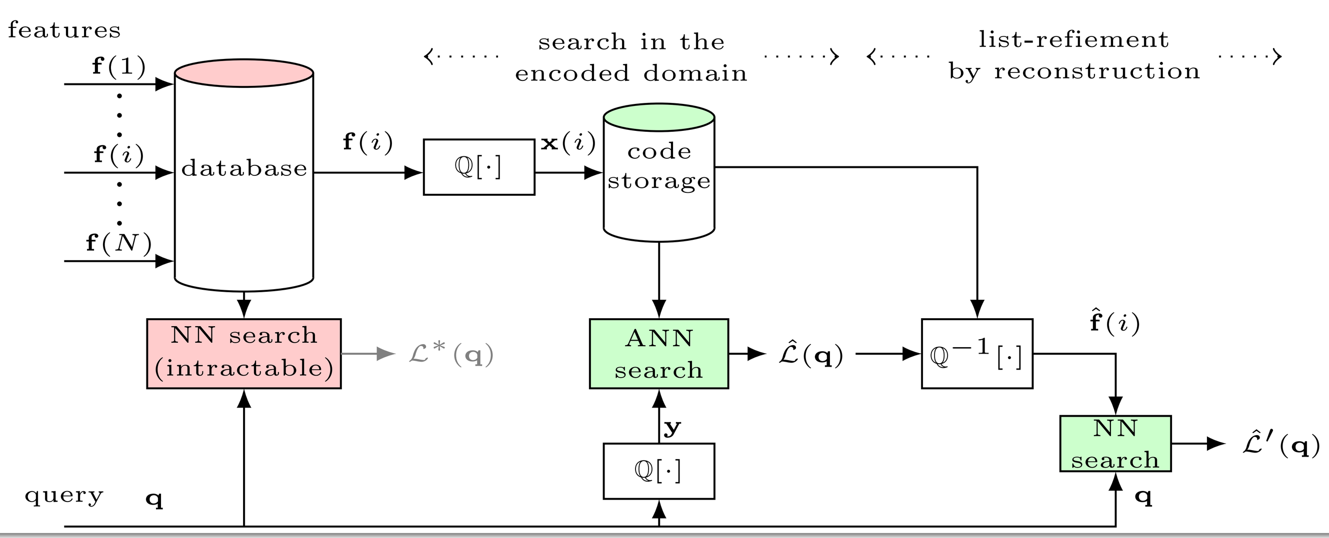
Problem formulation: ANN search

Similarity search:

- (Exact) Nearest Neighbor (NN) search: $\mathcal{L}(\mathbf{q}) = \{1 \leqslant i \leqslant N | d_E(\mathbf{f}(i), \mathbf{q}) \leqslant \epsilon n\}$
- Approximate Nearest Neighbor (ANN) search: $\hat{\mathcal{L}}(\mathbf{q}) = \{1 \leqslant i \leqslant N | d_H(\mathbb{Q}[\mathbf{f}(i)], \mathbb{Q}[\mathbf{q}]) \leqslant \epsilon\}$
- Our solution: List-refinement with reconstruction $\hat{\mathcal{L}}'(\mathbf{q}) = \{i \in \hat{\mathcal{L}}(\mathbf{q}) | d_E(\mathbb{Q}^{-1}[\mathbb{Q}[\mathbf{f}(i)]], \mathbf{q}) \leqslant \epsilon\}$

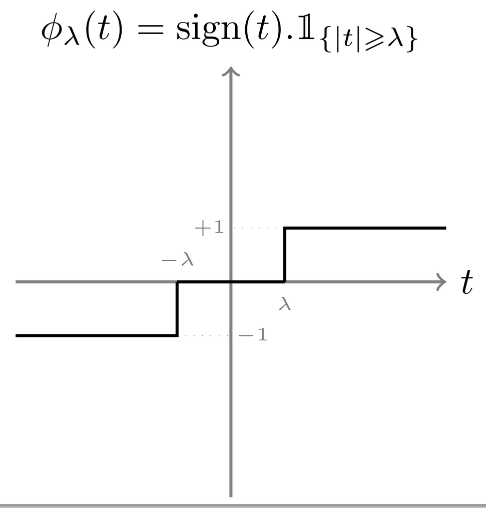
Compression:

- Encoding: $\mathbf{x} = \mathbb{Q}[\mathbf{f}]$
- Reconstruction: $\hat{\mathbf{f}} = \mathbb{Q}^{-1}[\mathbf{x}]$
- Rate: $\mathcal{R} = \frac{1}{n}\mathbb{E}[\#$ bits used to represent $\mathbf{x}]$
- Distortion: $\mathcal{D} = \mathbb{E}[d_E(\mathbf{F}, \hat{\mathbf{F}})]$ $d_E(\mathbf{a}, \mathbf{b}) \triangleq \frac{1}{n}||\mathbf{a} - \mathbf{b}||_2^2$



Single-layer Sparse Ternary Codes (STC)

$\phi_\lambda(t) = \text{sign}(t).\mathbb{1}_{\{|t| \geqslant \lambda\}}$



Encoding:

$\mathbf{x} = \phi_\lambda(\mathrm{A}\mathbf{f}) \odot \boldsymbol{\beta}$

(projection + ternarization + re-weighting)

Reconstruction:

$\hat{\mathbf{f}} = \mathrm{B}\mathbf{x} = \mathrm{B}\phi_\lambda(\mathrm{A}\mathbf{f}) \odot \boldsymbol{\beta}$

(projection)

Optimizing single-layer STC

Back-projection B:

- Decompose $\mathrm{B} = (\mathrm{A}^T\mathrm{A})^{-1}\mathrm{A}^T\mathrm{B}'$, optimize B':

$$\mathrm{B}' = \underset{\mathrm{B}'}{\text{argmin}} ||\mathrm{F} - (\mathrm{A}^T\mathrm{A})^{-1}\mathrm{A}^T\mathrm{B}'\mathrm{X}||_{\mathcal{F}}^2$$
$$= \underset{\mathrm{B}'}{\text{argmin}}\, \text{Tr}\Big[- 2\mathrm{A}\mathrm{A}^T\mathrm{A}\mathrm{F}\mathrm{X}^T\mathrm{B}'^T + \mathrm{B}'\mathrm{X}\mathrm{X}^T\mathrm{B}'^T\mathrm{A}\mathrm{A}^T\Big].$$
$$\Rightarrow \mathrm{B}' = \mathrm{A}\mathrm{F}\mathrm{X}^T(\mathrm{X}\mathrm{X}^T)^{-1}.$$

Projection A:

- To have un-correlated $\mathbf{X}$:

$$\mathrm{C}_F \triangleq \frac{1}{n}\mathbb{E}[\mathbf{F}\mathbf{F}^T] = \mathrm{U}_F\Sigma_F\mathrm{U}_F^T$$

- Simply as in PCA: $\mathrm{A} = \mathrm{U}_F^T$
- $\tilde{\mathbf{X}} \triangleq \mathrm{A}\mathbf{F} \sim \mathcal{N}(\mathbf{0}, \Sigma_F)$
- With this choice of A: $\Rightarrow \mathrm{B}' = \mathbb{I}_n$, $\mathrm{B} = \mathrm{A}^T$

- Calculation of distortion:

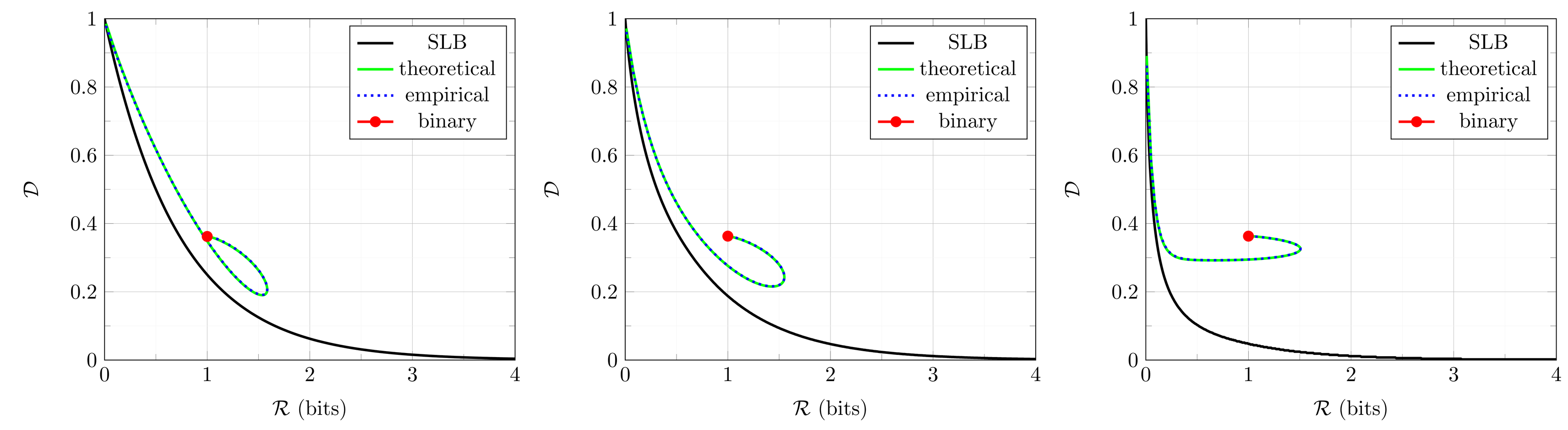
$$\mathcal{D} = \mathbb{E}\big[d_E(\mathbf{F}, \hat{\mathbf{F}})\big] = \frac{1}{n}\mathbb{E}\big[||\mathbf{F} - \mathrm{A}^T\mathbf{X}||_2^2\big] = \frac{1}{n}\mathbb{E}\big[||\mathrm{A}\mathbf{F} - \mathbf{X})||_2^2\big] = \frac{1}{n}\mathbb{E}\big[||\tilde{\mathbf{X}} - \phi_\lambda(\tilde{\mathbf{X}}) \odot \boldsymbol{\beta}||_2^2\big].$$

- Distortion per each dimension ($\mathcal{D} = \sum_{i=1}^n D_i$):

$$D_i = \mathbb{E}\big[(\tilde{X}_i - \beta_i\phi_\lambda(\tilde{X}_i))^2\big] = \int_{-\infty}^{-\lambda}(\tilde{x}_i + \beta_i)^2 p(\tilde{x}_i)d\tilde{x}_i + \int_{-\lambda}^{+\lambda}\tilde{x}_i^2 p(\tilde{x}_i)d\tilde{x}_i + \int_{+\lambda}^{+\infty}(\tilde{x}_i - \beta_i)^2 p(\tilde{x}_i)d\tilde{x}_i$$
$$\Longrightarrow D_i = \sigma_i^2 + 2\beta_i^2\mathcal{Q}\Big(\frac{\lambda}{\sigma_i}\Big) - \frac{4\beta_i\sigma_i}{\sqrt{2\pi}}\exp\Big(\frac{-\lambda^2}{2\sigma_i^2}\Big)$$

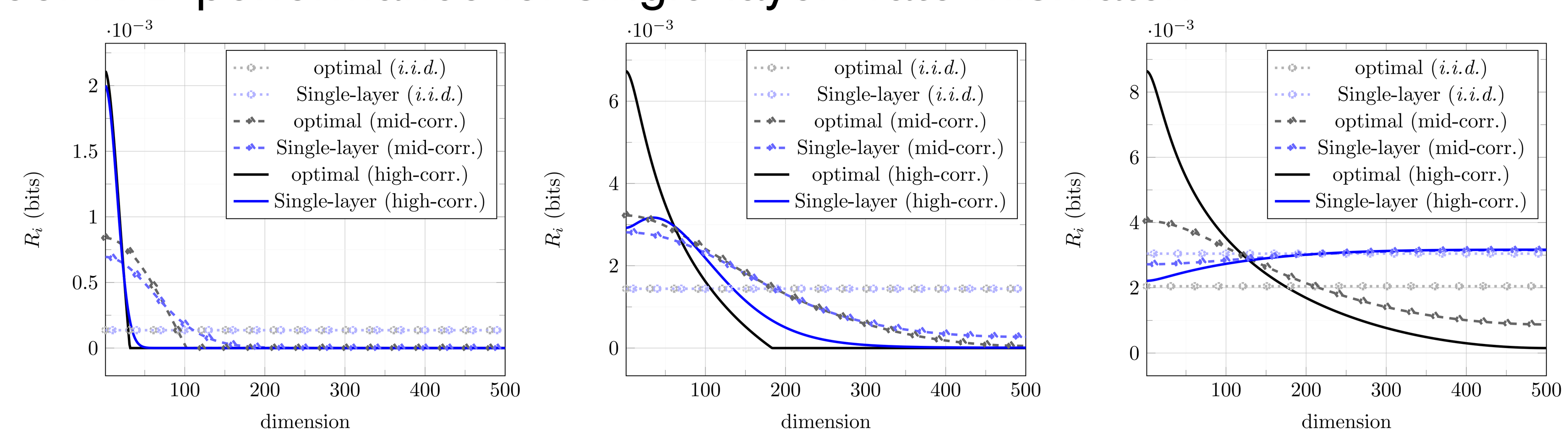
- Optimal Re-weighting vector: $\beta_i^* = \text{argmin}_{\beta_i} D_i = \dfrac{\sigma_i \exp\Big(\frac{-\lambda^2}{2\sigma_i^2}\Big)}{\sqrt{2\pi}\mathcal{Q}\Big(\frac{\lambda}{\sigma_i}\Big)}.$
- Rate: $\mathcal{R} = \frac{1}{n}H_t(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n H_t(X_i) = -\frac{1}{n}\sum_{i=1}^n \Big(2\alpha_i\log_2(\alpha_i) + (1 - 2\alpha_i)\log_2(1 - 2\alpha_i)\Big)$
- Sparsity per each dimension: $\alpha_i \triangleq \mathbb{P}[X_i = +\beta_i] = \mathbb{P}[X_i = -\beta_i]$

R-D performance of single-layer STC on AR(1) Gaussian sources



(a) *i.i.d.* (b) mid-correlation) (c) high-correlation

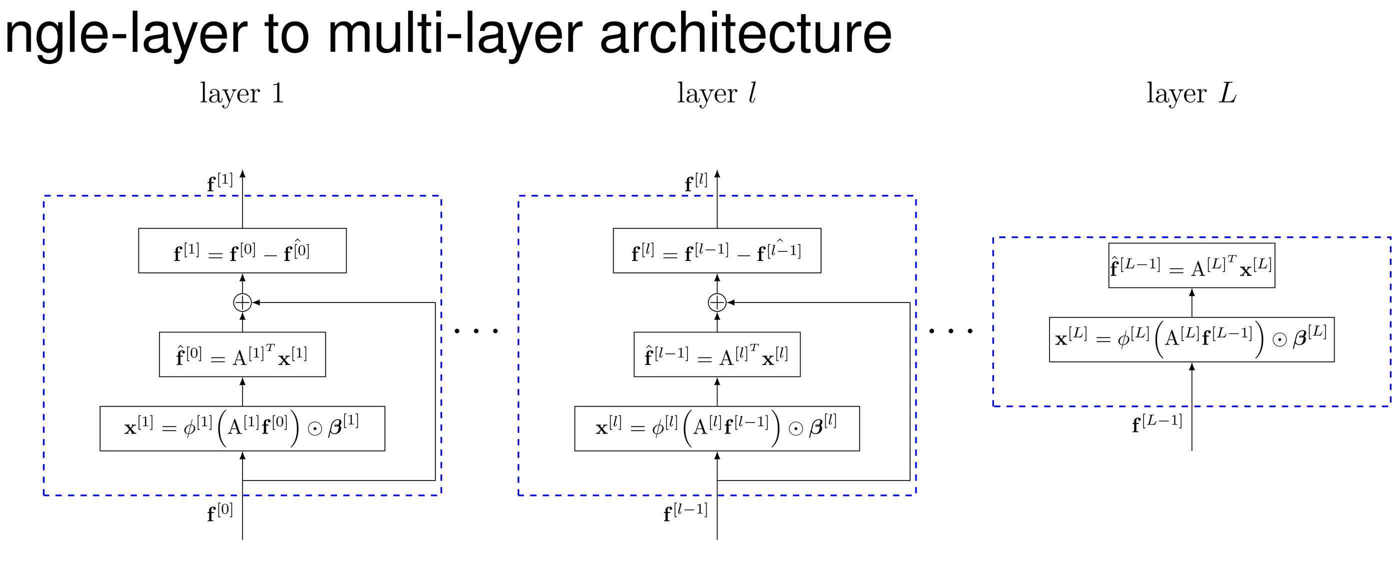
Poor R-D performance for single-layer: rate mismatch



(a) low-rate (b) mid-rate (c) high-rate

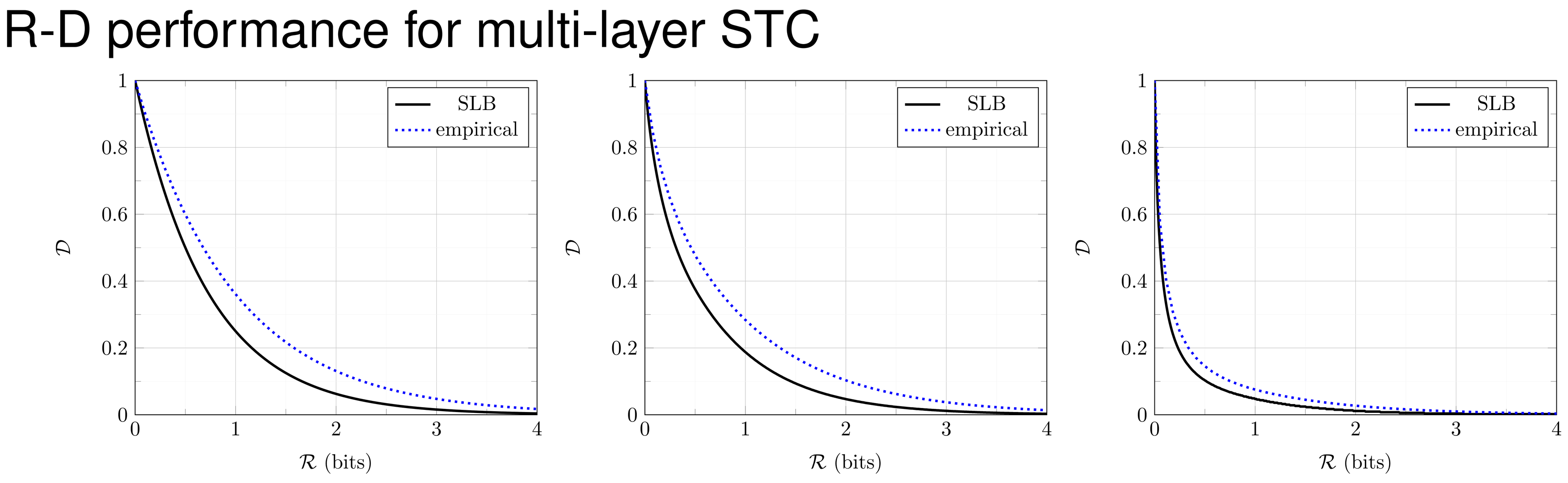
- Optimal rate allocation is calculated using the "reverse water-filling" paradigm from information theory.
- At higher rates, rate allocation deviates largely from optimal assignment.
- Binary encoding is a special case of ternary encoding with zero sparsity and hence rate is very high.

From single-layer to multi-layer architecture



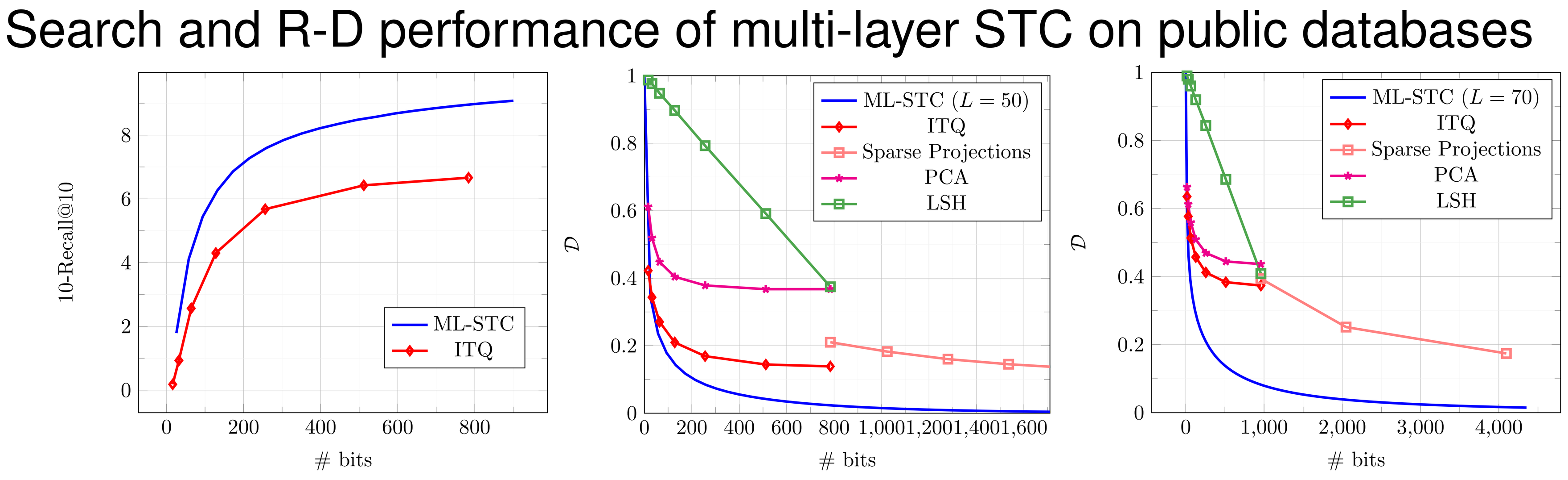
$\hat{f} = \hat{f}^{[0]} + \cdots + \hat{f}^{[l]} + \cdots + \hat{f}^{[L]}$

R-D performance for multi-layer STC



(a) low-correlation (b) mid-correlation) (c) high-correlation

Search and R-D performance of multi-layer STC on public databases



(a) MNIST (b) MNIST (c) Gist-1M

Conclusions

- Single-layer encoding is insufficient to provide good R-D performance at high rates.
- Residual-based multi-layer encoding can provide reasonable R-D performance.
- Since binary-encoding has rate mismatch, it cannot benefit from multi-layer encoding.
- Ternary encoding with high sparsity has low rate mismatch and can benefit from multi-layer encoding.
- Future work: Joint learning of all layers.
- Python implementation: https://github.com/sssohrab/DSW2018