Supplemental Materials for “A Scalable Optimization Mechanism for Pairwise based Discrete Hashing”

Xiaoshuang Shi, Fuyong Xing, Zizhao Zhang, Manish Sapkota, Zhenhua Guo, and Lin Yang

Proposition 1. *When $\mathbf{H}_l = \mathbf{H}_{l-1}$, the optimal solution of Eq. (6) is also the optimal solution of Eq. (4) w.r.t $\mathbf{H}$.*

Proof. Obviously, if $\mathbf{H}_l = \mathbf{H}_{l-1}$, it is the optimal solution of Eq. (6). Then we can consider the following formulation:

$$\underset{\mathbf{H}_l,\mathbf{H}_{l-1}}{min} \left\|\mathbf{H}_l\mathbf{H}_{l-1}^T - \mathbf{S}\right\|_F^2 \le \underset{\mathbf{H}}{min} \left\|\mathbf{H}\mathbf{H}^T - \mathbf{S}\right\|_F^2 \tag{A1}$$

Similar to one major motivation of asymmetric discrete hashing algorithms [1] [2], in Eq. (6), the feasible region of $\mathbf{H}_l$, $\mathbf{H}_{l-1}$ in the left term is more flexible than $\mathbf{H}$ in the right term (Eq. (4)), i.e. the left term contains both two cases $\mathbf{H}_l \neq \mathbf{H}_{l-1}$ and $\mathbf{H}_l = \mathbf{H}_{l-1}$. Only when $\mathbf{H}_l = \mathbf{H}_{l-1}$, $\underset{\mathbf{H}_l,\mathbf{H}_{l-1}}{min} \left\|\mathbf{H}_l\mathbf{H}_{l-1}^T - \mathbf{S}\right\|_F^2 = \underset{\mathbf{H}}{min} \left\|\mathbf{H}\mathbf{H}^T - \mathbf{S}\right\|_F^2$. It suggests that when $\mathbf{H}_l = \mathbf{H}_{l-1}$, it is the optimal solution of Eq. (4). Therefore, when $\mathbf{H}_l = \mathbf{H}_{l-1}$, it is the optimal solution of both Eq. (4) and Eq. (6). □

Theorem 1. *Given a discrete matrix $\mathbf{H} \in \{-1,1\}^{n\times m}$ and a real nonzero matrix $\mathbf{Z} \in \mathbb{R}^{m\times m}$, $\underset{\mathbf{P}}{min} \|\mathbf{H} - \mathbf{P}\mathbf{Z}\|_F^2 + \left\|\mathbf{P}\mathbf{\Gamma}^{\frac{1}{2}}\right\|_F^2 = Tr\left\{\mathbf{H}(\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z})\mathbf{H}^T\right\}$, where $\mathbf{\Gamma} \in \mathbb{R}^{m\times m}$ is a positive-definite diagonal matrix and $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is an identity matrix.*

Proof. It is easy to verify that $\mathbf{P}^* = \mathbf{H}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}$ is the global optimal solution to the problem $\underset{\mathbf{P}}{min} \|\mathbf{H} - \mathbf{P}\mathbf{Z}\|_F^2 + \left\|\mathbf{P}\mathbf{\Gamma}^{\frac{1}{2}}\right\|_F^2$. Substituting $\mathbf{P}^*$ into the above objective, its minimum value is $Tr\left\{\mathbf{H}(\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z})\mathbf{H}^T\right\}$. Therefore, Theorem 1 is proved. □

Theorem 1 suggests that when $\mathbf{H}_{l-1}^T\mathbf{H}_{l-1} = \gamma(\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z})$, the quadratic problem in Eq. (7) can be linearized as a regression type. We show the details in Theorem 2.

Theorem 2. *When $\mathbf{H}_{l-1}^T\mathbf{H}_{l-1} = \gamma(\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z})$, where γ is a constant, the problem in Eq. (7) can be reformulated as:*

$$\begin{aligned} &\underset{\mathbf{H},\mathbf{P}}{min}\ \gamma(\|\mathbf{H} - \mathbf{P}\mathbf{Z}\|_F^2 + \left\|\mathbf{P}\mathbf{\Gamma}^{\frac{1}{2}}\right\|_F^2) \\ &-2\lambda Tr\left\{\mathbf{H}\mathbf{H}_{l-1}^T\mathbf{S}^T\right\}, s.t.\ \mathbf{H} \in \{-1,1\}^{n\times m}. \end{aligned} \tag{8}$$

Proof. Based on the condition $\mathbf{H}_{l-1}^T\mathbf{H}_{l-1} = \gamma(\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z})$ and Theorem 1, substituting $\mathbf{P}^* = \mathbf{H}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}$ into the objective of Eq. (8), whose objective value is equal to that of Eq. (7), Therefore, Theorem 2 is proved. □

Theorem 3. *Suppose that a full rank matrix $\mathbf{H}_{l-1}^T\mathbf{H}_{l-1} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T$, where $\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$ is a positive diagonal matrix and $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_m$. If $\gamma \ge \Lambda_{ii}^2$ and $\Gamma_{ii} > 0$, $(1 \le i \le m)$ and a real nonzero matrix $\mathbf{Z} = \mathbf{V}\mathbf{\Delta}\mathbf{U}^T$ satisfies the conditions: $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_m$, and $\mathbf{\Delta} \in \mathbb{R}^{m\times m}$ is a non-negative real diagonal matrix with the i-th diagonal element being $\Delta_{ii} = \sqrt{\frac{\gamma\Gamma_{ii}}{\Lambda_{ii}^2} - \Gamma_{ii}}$.*

Proof. Based on singular value decomposition (SVD), there exist matrices $\mathbf{V}$ and $\mathbf{U}_\mathbf{Z}$, satisfying the conditions $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_m$ and $\mathbf{U}_\mathbf{Z}\mathbf{U}_\mathbf{Z}^T = \mathbf{U}_\mathbf{Z}^T\mathbf{U}_\mathbf{Z} = \mathbf{I}_m$, such that a real nonzero matrix $\mathbf{Z}$ is represented by $\mathbf{Z} = \mathbf{V}\mathbf{\Delta}\mathbf{U}_\mathbf{Z}^T$, where $\mathbf{\Delta}$ is a non-negative real diagonal matrix. Then $\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z} = \mathbf{U}_\mathbf{Z}(\mathbf{I}_m - \mathbf{\Delta}(\mathbf{\Delta}^2 + \mathbf{\Gamma})^{-1}\mathbf{\Delta})\mathbf{U}_\mathbf{Z}^T$. Note that when the vectors in $\mathbf{V}$ and $\mathbf{U}_\mathbf{Z}$ corresponds to the zero diagonal elements, they can be constructed by employing a Gram-Schmidt process such that $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_m$ and $\mathbf{U}_\mathbf{Z}\mathbf{U}_\mathbf{Z}^T = \mathbf{U}_\mathbf{Z}^T\mathbf{U}_\mathbf{Z} = \mathbf{I}_m$, and these constructed vectors are not unique.

Since $\mathbf{H}_{l-1}^T\mathbf{H}_{l-1} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T$ and $\mathbf{I}_m - \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Gamma})^{-1}\mathbf{Z} = \frac{\mathbf{H}_{l-1}^T\mathbf{H}_{l-1}}{\gamma}$, it can have $\gamma\Gamma_{ii}(\Delta_{ii}^2 + \Gamma_{ii})^{-1} = \Lambda_{ii}^2$ when $\mathbf{U}_\mathbf{Z} = \mathbf{U}$. Since there exists $0 < \Gamma_{ii}(\Delta_{ii}^2 + \Gamma_{ii})^{-1} \le 1$, γ should satisfy: $\gamma \ge \Lambda_{ii}^2$ and $\Gamma_{ii} > 0$. Additionally, based on $\gamma\Gamma_{ii}(\Delta_{ii}^2 + \Gamma_{ii})^{-1} = \Lambda_{ii}^2$, there exists $\Delta_{ii} = \sqrt{\frac{\gamma\Gamma_{ii}}{\Lambda_{ii}^2} - \Gamma_{ii}}$. Therefore, Theorem 3 is proved. □

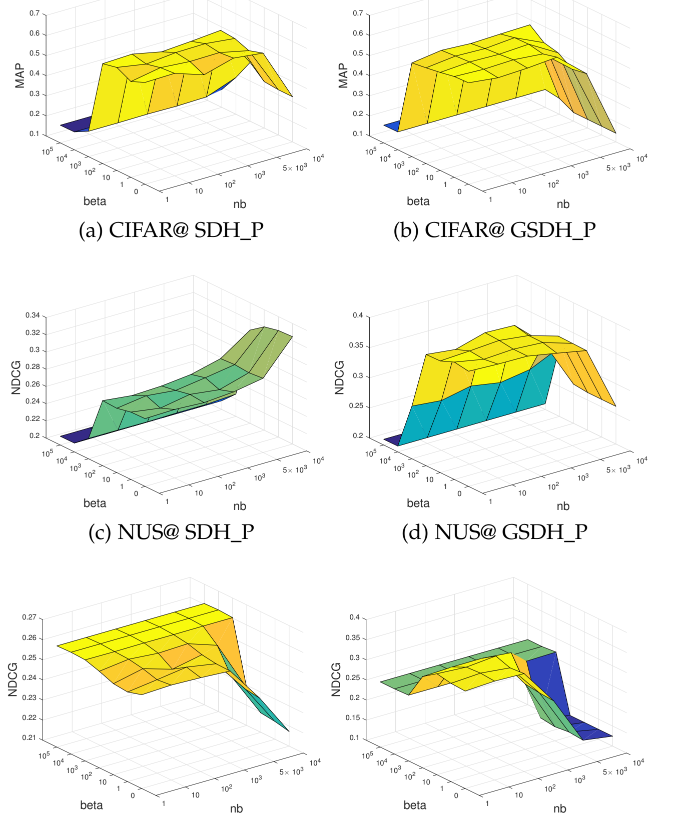
Theorem 4. *For the inner t-th iteration embedded in the outer l-th iteration, the problem in Eq. (9) can be reformulated as the*

X. Shi and L. Yang are with the J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA, e-mail: (xsshi2015@ufl.edu, lin.yang@bme.ufl.edu).

F. Xing and M. Sapkota is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA, e-mail: (f.xing@ufl.edu and manish.sapkota@gmail.com).

Z. Zhang is with the Department of Computer Science and Engineering, University of Florida, Gainesville, FL, USA, e-mail: (mr.zizhaozhang@gmail.com).

Z. Guo is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong, China, e-mail: (zhenhua.guo@sz.tsinghua.edu.cn).



(a) CIFAR@ SDH_P (b) CIFAR@ GSDH_P

(c) NUS@ SDH_P (d) NUS@ GSDH_P

(e) COCO@ SDH_P (f) COCO@ GSDH_P

Fig. A1. The influence of parameters n_b and β for SDH_P and GSDH_P on CIFAR-10, NUS-WIDE and COCO databases.

following problem:

$$\max_{\mathbf{H}_l} Tr\left\{\mathbf{H}_l((\gamma\mathbf{I}_m - \mathbf{H}_{l-1}^T\mathbf{H}_{l-1})\mathbf{H}_{l_{t-1}}^T + \lambda\mathbf{H}_{l-1}^T\mathbf{S})\right\}, s.t.\ \mathbf{H}_l \in \{-1,1\}^{n\times m}, \tag{10}$$

where $\mathbf{H}_l \in \mathbb{R}^{n\times m}$ *denotes binary codes* $\mathbf{H}$ *in the outer l-th iteration, and* $\mathbf{H}_{l_{t-1}}$ *represents the obtained binary codes* $\mathbf{H}$ *at the inner t-1-th iteration embedded in the outer l-th iteration.*

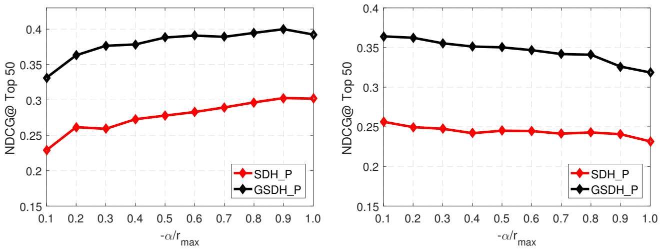
Proof. In Eq. (9), for the inner t-th iteration embedded in the outer l-th iteration, fixing $\mathbf{H}$ as $\mathbf{H}_{l_{t-1}}$, it is easy to obtain $\mathbf{P}_{l_t} = \mathbf{H}_{l_{t-1}}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T+\mathbf{\Gamma})^{-1}$. Substituting $\mathbf{P}_{l_t}$ into Eq. (9), it becomes:

$$\max_{\mathbf{H}_l} Tr\left\{\mathbf{H}_l\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T+\mathbf{\Gamma})^{-1}\mathbf{Z}\mathbf{H}_{l_{t-1}}^T\right\} + \frac{\lambda}{\gamma}Tr\left\{\mathbf{H}_l\mathbf{H}_{l-1}^T\mathbf{S}\right\},\ s.t.\ \mathbf{H}_l \in \{-1,1\}^{n\times m}. \tag{A2}$$

Based on Theorem 2 and its proof, there exists $\gamma\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T+\mathbf{\Gamma})^{-1}\mathbf{Z} = \gamma\mathbf{I}_m - \mathbf{H}_{l-1}^T\mathbf{H}_{l-1}$. Substituting it into Eq. (A2), the optimization problem becomes Eq. (10). Therefore, Theorem 4 is proved. □

Convergence Analysis

Empirically, when $n >> n_b$, the proposed algorithms can converge to at least a local optima, although they cannot be theoretically guaranteed to converge in all cases. Here, we explain why gradually updating each batch of binary codes is beneficial to the convergence of hash code learning.



(a) NUS-WIDE (b) COCO

Fig. A2. The influence of α for SDH_P and GSDH_P on NUS-WIDE and COCO databases.

In Eq. (10), with updating one batch of $\mathbf{H}$, i.e. $\mathbf{H}_b \in \{-1,1\}^{n_b\times m}$, Eq. (10) becomes:

$$\max_{\mathbf{H}_{bl}} Tr\left\{\mathbf{H}_{bl}((\gamma\mathbf{I}_m - \mathbf{H}_{l-1}^T\mathbf{H}_{l-1})\mathbf{H}_{bl_{t-1}}^T + \lambda\mathbf{H}_{l-1}^T\mathbf{S}_b)\right\}, s.t.\ \mathbf{H}_{bl} \in \{-1,1\}^{n_b\times m}, \tag{A3}$$

The hash code matrix $\mathbf{H}$ can be represented as $\mathbf{H} = \left[\mathbf{H}_b; \widetilde{\mathbf{H}}\right]$, where $\widetilde{\mathbf{H}} \in \{-1,1\}^{(n-n_b)\times m}$. Since $n >> n_b$, the objective of Eq. (A3) is determined by:

$$\max_{\mathbf{H}_{bl}} Tr\left\{\mathbf{H}_{bl}((\gamma\mathbf{I}_m - \widetilde{\mathbf{H}}_{l-1}^T\widetilde{\mathbf{H}}_{l-1})\mathbf{H}_{bl_{t-1}}^T + \lambda\widetilde{\mathbf{H}}_{l-1}^T\mathbf{S}_b)\right\}, s.t.\ \mathbf{H}_{bl} \in \{-1,1\}^{n_b\times m}, \tag{A4}$$

Based on Theorem 5, the inner loop can theoretically guarantee the convergence of the objective in Eq. (10), and thus the optimal solution $\mathbf{H}_{bl}^*$ of Eq. (A4) can be obtained by the inner loop. Then it has:

$$Tr\left\{\mathbf{H}_{bl}^*((\gamma\mathbf{I}_m - \widetilde{\mathbf{H}}_{l-1}^T\widetilde{\mathbf{H}}_{l-1})\mathbf{H}_{bl}^{*T} + \lambda\widetilde{\mathbf{H}}_{l-1}^T\mathbf{S}_b)\right\} \geq Tr\left\{\mathbf{H}_{bl-1}((\gamma\mathbf{I}_m - \widetilde{\mathbf{H}}_{l-1}^T\widetilde{\mathbf{H}}_{l-1})\mathbf{H}_{bl-1}^T + \lambda\widetilde{\mathbf{H}}_{l-1}^T\mathbf{S}_b)\right\} \tag{A5}$$

Because of $n >> n_b$, Eq. (A5) usually leads to

$$Tr\left\{\mathbf{H}_{bl}^*((\gamma\mathbf{I}_m - \widehat{\mathbf{H}}_l^T\widehat{\mathbf{H}}_l)\mathbf{H}_{bl}^{*T} + \lambda\widehat{\mathbf{H}}_l^T\mathbf{S}_b)\right\} \geq Tr\left\{\mathbf{H}_{bl-1}((\gamma\mathbf{I}_m - \mathbf{H}_{l-1}^T\mathbf{H}_{l-1})\mathbf{H}_{bl-1}^T + \lambda\mathbf{H}_{l-1}^T\mathbf{S}_b)\right\} \tag{A6}$$

where $\widehat{\mathbf{H}}_l = \left[\mathbf{H}_{bl}^*; \widetilde{\mathbf{H}}_{l-1}\right]$ and $\mathbf{H}_{l-1} = \left[\mathbf{H}_{bl-1}; \widetilde{\mathbf{H}}_{l-1}\right]$. Eq. (A6) suggests that when $n_b << n$, updating each batch matrix can usually make the objective of Eq. (10) gradually converge to at least a local optima.

Parameter Influence

Here, we first evaluate two essential parameters: batch size n_b and regularization parameter β, where n_b and β determine the batch number f and the parameter γ, respectively. Additionally, we also evaluate the effects of α on multi-label databases: NUS-WIDE and COCO. Similar to previous experiments, we uniformly select 10K training and 1K query images from CIFAR-10, NUS-WIDE and COCO databases, and then encode each image features into 16-bit binary codes. Fig. A1 presents the influence of $n_b \in \{1, 10, 10^2, 10^3, 5\times 10^3, 10^4\}$ and $\beta \in \{0, 1, 10, 10^2, 10^3, 10^4, 10^5\}$ on retrieval performance in term of MAP over top 500 returned samples on CIFAR-10 and NDCG over top 50 retrieved samples on NUS-WIDE and COCO. It suggests that both SDH_P and GSDH_P can obtain the best or sub-best retrieval performance when $n_b \in [1, 100]$ and $\beta \in [0, 100]$ on all the three databases. Because the selection of α depends on r_{max}, Fig. A2 displays the influence of $\frac{-\alpha}{r_{max}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ on NDCGs with images selected from NUS-WIDE and COCO databases. It shows that the NDCGs of SDH_P and GSDH_P increase on NUS-WIDE when $\frac{-\alpha}{r_{max}} \in [0.1, 0.9]$, while their NDCGs almost decrease on the COCO database with $\frac{-\alpha}{r_{max}} \in [0.1, 1.0]$. Similar findings can be observed at other bits, we do not show them for brevity. In our experiments, without loss of generality, we empirically set $\beta = 10$, $n_b = 100$ and $\alpha = -\frac{r_{max}}{2}$ for the proposed algorithms.

References

[1] B. Neyshabur, N. Srebro, R. R. Salakhutdinov, Y. Makarychev, and P. Yadollahpour, "The power of asymmetry in binary hashing," in *Advances in Neural Information Processing Systems*, 2013, pp. 2823–2831.

[2] X. Shi, F. Xing, K. Xu, M. Sapkota, and L. Yang, "Asymmetric discrete graph hashing," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2541–2547.