1. Convergence Analysis

We briefly analyze that the continuation optimization in Algorithm 1 will decrease the loss of HashNet (4) in each stage and in each iteration until converging to HashNet with sign activation function that generates exactly binary codes.

Let \( L_{ij} = w_{ij} \log (1 + \exp (\alpha (h_i, h_j))) - \alpha s_{ij} (h_i, h_j) \) and \( L = \sum_{s_{ij} \in S} L_{ij} \), where \( h_i \in \{-1, +1\}^K \) are binary hash codes. Note that when optimizing HashNet by continuation in Algorithm 1, network activation in each stage \( t \) is \( g = \tanh(\beta_t z) \), which is continuous in nature and will only become binary when convergence \( \beta_t \to \infty \). Denote by \( J_{ij} = w_{ij} \log (1 + \exp (\alpha (g_i, g_j))) - \alpha s_{ij} (g_i, g_j) \) and \( J = \sum_{s_{ij} \in S} J_{ij} \) the true loss we optimize in Algorithm 1, where \( g_i \in \mathbb{R}^K \) and note that \( h_i = \text{sgn}(g_i) \). We will show that HashNet loss \( L(h) \) descends when minimizing \( J(g) \).

Lemma 1. Denote by \( h_i = \text{sgn}(g_i) \), \( h_i' = \text{sgn}(g_i') \), then
\[
\left\{ \begin{array}{ll}
(h'_i, h'_j) \geq (h_i, h_j), & s_{ij} = 1, \\
(h'_i, h'_j) \leq (h_i, h_j), & s_{ij} = 0.
\end{array} \right.
\] (3)

Proof. Since \( (h_i, h_j) = \sum_{k=1}^K h_{ik} h_{jk} \), Lemma 1 can be proved by verifying that \( h'_{ik} h'_{jk} \geq h_{ik} h_{jk} \) if \( s_{ij} = 1 \) and \( h'_{ik} h'_{jk} \leq h_{ik} h_{jk} \) if \( s_{ij} = 0 \). Thus \( h'_{ik} \geq h_{ik} \), \( h'_{jk} \leq h_{jk} \).

Case 1. \( s_{ij} = 0 \).

(1) If \( g_{ik} < 0, g_{jk} > 0 \), then \( \frac{\partial J}{\partial g_{jk}} > 0 \). Thus \( h'_{ik} \geq h_{ik} = -1 \), \( h'_{jk} \geq h_{jk} = 1 \). And we have \( h'_{ik} h'_{jk} = -1 = h_{ik} h_{jk} \).

(2) If \( g_{ik} > 0, g_{jk} < 0 \), then \( \frac{\partial J}{\partial g_{jk}} < 0 \). Thus \( h'_{ik} \geq h_{ik} = 1 \), \( h'_{jk} \leq h_{jk} = -1 \). And we have \( h'_{ik} h'_{jk} = -1 = h_{ik} h_{jk} \).

(3) If \( g_{ik} < 0, g_{jk} < 0 \), then \( \frac{\partial J}{\partial g_{ik}} < 0 \). Thus \( h'_{ik} \geq h_{ik} = -1 \), \( h'_{jk} \geq h_{jk} = 1 \). So \( h'_{ik} \) and \( h'_{jk} \) may be either +1 or −1 and we have \( h'_{ik} h'_{jk} \leq 1 = h_{ik} h_{jk} \).

(4) If \( g_{ik} > 0, g_{jk} > 0 \), then \( \frac{\partial J}{\partial g_{ik}} > 0 \). Thus \( h'_{ik} \leq h_{ik} = 1 \), \( h'_{jk} \leq h_{jk} = 1 \). So \( h'_{ik} \) and \( h'_{jk} \) may be either +1 or −1 and we have \( h'_{ik} h'_{jk} \leq 1 = h_{ik} h_{jk} \).

Case 2. \( s_{ij} = 1 \). It can be proved similarly as Case 1.

Theorem 2. Loss \( L \) decreases when optimizing loss \( J(g) \) by the stochastic gradient descent (SGD) within each stage.

Proof. The gradient of loss \( L \) w.r.t. hash codes \( (h_i, h_j) \) is
\[
\frac{\partial L}{\partial (h_i, h_j)} = w_{ij} \alpha \left( \frac{1}{1 + \exp (-\alpha (h_i, h_j))} - s_{ij} \right) g_{jk},
\] (4)
We observe that
\[
\left\{ \begin{array}{ll}
\frac{\partial L}{\partial (h_i, h_j)} \leq 0, & s_{ij} = 1, \\
\frac{\partial L}{\partial (h_i, h_j)} \geq 0, & s_{ij} = 0.
\end{array} \right.
\] (5)

By substituting Lemma 1: if \( s_{ij} = 1 \), then \( (h'_i, h'_j) \geq (h_i, h_j) \), and thus \( L(h'_i, h'_j) \leq L(h_i, h_j) \); if \( s_{ij} = 0 \), then \( (h'_i, h'_j) \leq (h_i, h_j) \), and thus \( L(h'_i, h'_j) \leq L(h_i, h_j) \).