# Effective Batch Size

Zhi Wang

September 26, 2021

## Contents



## 1 Choosing all labeled items

Given a box with a $i$ labeled balls, you pick one of them and then put it back. Do it for $j$ times. How many ways are there such that *all* of them are picked at least once?

**Definition.** Denote the answer to the question as $F(i, j)$. This is only defined for positive integers $i, j$ such that $i > 0, j \geq i$.

Let's start from $i = 1$. Obviously,

$$F(1, j) = 1, \tag{1}$$

for all $j \geq 1$.

For $i > 1$, it is the number of all possibilities $i^j$ minus the cases where exactly $l \in (0, i)$ balls are picked:

$$F(i,j) = i^j - \sum_{l=1}^{i-1} \binom{i}{l} F(l,j), \tag{2}$$

where $\binom{n}{k}$ is the binomial coefficients $\frac{n!}{k!(n-k)!}$, for any non-negative integers $n, k$. This calculates how many ways to choose $k$ balls out of $n$.

**Proposition 1.1.** *The number of ways choosing $i \in \mathbb{Z}^+$ labeled items $j \in [i, +\infty) \cap \mathbb{Z}^+$ times such that all of them are chosen at least once is given by*

$$F(i,j) = \sum_{k=0}^{i-1} (-1)^k \binom{i}{k} (i-k)^j . \tag{3}$$

*Proof.* For $i = 1$,

$$\sum_{k=0}^{i-1} (-1)^k \binom{i}{k} (i-k)^j = (-1)^0 \binom{1}{0} (1-0)^j = 1,$$

satisfying (1).

If the property holds up to $i$, then for $i+1$, we get from (2)

$$\begin{aligned} F(i+1,j) &= (i+1)^j - \sum_{l=1}^{i} \binom{i+1}{l} F(l,j) \\ &= (i+1)^j - \sum_{l=1}^{i} \binom{i+1}{l} \sum_{k=0}^{l-1} (-1)^k \binom{l}{k} (l-k)^j \\ &= (i+1)^j - \sum_{p=1}^{i} \sum_{k=0}^{i-p} (-1)^k \binom{i+1}{p+k} \binom{p+k}{k} p^j, \end{aligned} \tag{4}$$

where $p = l - k$.

Since $k \geq 0, l \leq i$, $p = l - k \leq i$. $\because k \leq l-1, \therefore p \geq 1$. $\because l = p + k \leq i, \therefore k \leq i - p$. Both summations have $\frac{i(i+1)}{2}$ terms.

The ratio of the $(k+1)$-th term to the $k$-th term of $\sum_{k=0}^{i-p} (-1)^k \binom{i+1}{p+k} \binom{p+k}{k}$ is

$$-\frac{i+1-p-k}{p+k+1} \frac{p+k+1}{k+1} = -\frac{i+1-p-k}{k+1}.$$

Therefore, the coefficient is calculated as

$$\begin{aligned}
&\sum_{k=0}^{i-p}(-1)^k\binom{i+1}{p+k}\binom{p+k}{k}\\
&=\binom{i+1}{p}\left[1-\frac{i-p+1}{1}+\frac{(i-p+1)(i-p)}{1\cdot 2}\cdots\right]\\
&=\binom{i+1}{p}\sum_{m=0}^{i-p}(-1)^m\binom{i-p+1}{m}\\
&=\binom{i+1}{p}\left[\sum_{m=0}^{i-p+1}(-1)^m\binom{i-p+1}{m}-(-1)^{i-p+1}\binom{i-p+1}{i-p+1}\right]\\
&=\binom{i+1}{p}\left[(1+(-1))^{i-p+1}-(-1)^{i-p+1}\right]\\
&=(-1)^{i-p}\binom{i+1}{p},
\end{aligned}\tag{5}$$

where binomial theorem is applied ($i-p+1\geq k+1\geq 1$). Plug (5) into (4), we get

$$F(i+1,j)=(i+1)^j+\sum_{p=1}^{i}(-1)^{i-p+1}\binom{i+1}{p}p^j.\tag{6}$$

On the other hand, set $p=i+1-k$ in (3):

$$\begin{aligned}
F(i+1,j)&=(i+1)^j+\sum_{k=1}^{i}(-1)^k\binom{i+1}{k}(i+1-k)^j\\
&=(i+1)^j+\sum_{p=1}^{i}(-1)^{i+1-p}\binom{i+1}{i+1-p}p^j\\
&=(i+1)^j+\sum_{p=1}^{i}(-1)^{i-p+1}\binom{i+1}{p}p^j.
\end{aligned}\tag{7}$$

Comparing (6) with (7), we finish the proof. □

*Remark.* This can also derived from the inclusion–exclusion principle.

*Remark.* The equation (5) can be derive alternatively

$$\begin{aligned}\binom{i+1}{p+k}\binom{p+k}{k} &= \frac{(i+1)!}{(p+k)!(i+1-p-k)!}\frac{(p+k)!}{k!p!} \\ &= \frac{(i+1)!}{p!(i+1-p)!}\frac{(i+1-p)!}{k!(i+1-p-k)!} \\ &= \binom{i+1}{p}\binom{i-p+1}{k}.\end{aligned} \tag{8}$$

# 2 The $F$ function

**Definition.** Extending the previous question, we define a new function $F\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$ given by

$$F(i,j) \coloneqq \sum_{k=0}^{i}(-1)^k\binom{i}{k}(i-k)^j\,, \tag{9}$$

assuming $0^0 = 1$.

**Proposition 2.1.**

$$F(0,0) = 1. \tag{10}$$

*Proof.*

$$F(0,0) = (-1)^0\frac{0!}{0!(0-0)!}(0-0)^0 = 0^0 = 1.$$

□

**Proposition 2.2.** *If $i \neq 0$,*

$$F(i,0) = 0. \tag{11}$$

*Proof.*

$$F(i,0) = \sum_{k=0}^{i}(-1)^k\binom{i}{k} = (1+(-1))^i = 0.$$

The binomial theorem only holds for $i \neq 0$. □

**Proposition 2.3.** *If $j \neq 0$,*

$$F(0,j) = 0. \tag{12}$$

*Proof.* Since 0 to the $j$-th power is still 0 when $j \neq 0$,

$$F(0,j) = (-1)^0 \frac{0!}{0!(0-0)!} (0-0)^j = 0.$$

□

**Proposition 2.4.** *If $i \in \mathbb{Z}^+, j \in \mathbb{Z}^+$,*

$$F(i,j) = \sum_{k=0}^{i-1} (-1)^k \binom{i}{k} (i-k)^j, \tag{13}$$

*therefore the definition of $F$ is consistent with previous section.*

*Proof.* When $k = i$, the term becomes $(-1)^i \frac{i!}{i!0!} (i-i)^j$, which is just zero if $j \neq 0$. □

**Proposition 2.5.**

$$F(i,j) = 0, \forall j \in [0,i) \cap \mathbb{N}. \tag{14}$$

*Proof.* From (11) we know this is true for all $i > j$ when $j = 0$.

If this is true for all $i > j$ when $j$ is a natural number smaller than a positive integer $k$, we need to prove that this is also true for all $i > j$ when $j = k$.

(7) should hold for all $i > 0$ and $j > 0$. Therefore,

$$F(i+1,j) - (i+1)^j = \sum_{p=1}^{i} (-1)^{i-p+1} \binom{i+1}{p} p^j. \tag{15}$$

For $i = 1$, we have

$$F(2,j) = 2^j + (-1)^i (i+1) = 2^j - 2, \tag{16}$$

which give zero if $j = 1$. For $i > 1$, since $p$ is nonzero,

$$\begin{aligned}
&F(i+1,j) - (i+1)^j \\
&= \sum_{p=1}^{i} (-1)^{i+1-p} \frac{(i+1)!}{(p-1)!(i+1-p)!} p^{j-1} \\
&= (i+1) \left[ (-1)^i + \sum_{p=2}^{i} (-1)^{i+1-p} \frac{i!}{(p-1)!(i+1-p)!} \sum_{l=0}^{j-1} \binom{j-1}{l} (p-1)^l \right] \\
&= (i+1) \left[ (-1)^i + \sum_{l=0}^{j-1} \binom{j-1}{l} \sum_{p'=1}^{i'} (-1)^{i'-p'+1} \binom{i'+1}{p'} (p')^l \right]
\end{aligned} \tag{17}$$

where $p' = p - 1, i' = i - 1$. $\because i > 1, \therefore i' > 0$.

When $j = 0$, the right hand side of (15) evaluates into

$$\begin{aligned} &\left[\sum_{p=0}^{i+1}(-1)^{i-p+1}(-1)^{2p}\binom{i+1}{p}\right] - (-1)^{i+1} - 1 \\ &= (-1)^{i+1}(1 + (-1))^{i+1} - (-1)^{i+1} - 1 \\ &= (-1)^i - 1. \end{aligned} \tag{18}$$

If $j = 1, i > 1$, by plugging (18) to (17), it becomes

$$\begin{aligned} &F(i+1, j) - (i+1)^1 \\ &= (i+1)\left[(-1)^i + \left((-1)^{i'} - 1\right)\right] = -(i+1). \end{aligned} \tag{19}$$

Therefore $F(i+1, 1)$ for all $i > 1$.

For $j > 1$,plug (15) and (19), (17) becomes

$$\begin{aligned} &F(i+1, j) - (i+1)^j \\ &= (i+1)\left[-1 + \sum_{l=1}^{j-1}\binom{j-1}{l}\left(F(i'+1, l) - (i'+1)^l\right)\right]. \end{aligned} \tag{20}$$

Consider the case where $i + 1 > j = k$. From mathematical induction, we know $F(i' + 1, l) = 0$ because $l < j = k$ and $i' + 1 = i \geq j > l$. Therefore,

$$\begin{aligned} &F(i+1, j) - (i+1)^j \\ &= (i+1)\left[-1 - \sum_{l=1}^{j-1}\binom{j-1}{l} i^l\right] \\ &= -(i+1)\left[\sum_{l=0}^{j-1}\binom{j-1}{l} i^l\right] \\ &= -(i+1)(i+1)^{j-1} \\ &= -(i+1)^j. \end{aligned} \tag{21}$$

Therefore $F(i+1, j) = 0$ for $i + 1 > j = k$. □

*Remark.* The $F$ function can be written as $F(i, j) = \Gamma(i+1)S_j^{(i)}$, where $S$ is the Stirling number of the second kind, and $\Gamma$ is the gamma function.

# 3 Effective batch size

**Definition.** Choose a ball from a black box with $m$ labeled indistinguishable balls and put it back after each choose. Repeat $n$ times. The expected value of the number of unique labels is given by $\mathrm{EBS}(m, n)$, where $m, n$ are positive integers.

**Lemma 3.1.**

$$\sum_{k=0}^{\min(m,n)} \binom{m}{k} F(k, n) = m^n, \tag{22}$$

*where $m \in \mathbb{N}, n \in \mathbb{N}$, following the convention that $0^0 = 1$.*

*Proof.* If $n = 0$, we know from (10) that left hand side is $\binom{m}{0} F(0, 0) = 1$. This is consistent with the right hand side.

When $n \neq 0$, for the term $k = 0$, by applying (12),

$$\binom{m}{0} F(0, n) = 0. \tag{23}$$

If $m = 0$ but $n \neq 0$, right hand side is $0^n = 0$. And left hand side is the sole term of $k = 0$, which also gives 0 (see (23)).

For $m > 0, n > 0$, since $F(k, n) = 0$ for $k > n$ (see (14)) and the $k = 0$ term is zero (see (23)), by plugging (13), we get

$$\begin{aligned}
&\sum_{k=0}^{\min(m,n)} \binom{m}{k} F(k, n) \\
&= \sum_{k=1}^{m} \binom{m}{k} F(k, n) \\
&= \sum_{k=1}^{m} \binom{m}{k} \sum_{l=0}^{k-1} (-1)^l \binom{k}{l} (k - l)^n \\
&= \sum_{k=1}^{m} \sum_{l=0}^{k-1} (-1)^l \frac{m!}{(m - k)!(k - l)!l!} (k - l)^n.
\end{aligned} \tag{24}$$

Given $0 \leq l \leq k-1$ and $1 \leq k \leq m$, we denote $k-l$ as $p$. Therefore $k = p+l$,

$1 \leq p \leq m$, and $0 \leq l \leq m-p$. The original formula becomes

$$\begin{aligned}
&\sum_{k=0}^{\min(m,n)} \binom{m}{k} F(k,n) \\
&= \sum_{p=1}^{m} \sum_{l=0}^{m-p} (-1)^l \frac{m!}{(m-p-l)!p!l!} p^n \\
&= \sum_{p=1}^{m} p^n \frac{m!}{p!(m-p)!} \sum_{l=0}^{m-p} (-1)^l \frac{(m-p)!}{(m-p-l)!l!} \\
&= m^n + \sum_{p=1}^{m-1} p^n \frac{m!}{p!(m-p)!} \left(1 + (-1)\right)^{m-p} \\
&= m^n.
\end{aligned} \tag{25}$$

Note that when $p = m$, $m - p$ is zero. The binomial theorem will generate meaningless $0^0$. Therefore the $p = m$ case must be treated specially. □

*Remark.* This is to say, selecting one item out of $m$ freely for $n$ times is the sum of getting $k$ unique items in the process for $k$ running from 0 to the maximum possible value $\min(m, n)$.

**Proposition 3.2.** *The EBS function is calculated by*

$$\mathrm{EBS}(m,n) = m\left[1 - \left(\frac{m-1}{m}\right)^n\right], \tag{26}$$

*where $m, n$ are positive integers.*

*Proof.* Picking exactly $k$ unique items for a chosen set of $k$ items is $F(k, n)$. There are $\binom{m}{k}$ ways to determine such a set of size $k$. Therefore, out of $m^n$ events, the probability of getting exactly $k$ unique items is

$$\binom{m}{k} F(k,n) m^{-n}.$$

Therefore, following the procedure of (24) and (25),

$$
\begin{aligned}
&\mathrm{EBS}(m,n)\\
&= \sum_{k=1}^{\min(m,n)} k\binom{m}{k}F(k,n)m^{-n}\\
&= m^{-n}\sum_{p=1}^{m} p^n \frac{m!}{p!(m-p)!}\sum_{l=0}^{m-p}(-1)^l\frac{(m-p)!}{(m-p-l)!l!}(p+l)\\
&= m^{-n}\sum_{p=1}^{m} p^{n+1} \frac{m!}{p!(m-p)!}\sum_{l=0}^{m-p}(-1)^l\frac{(m-p)!}{(m-p-l)!l!}\\
&+ m^{-n}\left[0+\sum_{p=1}^{m-1} p^n \frac{m!}{p!(m-p)!}\sum_{l=1}^{m-p}(-1)^l\frac{(m-p)!}{(m-p-l)!(l-1)!}\right]\\
&= m^{-n}m^{n+1} + m^{-n}\\
&\left[-m(m-1)^n - \sum_{p=1}^{m-2} p^n \frac{m!}{p!(m-p)!}(m-p)\sum_{l'=0}^{m-p-1}(-1)^{l'}\frac{(m-p-1)!}{(m-p-1-l')!(l')!}\right]\\
&= m - m\frac{(m-1)^n}{m^n} - \sum_{p=1}^{m-2}\frac{p^n}{m^n}\frac{m!}{p!(m-p-1)!}(1+(-1))^{m-p-1}\\
&= m\left[1-\left(\frac{m-1}{m}\right)^n\right],
\end{aligned}
\tag{27}
$$

where $l' = l-1$. Note that $p \le m-2$, so $m-p-1 > 0$. □

*Remark.* $\left[1-\left(\frac{m-1}{m}\right)^n\right]$ is the probability of picking a specific item at least once. Therefore this value multiplied by the number of items is the expected number of unique item being picked.

**Corollary 3.3.** *The effective batch size is the number of draws when there are infinite number of items to choose from:*

$$\lim_{m\to\infty}\mathrm{EBS}(m,n) = n, \tag{28}$$

*where $n$ is a finite positive number.*

*Proof.* Let $x$ be $\frac{1}{m}$, then

$$\lim_{m\to\infty}\mathrm{EBS}(m,n) = \lim_{x\to 0}\frac{1-(1-x)^n}{x} = \lim_{x\to 0}\frac{n(1-x)^{n-1}}{1} = n. \tag{29}$$

□

**Corollary 3.4.** *With infinite draws, the whole set of items will be picked.*

$$\lim_{n \to +\infty} \mathrm{EBS}(m, n) = m, \tag{30}$$

*where $m$ is a finite positive number.*

*Proof.* This is obvious since $\left|\frac{m-1}{m}\right| < 1$ when $m > 0$. □

**Corollary 3.5.** *When $m(m-1) \neq 0$,*

$$\lim_{n \to 0} \mathrm{EBS}(m, n) = 0. \tag{31}$$

**Corollary 3.6.** *When $n(n-1) \neq 0$,*

$$\lim_{m \to 0} \mathrm{EBS}(m, n) \to \infty. \tag{32}$$

# 4 Different levels of permutation

There are different levels of permutation in distributed learning (with $n$ being the number of GPUs) to decrease repetition (and somehow increase "effective batch size"):

1. Naïve sampling (no permutation, no blocks).
2. No permutation of data set, but divide the data set into $n$ blocks such that each GPU only samples within its corresponding block.
3. Permute the data set only once at the beginning, then divide it into $n$ blocks.
4. Permute the data set at the beginning of each epoch, and divide into $n$ blocks.
5. Permute every iteration, and divide into $n$ blocks.

In terms of how the program fetches data, there are different patterns, from more repetition to less:

1. possibly repeating inputs even in a local batch;

2. no repeating input in each local batch;

3. no repeating input in each epoch within GPU, but might be repetitive across GPUs, i.e., for each GPU each epoch, there will be $m/n$ unique inputs ($m$ is the size of data set);

4. no repetition in a global batch;

5. no repetition in a global epoch, i.e., in each epoch, all inputs are guaranteed to sample exactly once.

Note that permutation level 1 + data fetching level 4 is equivalent to permutation level 5 + data fetching level 2.

Permutation level 4 + data fetching level 3 is permutation level 1 + data fetching level 5.

Does the dataloader tend to get repeated samples in distributed learning?

# 5 Effective batch size in distributed learning

The effective global batch size (EGBS) is the averaged or expected unique samples per global batch.

The effective data set size (EDS) is the averaged or expected unique samples per epoch.

We might want low EDS but high EGBS because the weight (parameter) update happens each iteration. The weight update is done by averaging within each *global* batch, i.e., all samples in a *global* batch is contributing to the weight update simultaneously.

However, the data in different global batches affect the weight update at different timings. Therefore, it is unnecessary to correlate samples in different iteration. Instead, we might want a low EDS to prevent overfitting (that is to say, we might want to prevent data fetching level 3 or 5).

If $n$ samples are randomly chosen from a data set of size $m$, such that all selections are *iid* and each input in the data set has equal probability of being sampled. Then, for data fetching level 1 and permutation level 1, effective global batch size =

$$\mathrm{EBS}(m, n \cdot l),$$

where $l$ is the local batch size. For data fetching level $1 + n$-blocking, effective global batch size =

$$\mathrm{EBS}(m/n, l) \cdot n.$$

# 6 Average degeneracy of bosons

Each boson-state fits infinite amount of bosons, unlike fermions. Therefore, the average degeneracy of $n$ Bosons filling $m$ states is $\mathrm{EBS}(m, n)$ at infinite temperature.