# Note on the Number of Balanced Independent Sets in the Hamming Cube

Jinyoung Park*

Department of Mathematics
Stanford University
Stanford, U.S.A.

jinypark@stanford.edu

Submitted: Jun 3, 2021; Accepted: Feb 4, 2022; Published: May 20, 2022



**Abstract**

Let $Q_d$ be the $d$-dimensional Hamming cube and $N = |V(Q_d)| = 2^d$. An independent set $I$ in $Q_d$ is called balanced if $I$ contains the same number of even and odd vertices. We show that the logarithm of the number of balanced independent sets in $Q_d$ is

$$(1 - \Theta(1/\sqrt{d}))N/2.$$

The key ingredient of the proof is an improved version of "Sapozhenko's graph container lemma."

**Mathematics Subject Classifications:** 05C69

## 1 Introduction

For a bipartite graph $G = X \coprod Y$ and an independent set $I$ in $G$, $I$ is said to be *balanced* if $|I \cap X| = |I \cap Y|$. We use $\mathrm{bis}(G)$ for the number of balanced independent sets of a graph $G$.

Write $Q_d$ for the $d$-dimensional Hamming cube and $N$ for $|V(Q_d)|(= 2^d)$. In this note we prove the following result on $\log \mathrm{bis}(Q_d)$. (All log's in this paper are in base 2.)

**Theorem 1.**

$$\log \mathrm{bis}(Q_d) = (1 - \Theta(1/\sqrt{d}))N/2. \tag{1}$$

*The author was supported directly by NSF grant DMS-1926686 and indirectly by NSF grant CCF-1900460.

It is easy to see that the rhs of (1) is a lower bound: Barber [2] showed that the size of a maximum balanced independent set in $Q_d$ is

$$\begin{cases} 2^{d-1} - 2\binom{d-2}{(d-2)/2} & \text{if } d \text{ is even;} \\ 2^{d-1} - \binom{d-1}{(d-1)/2} & \text{if } d \text{ is odd,} \end{cases}$$

and collecting balanced subsets of a maximum BIS gives the lower bound. So the main task of this paper is to show the rhs of (1) is also an upper bound.

*Background.* The asymptotics for the number of (ordinary) independent sets in $Q_d$, $i(Q_d)$, was first given by Korshunov and Sapozhenko [11]:

**Theorem 2.**

$$i(Q_d) \sim 2\sqrt{e}2^{N/2}. \tag{2}$$

(The above asymptotics are substantially refined by Jenssen and Perkins in [8].) Note that the rhs of (2) is an asymptotic lower bound on $i(Q_d)$: writing $Q_d = \mathcal{E} \coprod \mathcal{O}$ (a few basic definitions are recalled below), any subset of $\mathcal{E}$ or $\mathcal{O}$ is an independent set, from which we have $2 \cdot 2^{N/2} - 1$ independent sets. The extra factor $\sqrt{e}$ reflects the contribution of independent sets most of whose vertices are even (odd, resp.), together with a (very) small number of odd (even, resp.) vertices. (See e.g. [4] for a more detailed description on this lower bound construction.)

Thus Theorem 2 implies that $i(Q_d)$ is asymptotically equal to this lower bound, and in particular, this implies all but a negligible fraction of independent sets in $Q_d$ are highly unbalanced. The natural problem of estimating $\text{bis}(Q_d)$ was suggested by T. Helmuth, M. Jenssen, and W. Perkins [6], and Theorem 1 answers this question at the level of asymptotics of the logarithm.

After the first draft of this paper was prepared, it was communicated to the author that the problem of estimating $\text{bis}(Q_d)$ was also considered by Galvin and Tetali [5] in their study of the mixing time of hard core model dynamics on $Q_d$, who obtained the weaker upper bound $\exp_2[(1 - \Omega(1/(\sqrt{d}\log d)))N/2]$. We point out that Galvin and Tetali take a roughly similar approach to the present approach, the key difference being that their Lemma 3.4 is weaker (by a factor of $\log d$) than our Lemma 11.

The key ingredient of the proof of Theorem 1 is Lemma 11, an improvement (see Remark 12) of "Sapozhenko's graph container lemma" from [12]. Sapozhenko's lemma and its variants have played a key role in resolving a number of asymptotic enumeration problems on the Hamming cube and related structures, e.g. [11, 8, 3, 9, 10, 7, 1]. The current improved version of the lemma is implicitly proved in [10, Lemma 6.3], but we give a self-contained proof in Section 3 to provide a convenient reference for future work.

It would be of interest finding finer asymptotics for $\text{bis}(Q_d)$. For example,

**Question 3.** What is optimal $C$ for which

$$\log \text{bis}(Q_d) = (1 - C/\sqrt{d})N/2?$$

Or, even more ambitiously,

**Question 4.** What is the asymptotics for $\operatorname{bis}(Q_d)$?

*Definitions.* We use $Q_d$ for the $d$-dimensional Hamming cube: that is, $V = V(Q_d)$ is the collection of binary strings of length $d$, and two vertices are adjacent iff they differ in exactly one coordinate. A vertex $v$ is even (odd, resp.) if $v$ contains an even (odd, resp.) number of 1's. We use $\mathcal{E}$ ($\mathcal{O}$, resp.) for the set of even (odd, resp.) vertices in $Q_d$ (so $Q_d = \mathcal{E} \coprod \mathcal{O}$). The collection of balanced independent sets in $Q_d$ is denoted by $\mathcal{B} = \mathcal{B}(d)$, and $I$ always denotes a balanced independent set.

As usual, $N(v)$ is the set of neighbors of $v$, and $N(A)$ is the set of vertices that are adjacent to at least one vertex in $A$. We use $[A]$ for the *closure* of $A$, namely, $[A] = \{v \in V : N(v) \subseteq N(A)\}$.

Finally, we refer to the logarithm of the number of possibilities for a choice as the *cost* of that choice.

*Outline.* In Section 2 we recall some basic tools. The main lemma (Lemma 11) and Theorem 1 are proved in Section 3 and Section 4 respectively.

## 2 Tools

The following is a well-known fact about the sum of binomial coefficients.

**Proposition 5.** *For any fixed $\alpha \in [0, 1/2]$ and $n \in \mathbb{Z}^+$,*

$$\sum_{i \leqslant \alpha n} \binom{n}{i} \leqslant 2^{H(\alpha)n},$$

*where $H(\alpha) := -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function.*

For a positive integer $m$, a *composition* of $m$ is a sequence $(a_1, \ldots, a_s)$ of positive integers summing to $m$. Recall the following basic fact:

**Proposition 6.** *The number of compositions of $m$ is $2^{m-1}$ and the number with at most $b \leqslant m/2$ parts is*

$$\sum_{i \leqslant b} \binom{m-1}{i} \leqslant \exp_2[b \log(em/b)].$$

Say $A \subseteq V$ is 2-*linked* if for any $u, v \in A$, there are vertices $u = u_0, u_1, \ldots, u_l = v$ in $A$ such that for each $i \in [l]$, $u_{i-1}$ and $u_i$ are at distance at most 2 in $Q_d$. The 2-*components* of $A$ are its maximal 2-linked subsets.

**Proposition 7** ([3], Lemma 1.6)**.** *For each fixed $k$, the number of $k$-linked subsets of $V$ of size $x$ containing some specified vertex is at most $2^{O(x \log d)}$.*

The next two results recall standardish isoperimetric inequalities for $Q_d$. Recall that $N = |V(Q_d)| = 2^d$.

**Proposition 8** ([4], Claim 2.5)**.** *For $A \subseteq \mathcal{E}$ (or $\mathcal{O}$) with $|A| \leqslant N/4$,*

$$\frac{|N(A)| - |A|}{|N(A)|} = \Omega(1/\sqrt{d}).$$

**Proposition 9** ([4], Lemma 2.6)**.** *For $A \subseteq \mathcal{E}$ (or $\mathcal{O}$),*

$$\textit{if } |A| < d^{O(1)}, \textit{ then } |A| \leqslant O(1/d)|N(A)|.$$

The next lemma recalls what we need from [12], and follows from Lemmas 5.3-5.5 in the excellent exposition due to Galvin [4]. For $A$ in the statement, we use $G = N(A)$ and $t = |G| - |[A]|$.

**Lemma 10.** *For $q, g \in \mathbb{Z}^+$, $q \leqslant N/4$, $g \geqslant d^4$ and*

$$\mathcal{G}(q, g) = \{A \subseteq \mathcal{E} : A \textit{ is 2-linked}, |[A]| = q \textit{ and } |G| = g\},$$

*there exist a family $\mathcal{W} = \mathcal{W}(q, g) \subseteq 2^{\mathcal{E}} \times 2^{\mathcal{O}}$ with*

$$|\mathcal{W}| = 2^{O(t \log^2 d/\sqrt{d})} \tag{3}$$

*and a function $\Phi = \Phi_{q,g} : \mathcal{G} \to \mathcal{W}$ such that for each $A \in \mathcal{G}$, $(S, F) := \Phi(A)$ satisfies:*

*(a)* $S \supseteq [A], F \subseteq G$;

*(b)* $|S| \leqslant |F| + O(t/(\sqrt{d} \log d))$.

## 3 Main Lemma

In this section we prove the following key lemma.

**Lemma 11.** *For $q, g$, and $\mathcal{G}(q, g)$ as in Lemma 10,*

$$\log |\mathcal{G}(q, g)| \leqslant g - \Omega(t).$$

*Remark* 12. For comparison, Sapozhenko's original graph container lemma says

$$\log |\mathcal{G}(q, g)| \leqslant g - \Omega(t/\log d), \tag{4}$$

so the main contribution of Lemma 11 is to improve the $\Omega(t/\log d)$-term in the rhs of (4) to $\Omega(t)$. This improvement plays a crucial role in the current work: the bound in (4) would give a weaker bound, $2^{g-\Omega(N/(\sqrt{d}\log d))}$, in (10).

*Proof of Lemma 11.* Given $q$ and $g$, Lemma 10 gives $\mathcal{W} = \mathcal{W}(q, g)$ at cost $O(t \log^2 d/\sqrt{d})$. So it suffices to show that given $(S, F) \in \mathcal{W}$, the cost of specifying $A \in \Phi^{-1}(S, F)$ is at most $g - \Omega(t)$.

Let $\gamma \in (0, 1)$ be a constant TBD. (We don't try to optimize $\gamma$.)

Case 1. If $|S| < g - \gamma t$, then we specify $A$ by picking a subset of $S$, which costs $|S| = g - \gamma t$.

Case 2. If $|S| \geqslant g - \gamma t$, then we first fix an arbitrary closed $A^* \in \Phi^{-1}(S, F)$ (we can simply pick *any* member of $\Phi^{-1}(S, F)$ and take its closure). Note that this choice is free, and $|A^*| = q$ by the definition of $\mathcal{G}(q, g)$.

The crucial observation is that (letting $G^* = N(A^*)$)

$$(G^* \setminus G, G \setminus G^*) \text{ determines } (G, [A]).$$

In what follows we first specify $G^* \setminus G$ and $G \setminus G^*$ from which we have $[A]$, and then specify $A \subseteq [A]$.

We first bound the cost of $G^* \setminus G$. Since $G^* \setminus G \subseteq G^* \setminus F$, the cost of $G^* \setminus G$ is at most (using Lemma 10 (b))

$$|G^* \setminus F| = |G^*| - |F| \leqslant |G| - |S| + O(t/(\sqrt{d} \log d)) \leqslant (1 + o(1))\gamma t. \tag{5}$$

Next, we bound the cost of $G \setminus G^*$. Observe that

$$G \setminus G^* = N([A] \setminus A^*) \setminus G^*,$$

because each $x \in G \setminus G^*$ has a neighbor in $[A]$ and none in $A^*$. So we may specify $G \setminus G^*$ by specifying a $Y \subseteq [A] \setminus A^* \subseteq S \setminus A^*$ with $G \setminus G^* = N(Y) \setminus G^*$. Moreover, we only need $Y \subseteq S \setminus A^*$ of size at most $|G \setminus G^*| \leqslant g - |F| \leqslant (1 + o(1))\gamma t$, by letting $Y$ contain one neighbor of $x$ for each $x \in G \setminus G^*$.

Now, since (again using Lemma 10 (b))

$$|S \setminus A^*| = |S| - q \leqslant |F| + o(t) - q \leqslant g - q + o(t) \leqslant (1 + o(1))t,$$

the cost of specifying $Y$ from $S \setminus A^*$ is at most

$$\log \binom{(1 + o(1))t}{(1 + o(1))\gamma t} \leqslant (1 + o(1))H(\gamma)t \tag{6}$$

where $H(\cdot)$ is the binary entropy function. Finally, once we have $[A]$, we specify $A$ by picking a subset of $[A]$, which costs

$$q = g - t. \tag{7}$$

Summing up (5), (6), and (7), we bound the total cost for Case 2 by

$$(1 + o(1))\gamma t + (1 + o(1))H(\gamma)t + (g - t). \tag{8}$$

Now, choose $\gamma$ so that (8) is less than (say) $g - t/2$, and the lemma follows. $\square$

## 4 Proof of Theorem 1

We show that the rhs of (1) is an upper bound on $\log \mathrm{bis}(Q_d)$. We first dispose of the minor cost for small independent sets.

**Proposition 13.** *There is a constant $\alpha \in (0, 1/2)$ such that*

$$|\{I \in \mathcal{B} : |I| \leqslant \alpha N\}| = 2^{(1-\Omega(1))N/2}. \tag{9}$$

*Proof.* The lhs of (9) is at most (with plenty of room)

$$\left[\sum_{0\leqslant k\leqslant \alpha N/2} \binom{N/2}{k}\right]^2 \leqslant 2^{H(\alpha)N}$$

(the inequality uses Proposition 5), and the rhs is less than $2^{(1-\Omega(1))N/2}$ for small enough constant $\alpha$. □

Let $\mathcal{B}' = \{I \in \mathcal{B} : |I| > \alpha N\}$ where $\alpha$ is the constant in Proposition 13. A natural way to specify a balanced independent set $I$ is to choose a set $A \subseteq \mathcal{E}$ and a set $B \subseteq \mathcal{O} \setminus N(A)$ so that $|A| = |B|$ (and take $I = A \cup B$). Moreover, since $[I \cap \mathcal{E}]$ and $[I \cap \mathcal{O}]$ have no edges between them, $A$ and $B$ must satisfy $\min\{|[A]|, |[B]|\} \leqslant N/4$ (because $|N(X)| \geqslant |X|$, $\forall X \subseteq \mathcal{E}$ or $\mathcal{O}$). Thus, $|\mathcal{B}'|$ is at most

$$\begin{aligned} &2 \times \sum_{g>\alpha N/2} \sum_{\substack{A\subseteq\mathcal{E}:|N(A)|=g\\ |A|\geqslant\alpha N/2\\ |[A]|\leqslant N/4}} |\{B \subseteq \mathcal{O} \setminus N(A)\}| \\ &= 2^{N/2+1} \sum_{g>\alpha N/2} 2^{-g} |\{A \subseteq \mathcal{E} : |N(A)| = g, |A| \geqslant \alpha N/2, |[A]| \leqslant N/4\}|. \end{aligned}$$

Our main task is to show that

$$\text{given } g = \Omega(N),\ |\{A \subseteq \mathcal{E} : |N(A)| = g, |A| \geqslant \alpha N/2, |[A]| \leqslant N/4\}| \leqslant 2^{g-\Omega(N/\sqrt{d})}, \tag{10}$$

from which it follows that (with Proposition 13)

$$|\mathcal{B}| \leqslant 2^{(1-\Omega(1))N/2} + 2^{N/2+1} \sum_{g>\alpha N/2} 2^{-\Omega(N/\sqrt{d})} = 2^{(1-\Omega(1/\sqrt{d}))N/2}.$$

In the rest of the paper, we show (10). In what follows, we always assume that $g$ and $A$ satisfy the restrictions in (10).

*Notation.*

Recall that a 2-component of $A$ is a maximal 2-linked subset of $A$ (see Section 2).

- $A_i$'s: 2-components of $A$.

- $G_i = N(A_i)$, $G = \cup_i G_i = N(A)$.
- $g_i = |G_i|$, $a_i = |A_i|$, $q_i = |[A_i]|$, $t_i = g_i - q_i$.
- $c(A) = \sum_i q_i$ (note that $|A| \leqslant c(A) \leqslant |[A]|$).
- $g = |G|$ $(= \sum_i g_i)$.
- $A_i$ (or simply $i$) is $\begin{cases} \textit{isolated} & \text{if } a_i = 1 \text{ (equiv. } g_i = d\text{);} \\ \textit{small} & \text{if } A_i \text{ is not isolated and } g_i < d^4\text{;} \\ \textit{large} & \text{otherwise.} \end{cases}$

Note that the classification in the above bullet point is entirely determined by $g_i$.

By Proposition 9,

$$\sum \{a_i\colon\ i \text{ isolated or small}\} = O(N/d),$$

so in particular, we have (since $|A| = \Omega(N)$)

$$\sum \{g_i\colon\ i \text{ large}\} > \sum \{a_i\colon\ i \text{ large}\} = \Omega(N). \tag{11}$$

*Proof of* (10). Observe that (since $|A| \leqslant c(A) \leqslant |[A]|$) it suffices to show that given $g$ as in (10) and $q$ with $\alpha N/2 \leqslant q \leqslant \min\{N/4, g\}$, the number of $A$'s in $\mathcal{E}$ with $c(A) = q$ and $|N(A)| = g$ is at most $2^{g - \Omega(N/\sqrt{d})}$ (since then summing this up over all $q$'s gives (10)).

Given $q$ and $g$, we first decompose $(q, g)$ into $\{(q_i, g_i)'s\}$ so that $\sum_i q_i = q$ and $\sum_i g_i = g$ (and then specify $A_i$'s satisfying $|[A_i]| = q_i$ and $|G_i| = g_i$). The number of elements in a decomposition $\{(q_i, g_i)'s\}$ is at most $g/d$, so Proposition 6 bounds the cost of the $g_i$'s by $(g/d)\log(ed)$ and that of the $q_i$'s by

$$\begin{cases} (g/d)\log(ed) & \text{if } (g >)\ q > 2g/d; \\ 2g/d & \text{if } q \leqslant 2g/d. \end{cases}$$

Therefore, the total cost of the specification of $q_i$'s and $g_i$'s is at most

$$O(g \log d/d). \tag{12}$$

□

**Lemma 14.** *Given $(q_i, g_i)$, if $i$ is isolated or small, then the cost of $A_i$ with $|[A_i]| = q_i$ and $|G_i| = g_i$ is at most $g_i$.*

*Proof.* The cost of an isolated $i$ is at most

$$\log(N/2) = d - 1 \leqslant g_i.$$

For a small $i$, we use Proposition 7 to bound the cost of $[A_i]$ by

$$\log(N/2) + O(q_i \log d)$$

($\log(N/2)$ is the cost for the 'specified vertex' in Proposition 7). Once we have $[A_i]$ we specify $A_i$ by choosing each subset of $[A_i]$, which costs $q_i$. Therefore, the total cost for small $i$'s is

$$\log(N/2) + O(q_i \log d) + q_i \leqslant g_i,$$

where the inequality follows from the fact that $g_i/2 \geqslant d-1$ and Proposition 9. □

Finally, given $(q_i, g_i)$ such that $i$ is large, Lemma 11 bounds the cost for $A_i$ with $|[A_i]| = q_i$ and $|G_i| = g_i$ by

$$g_i - \Omega(t_i) \tag{13}$$

(here we need the assumption that $(q_i \leqslant)$ $q \leqslant N/4$ to apply Lemma 11).

Summing up the costs in (12), Lemma 14 and (13), we have the cost for $A$ at most

$$O(g \log d/d) + g - \sum\{\Omega(t_i)\colon\ i \text{ large}\}. \tag{14}$$

Now Proposition 8 gives $t_i = \Omega(g_i/\sqrt{d})$ for all $i$, so by (11) we bound (14) by

$$g - \Omega(N/\sqrt{d}).$$


### Acknowledgements

The author is grateful to Matthew Jenssen for insightful conversations and helpful comments on the first draft of this paper.


## References


[1] J. Balogh, R. Garcia and L. Li, Independent sets in the middle two layers of Boolean lattice, *J. Combin. Theory Ser. A* 178 (2021), Paper No. 105341.

[2] B. Barber, A note on balanced independent sets in the cube, *Australas. J. Combin.* 52 (2012), 205-207.

[3] D. Galvin, On homomorphisms from the Hamming cube to Z, *Israel J. Math.* 138 (2003), 189-213.

[4] D. Galvin, Independent sets in the discrete hypercube, arXiv:1901.01991.

[5] D. Galvin and P. Tetali, Slow mixing of Glauber dynamics for the hard-core model on regular bipartite graphs, *Random Structures Algorithms* 28 (2006), no. 4, 427-443.

[6] M. Jenssen, personal comunication

[7] M. Jenssen and P. Keevash, Homomorphisms from the torus, arXiv:2009.08315.

[8] M. Jenssen and W. Perkins, Independent sets in the hypercube revisited, *J. Lond. Math. Soc.* (2) 102 (2020), no. 2, 645-669.

[9] J. Kahn and J. Park, The number of 4-colorings of the Hamming cube, *Israel J. Math.* 236 (2020), no. 2, 629-649.

[10] J. Kahn and J. Park, The number of maximal independent sets in the Hamming cube, *Combinatorica* (2022).

[11] A.D. Korshunov and A.A. Sapozhenko, The number of binary codes with distance 2. (Russian) *Problemy Kibernet.* No. 40 (1983), 111-130.

[12] A.A. Sapozhenko, On the number of connected subsets with given cardinality of the boundary in bipartite graphs. (Russian) *Metody Diskret. Analiz.* No. 45 (1987), 42-70.