20240905

[illegible]

[illegible]

[illegible]

IDENTIFICATION OF THE MINIMIZER

SORT THE PROBABILITY $p_1 \geq p_2 \geq \dots$ in DECREASING ORDER, THEN ASSIGN TO p_i the i-th string in lexicographic order

$$\mathcal{E}(i) = s_i$$ (i_s is [illegible])

THEN THIS MINIMIZES

$\mathbb{E}[\text{length message}]$ over all encodings $\mathcal{E}$.

PROOF

[illegible] $\mathcal{E}$

[illegible] $\text{len}(\mathcal{E}(p_i)) > \text{len}(s_i)$

[illegible] $j > i_{\mathcal{E}(p_i)}$

[illegible] $\text{len}(\mathcal{E}(p_i)) > \text{len}(s_i)$

$\Rightarrow i_{\mathcal{E}(p_i)} > i$

[illegible]

$p_i \,\text{len}(s_i) + p_j \,\text{len}(s_j)$

$< p_i \,\text{len}(s_j) + p_j \,\text{len}(s_i)$

([illegible] $p_i > p_j$)

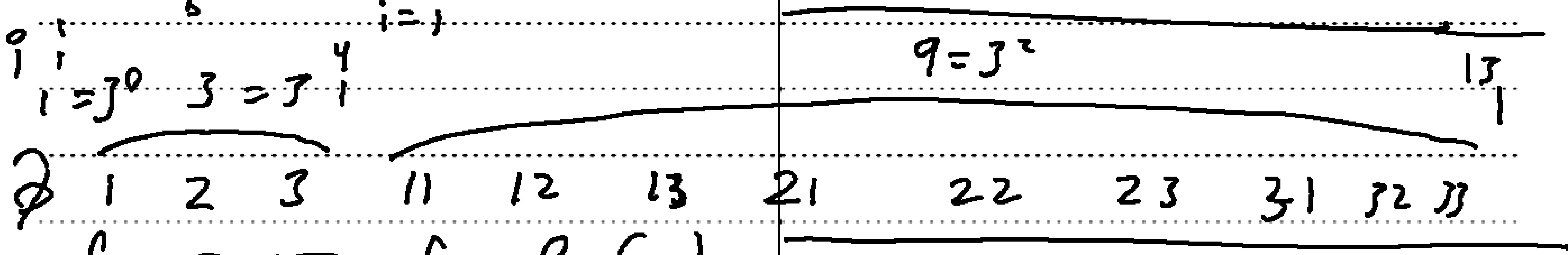
[illegible]

$\Rightarrow$ [illegible] $\square$

[illegible]

[illegible]

$$S_b(p) = \sum_{i=1}^{\infty} p_i \, len(s_i)$$



[illegible] $len(s_i)$

[illegible] $0, 1, 1+b, 1+b+b^2, 1+b+b^2+b^3, \dots$

$$= \frac{b^n-1}{b-1}$$

$$y = \frac{b^n-1}{b-1}$$

$$\Rightarrow (b-1)y + 1 = n\log_b(b$$

[illegible]

$$\Rightarrow \frac{\log((b-1)y+1)}{\log(b)} = n$$

$$\Rightarrow len(s_i) = \left\lfloor \frac{\log((b-1)i+1)}{\log(b)} \right\rfloor$$

$$\geq \left\lfloor \frac{\log((b-1)i)}{\log(b)} \right\rfloor$$

$$\geq \lfloor \log_b(i) \rfloor$$

$\Rightarrow S_b(p)$

$$= \sum_{i=1}^{\infty} p_i \lfloor \log_b((b-1)i+1) \rfloor$$

$$\geq \sum_{i=1}^{\infty} p_i \lfloor \log_b(i) \rfloor$$

[illegible]

[illegible]

$$\lfloor \log_b((b-1)i+1) \rfloor$$
$$\geq \log_b[(b-1)i+1] - 1$$
$$\geq \log_b(i) - 1$$

[illegible]

$$S(p) = \sum p_i \log(i)$$
$$\Rightarrow S_b(p) \geq \frac{S(p)}{\log(b)} - 1$$

[illegible] "-1"

$$\lfloor \log_b((b-1)i+1) \rfloor$$

[illegible]

[illegible]

$$\left[\frac{b^n-1}{b-1}, \frac{b^{n+1}-1}{b-1}\right)$$

$$\sum_{i=\frac{b^n-1}{b-1}}^{\frac{b^{n+1}-1}{b-1}-1} (\log_b(i) - n)$$

20240906

$$= -n\left[\frac{b^{n+1}-1}{b-1} - \frac{b^n-1}{b-1}\right] + \sum \log_b(i)$$
$$= -nb^n + \sum_L^U \log_b(i)$$
$$= -nb^n + \log_b(U!) - \log_b((L-1)!)$$

[illegible]

$$\log_b(n!) = n\log_b(n) - n\log_b(e) + O(\log_b n)$$

[illegible]

$$= -nb^n + U\log_b(U) - U\log_b(e) - (L-1)\log_b(L-1) + (L-1)\log_b(e) + O(\log_b U)$$

$$= -nb^n + \left(\frac{b^{n+1}-1}{b-1}-1\right) \times \log_b\left(\frac{b^{n+1}-1}{b-1}-1\right) - \left(\frac{b^n-1}{b-1}-1\right) \times \log_b\left(\frac{b^n-1}{b-1}-1\right) + O(b^n)$$

[illegible]

[illegible]

[illegible]

[illegible] (1 −

[illegible])

[illegible]

p_i [illegible]

$$S_b(p) \geq \frac{S(p)}{\log(b)}$$

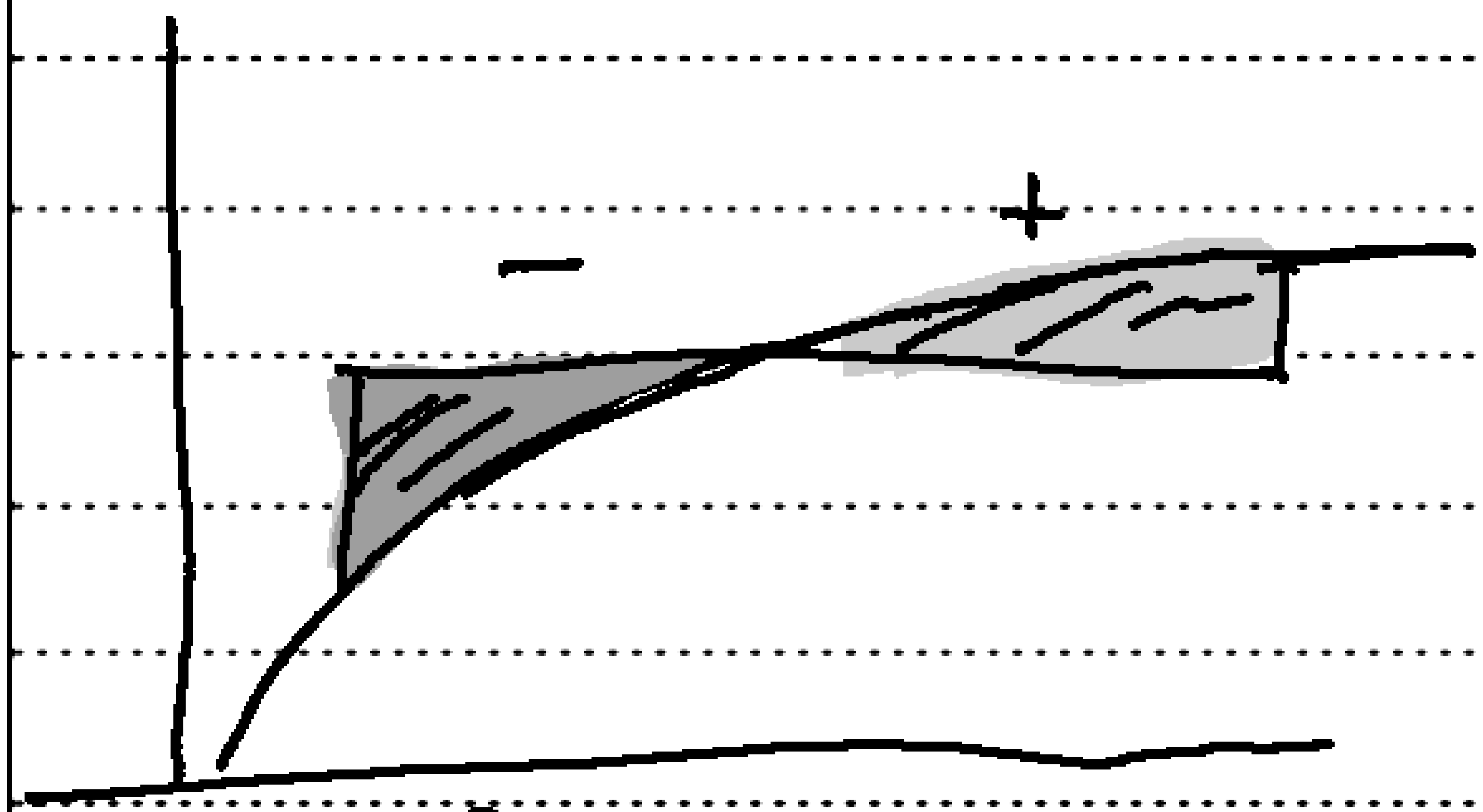
[illegible] "−1"

[illegible]

$$\int_L^{U-1} \log_b(x)\,dx - nb^n$$

$$= (U-1)\log_b(U-1) - L\log_b(L) - nb^n$$

$$= \left(\frac{b^{n+1}-1}{b-1}-1\right)\log_b\left(\frac{b^{n+1}-1}{b-1}-1\right) - \left(\frac{b^n-1}{b-1}\right)\log_b\left(\frac{b^n-1}{b-1}\right) - nb^n$$



[illegible]

[illegible]

[illegible]

$$[\log_3(1)+\log_3(2)+\log_3(3)]-3 = -1.37 < 0$$

$$[\log_3(4)+\dots+\log_3(12)] - 2\cdot 9 = -1.43 < 0$$

[illegible]

$$\int_{e^{n-1}}^{e^n} \log(x)\,dx = ne^n-(n-1)e^{n-1} = [ne-(n-1)]e^{n-1}$$

[illegible]

[illegible]

[illegible]

Claim: Entropy relationship

$S(p) < H(p)$

proof:

[illegible] $p_i \le \frac{1}{i}$

$\Rightarrow -\log(p_i) \ge \log(i)$

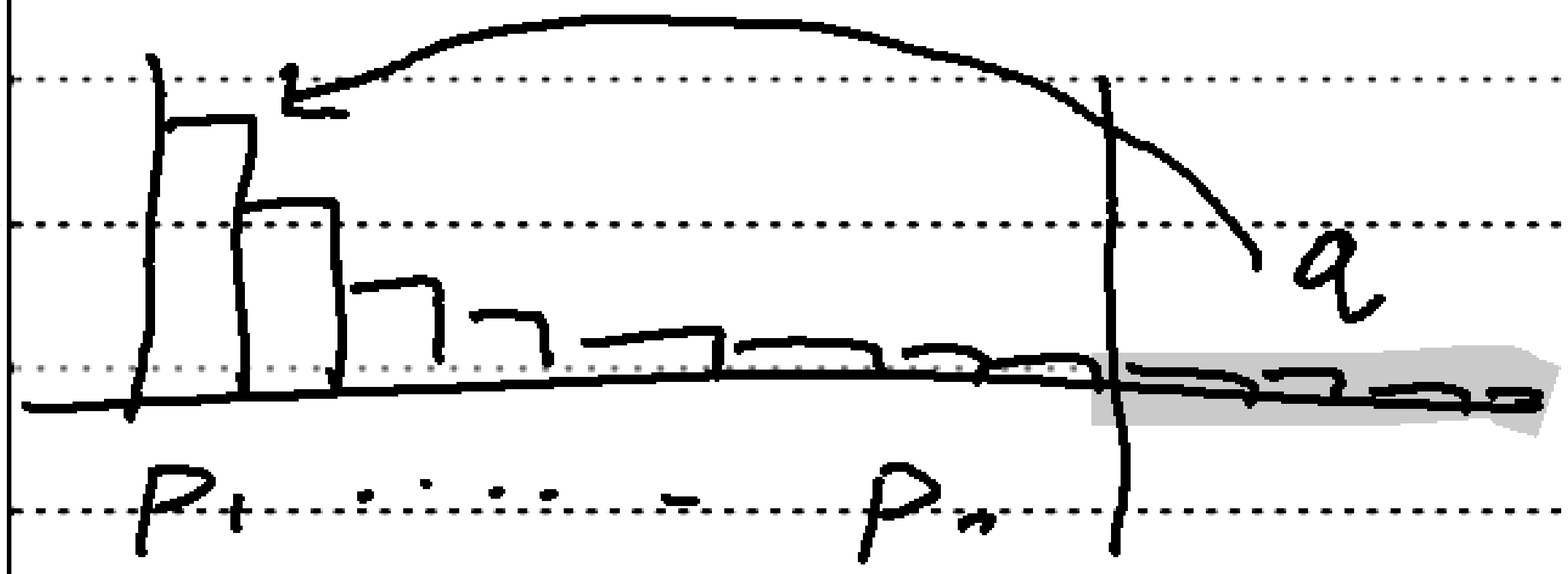
[illegible] $\sum \frac{1}{i} = \infty$

[illegible] $\exists i_0$ s.t. $-\log(p_i) > \log(i)$

$$\Rightarrow S(p) = \sum p_i \log(i) < -\sum_i p_i \log(p_i) = H(p) \quad \blacksquare$$

[illegible]

[illegible]



[illegible] $\log(i)$

[illegible] $\log(i) = \infty$

$S(p)$ $S(p'')$

$$\mathbb{E}[\log(i)] = \mathbb{E}[\log(i)\mathbb{1}_{i \le k}] + \mathbb{E}[\log(i)\mathbb{1}_{i > k}]$$

$$\le S(\hat{p}) + \mathbb{E}[\log(i) \mid i > k]\, q$$

[illegible]

$$\mathbb{P}\{i > k\} = \mathbb{P}\{\log(i) \ge \log(k)\} \le \frac{\mathbb{E}[\log(i)]}{\log(k)} = \frac{S(p)}{\log(k)}$$

$\cancel{\mathbb{E}[\log(i)\mid i>k]\,\mathbb{P}\{i>k\}}$

[illegible]

$S(p) \le S(\hat{p}) + \frac{S(p)}{k}$

$\times \underbrace{\mathbb{E}[\log(i)\mid i>k]}$

[illegible]

$S(p)$

[illegible]

$\log(N)$ [illegible] N [illegible]

[illegible]

$\Rightarrow S(p) \le S(\hat{p}) + q\log(N)$

$\Rightarrow q \ge \frac{S(p)-S(\hat{p})}{\log(N)}$

[illegible] $\mathbb{E}[\log(i)\mid i>k]$

$= \mathbb{E}[\log(i-k)\mid i>k]$

$+\mathbb{E}\left[\log\left(1+\frac{k}{i-k}\right)\middle| i>k\right]$

$\le \underbrace{S(p\mid i>k)}_{\text{conditional entropy}} + \log(1+k)$

$\log(k) \le \frac{S(\hat{p})}{2}$ [illegible]

$\Rightarrow$

$\mathbb{E}[\log(i)\mid i>k]$

$\le S(p\mid i>k) + O\left(\frac{S(\hat{p})}{q}\right)$

[illegible]

Claim (FANO)

$$\mathbb{P}\{\text{error}\} \ge \frac{S(p)-S(\hat{p})}{\log(N)}$$

$$\ge \frac{S(p)-\mathbb{E}(\text{len})}{\log(N)}$$

[illegible]

[illegible]

[illegible]

$$S(p) = \sum_{i=1}^{N} \frac{1}{N} \log(i)$$

$$= \frac{1}{N} \log(N!)$$

$$= \frac{1}{N}\left(N\log(N) - N + O(\log(N))\right)$$

$$= \log(N) - 1 + O\left(\frac{\log(N)}{N}\right)$$

[illegible]

[illegible]

Claim

If $X \sim Unif(N)$

$S(X) = \log(N) - 1 + o(1)$

proof: above [redacted]

[illegible]

[illegible] f, [illegible]

$$S(f(X)) \le S(X) \; ?$$

[illegible]:

[illegible]

$P\{\text{error}\} \ge 0$

Claim

$S(f(x)) \le S(x)$ for any deterministic f

[illegible]

[illegible]

$X, Y \sim Unif(2)$

$(X,Y) \sim Unif(4)$

$S(X) = \log(\sqrt{2})$

$S(Y) = \log(\sqrt{2})$

$S(X,Y) = \log(\sqrt[4]{24})$

$$\underset{\substack{\parallel \\ 0.34}}{\log(\sqrt[4]{24})} > 2\underset{\substack{\parallel \\ 0.30}}{\log(\sqrt{2})}$$

Claim

If X, Y indep, we don't have

$S(X,Y) = S(X) + S(Y)$

[illegible]

Claim "Kraft"

As long as

$\#\{\mathcal{E}(i) \mid \mathrm{len}(\mathcal{E}(i)) = k\} \le b^k$

then the code exists.

[illegible]

FREELY DELIMITED CODES

[illegible]

Claim

If $|\mathrm{supp}(X)| = N$, then $S(X)$ is maximized by $X \sim \mathrm{Unif}(N)$

proof: [illegible]

[illegible] p_N [illegible] $\log(N)$ [illegible] $S(N)$, [illegible] p_i [illegible] $S(p)$.

▨

Claim

$S(X) \ge 0$ with $S(X) = 0$ iff X is constant

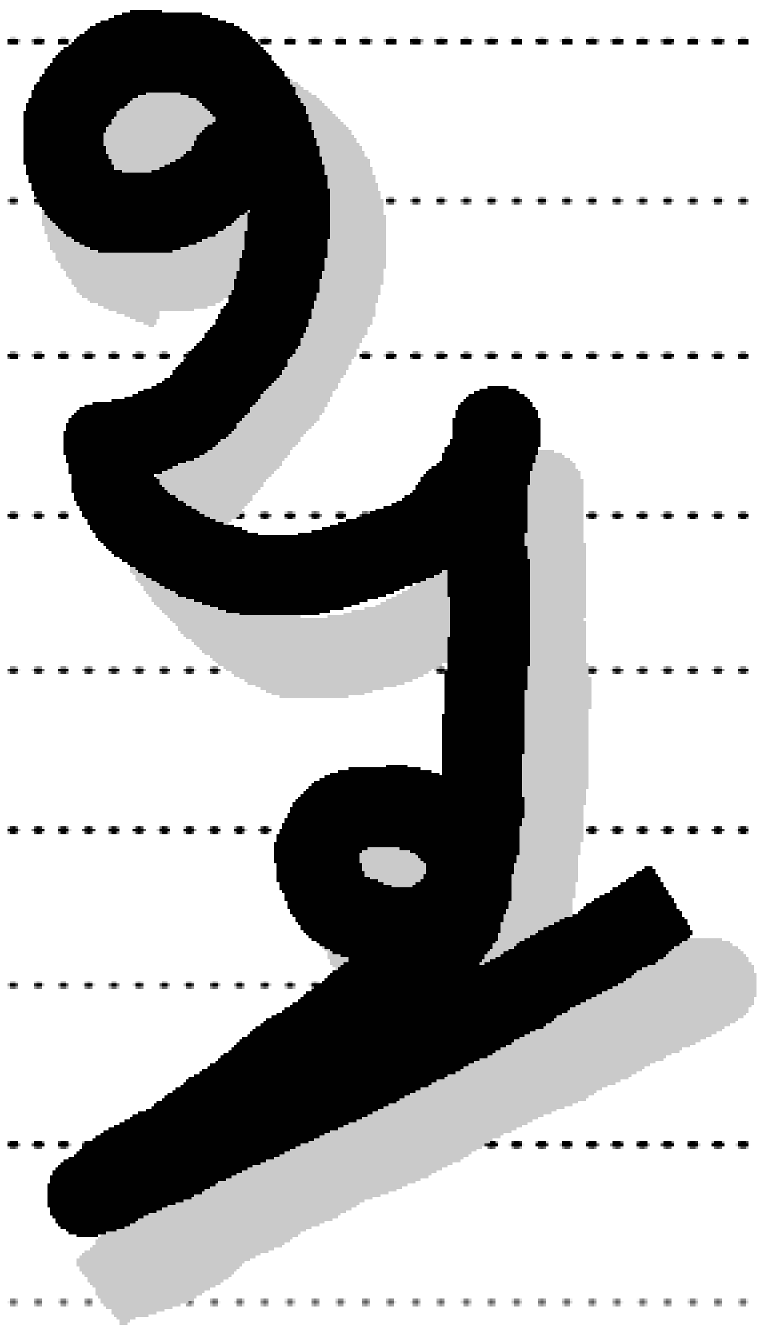
proof

[illegible] ▨

[illegible]

[illegible] !

~~[illegible]~~



[illegible] …

[illegible]

[illegible] ?

[illegible] log(N) [illegible] "N," [illegible]

[illegible]

$$p_i = \begin{cases} 1-\varepsilon & i=1 \\ \frac{\varepsilon}{N-1} & \text{otherwise.} \end{cases}$$

$$S(p) = \frac{\varepsilon}{N-1}\left(\log(2)+\dots+\log(N)\right)$$

$$= \frac{\varepsilon}{N-1}\log(N!)$$

$$= \frac{\varepsilon}{N-1}\left(N\log(N) - N + O(\log(N))\right)$$

$$= \frac{N\varepsilon}{N-1}\log(N) - \frac{\varepsilon N}{N-1} + O\left(\frac{\varepsilon\log(N)}{N-1}\right)$$

$$= \varepsilon\log(N) + O(\varepsilon)$$

[illegible] ε, [illegible] $i=0$:

$$S(\hat{p}) = 0$$

[illegible]

$$S(p) - S(\hat{p}) = \varepsilon\log(N)$$

$$\Rightarrow \frac{S(p)-S(\hat{p})}{\log(N)} = \varepsilon$$

$\Rightarrow$ [illegible]

Claim

The inequality

$$\mathbb{P}\{\text{error}\} \geq \frac{S(p)-S(\hat{p})}{\log(N)}$$

is sharp.

proof: [illegible] ∎

[illegible]

[illegible] $\varepsilon \log(N) = c$ [illegible]

[illegible] $S(p)$ [illegible]

[illegible]

[illegible] $\mathbb{P}\{\text{error}\} = \varepsilon$ [illegible]

g . . .

[illegible]

Claim (FANO FRONT)

Define

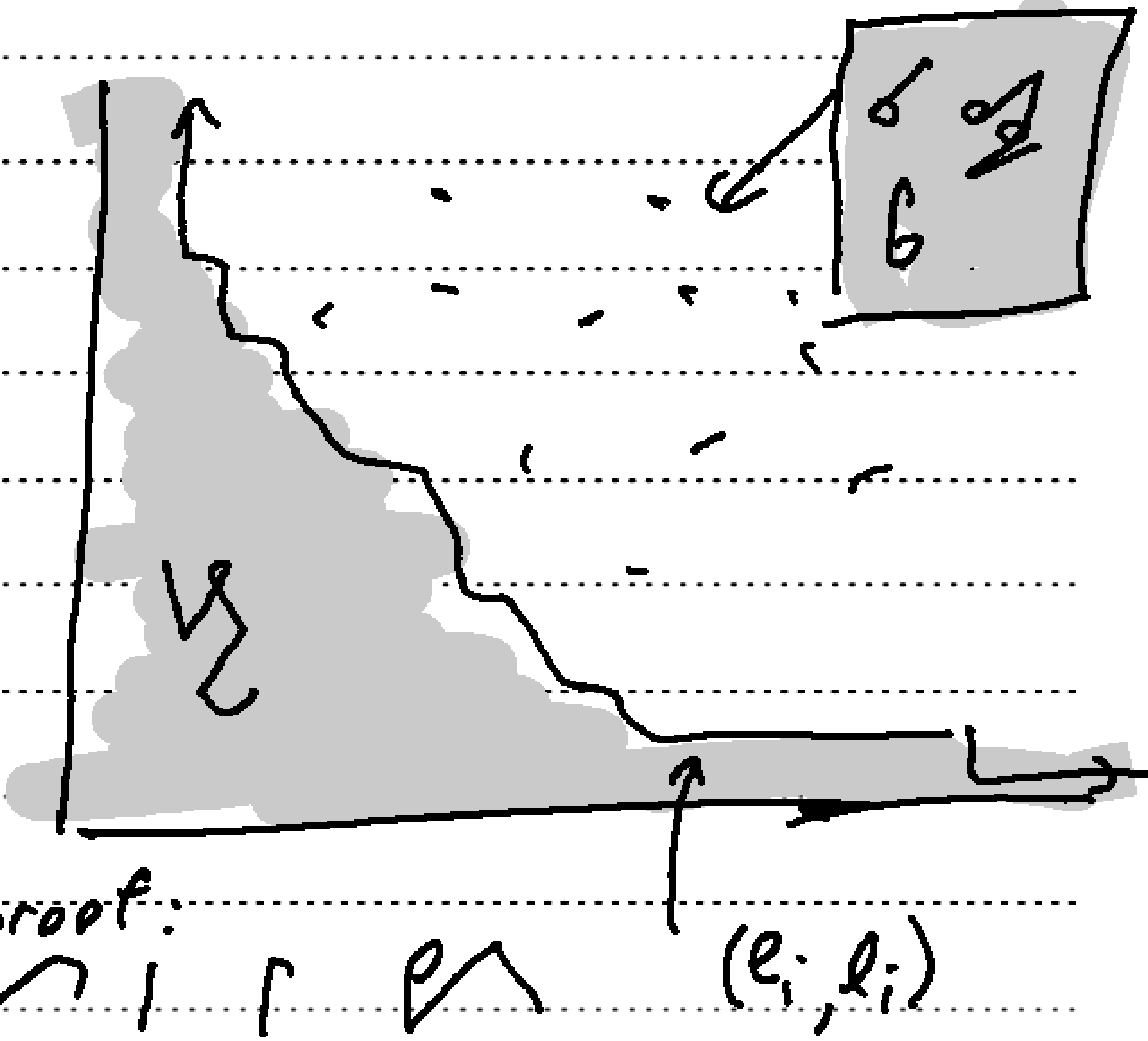
$$e_N = 1 - \sum_{i=1}^{N} p_i$$

$$\ell_n = \sum_{i=1}^{N} p_i \log(i)$$

then

$$\{(e_i, \ell_i) : i = 1, \dots, N\}$$

is the Pareto front for all possible error, expected length pairs



proof: [illegible] ∎

[illegible]

Claim

Given p (assume in order) $S(p+\Delta)$ with $\|\Delta\|_1 = \varepsilon$ is minimized when $(\sum \Delta_i = 0)$

$$\Delta = (\varepsilon, 0, \dots, 0, \underbrace{p_j + \dots + p_N - \varepsilon}_{\substack{0 \leq \\ \leq -p_{j-1}}}, -p_j, \dots, -p_{N-1}, -p_N)$$

proof

LETS FIRST HANDLE THE CASE THAT ALL p_i are distinct so the order is unique

TAKE ANY $\tilde{\Delta} \neq \Delta$,

we will show that it can always be reduced.

First: We show that $(p+\Delta)_i$ may be assumed to be max^l for $i=0$.

$\overset{\text{max}^l}{\downarrow}$

Suppose $(p+\Delta)_j > (p+\Delta)_0$

Since $p_0 > p_j \Rightarrow \Delta_j > \Delta_0$

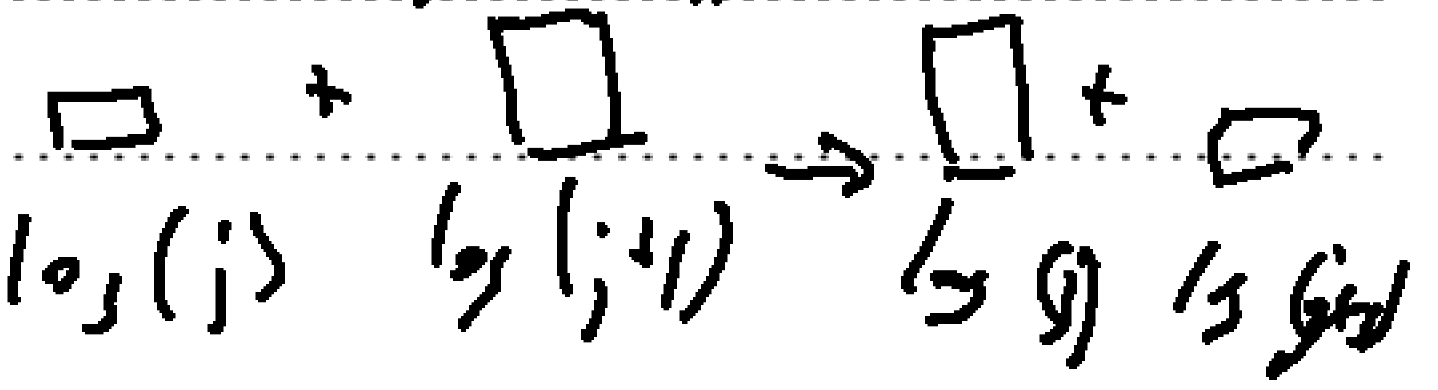
$\Rightarrow p_0 + \Delta_j > p_j + \Delta_0$

so if swapped, then $p_0 + \Delta_j$ is max^l

if $(p_j + \Delta_0) \geq 0$ Then we are done

$(p_j + \Delta_0) < (p_0 + \Delta_0)$

so we remove the mode for $\log(j)$, but it might be out of order, if it is, bubbling it later order makes it better



so say $(p_j + \Delta_0) < 0$

then, further need to

Ok or maybe this:

suppose

$\Delta = (\Delta_1, \ldots, \Delta_n)$
with $\{i_k\}_{k=1}^{K}$ the
indices with $\Delta_{i_k} > 0$

form

$\tilde{\Delta} = (\Delta_1 + \sum_{i_k \neq 1} \Delta_{i_k}$.

Ok, simpler:

SAME SETUP:

① WLOG, we may assume $(p+\Delta)_1$ is a max$^{\underline{l}}$ element.

assume not and $p_j+\Delta_j > p_1+\Delta_1$

Consider

$$\tilde{\Delta}_i = \begin{cases} \Delta_1 + (p_j+\Delta_j) - (p_1+\Delta_1) = \Delta_j + (p_j - p_1) \\ \Delta_j + (p_1+\Delta_1) - (p_j+\Delta_j) = \Delta_1 + (p_1 - p_j) \\ \Delta_i \quad \text{otherwise} \end{cases}$$

Note: this swaps the values of 1 & j since

$$(p+\Delta)_0 = (p+\tilde{\Delta})_j$$

& $(p+\Delta)_j = (p+\tilde{\Delta})_0$

so $S(p+\tilde{\Delta}) = S(p+\Delta)$

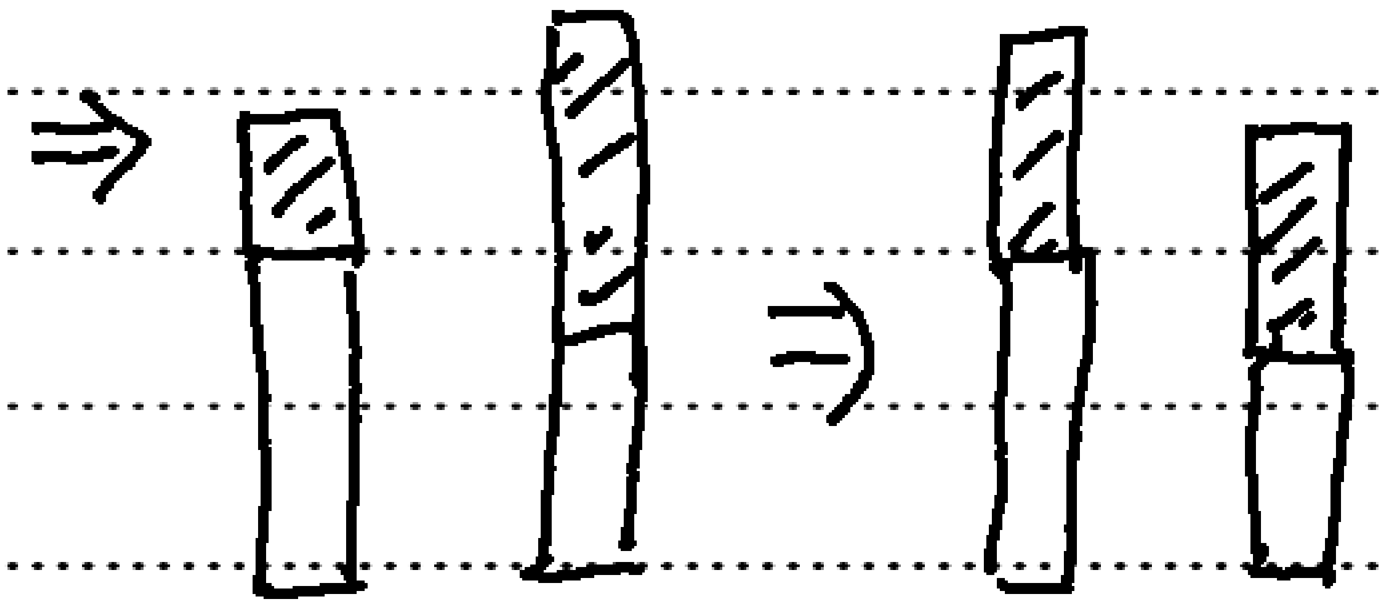
& $\sum_i \tilde{\Delta}_i = 0$

Finally, we need to show that $\|\tilde{\Delta}\|_1 \leq \|\Delta\|_1$

since $p_1 \geq p_j$ and

(*) $p_j+\Delta_j > p_1+\Delta_1$

20240908



if $\Delta_1, \Delta_j \geq 0$ then

$$|\tilde{\Delta}_1| + |\tilde{\Delta}_j| = |\Delta_j + (p_j - p_1)| + |\underset{\overset{\vee}{0}}{\Delta_1} + (p_1 \underset{\overset{\vee}{0}}{-} p_j)|$$

$> \Delta_1 \geq 0$

$$= \Delta_j + \Delta_1 + (p_j - p_1) + (p_1 - p_j)$$

$$= \Delta_j + \Delta_1 = |\Delta_1| + |\Delta_j|$$

if $\Delta_1 \geq 0$, $\Delta_j < 0$

this would mean

$$p_1 + \Delta_1 \geq p_1 \geq p_j > p_j + \Delta_j \quad \times.$$

if $\Delta_1 < 0$, $\Delta_j \geq 0$

$|\tilde{\Delta}_1| + |\tilde{\Delta}_j|$

$= |\Delta_j + (p_j - p_1)| + |\Delta_1 + (p_1 - p_j)|$

Example (of this case)

$p = (.9, .1)$

$\Delta = (-0.5, 0.5)$

$\|\Delta\|_1 = 1$

$p + \Delta = (.4, .6)$

$\tilde{\Delta} = (-0.3, 0.3)$

$\|\tilde{\Delta}\|_1 = 0.6 < \|\Delta\|_1$

$\Delta_j + (p_j - p_1) > \Delta_1$

$\Delta_1 + (p_1 - p_j) < \Delta_j$

claim $\Delta_j + (p_j - p_1) \leq 0$

& $\Delta_1 + (p_1 - p_j) \geq 0$

No

Ex $p = (.9, .1)$

$\Delta = (-0.9, 0.9)$

$p + \Delta = (0, 1)$

$\tilde{\Delta} = (.1, -.1)$

so the claim is false

$\Delta_j + (p_j - p_1) < -\Delta_1$

$\Delta_1 + (p_1 - p_j) > -\Delta_j$

?? Maybe ??

No

Example

$(\frac{1}{N}, \ldots, \frac{1}{N}) \to (0, \ldots, 1)$

$\Delta = (-\frac{1}{N}, \ldots, 1)$

$\tilde{\Delta} = (1, \ldots, -\frac{1}{N})$

so claim is false

but

$\Delta_j > \Delta_j + (p_j - p_1) > \Delta_1$

$\Delta_j > \Delta_1 + (p_1 - p_j) > \Delta_1$

is true?

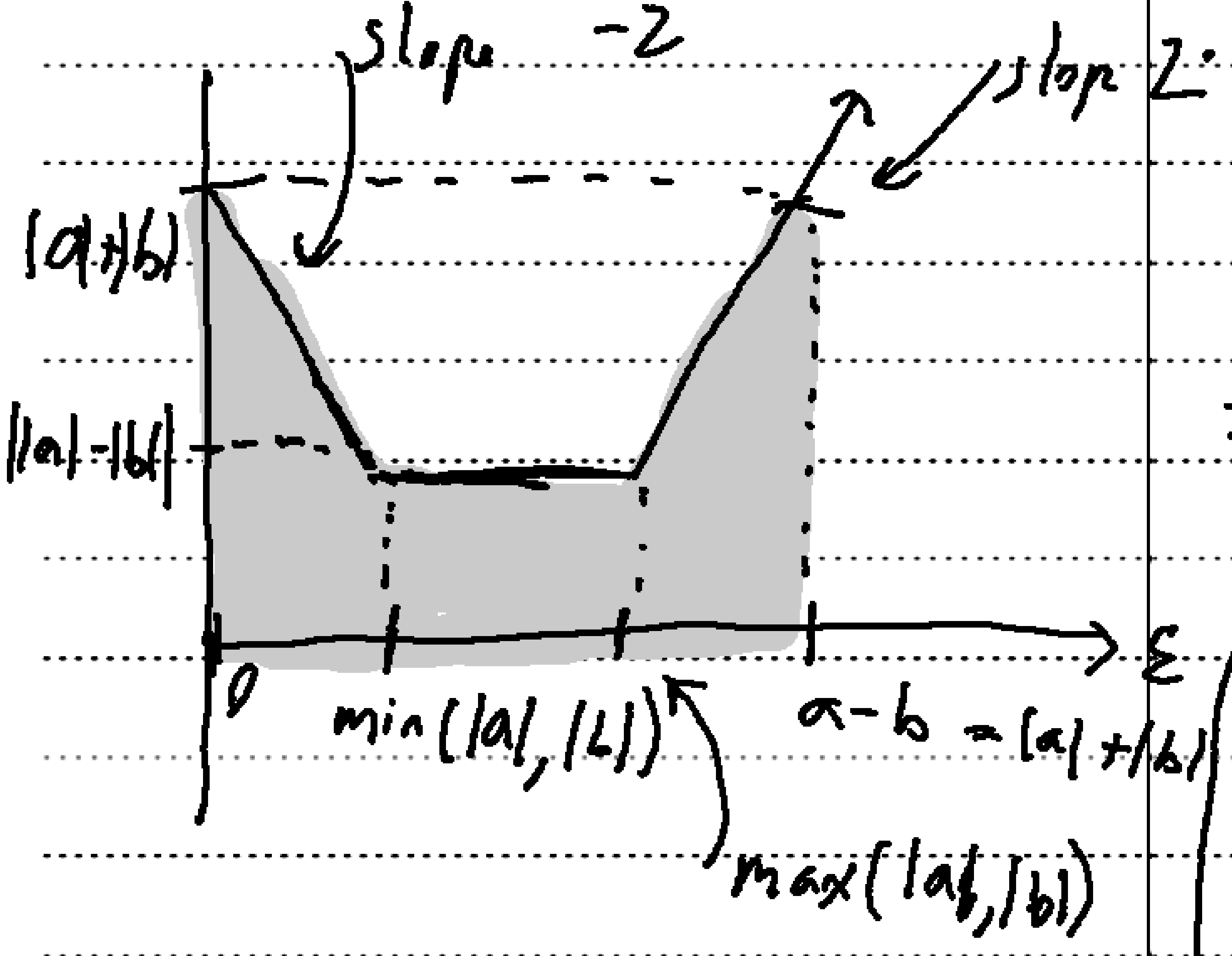
yes since $p_j - p_1 < 0$ & $p_1 - p_j > 0$

yes

Lemma $\underset{a}{\overset{\Delta_j}{\downarrow}} > 0 > \underset{b}{\overset{\Delta_1}{\downarrow}}$

$|a - \varepsilon| + |b + \varepsilon| \leq |a| + |b|$

$\forall \varepsilon \in [0, a - b]$



∎

since

$\Delta_j > \Delta_j + (p_j - p_1) > \Delta_1$

$\Delta_j > \Delta_1 + (p_1 - p_j) > \Delta_1$

⇒ we are in the regime where the above inequality holds

$\Rightarrow \|\tilde{\Delta}\|_1 \leq \|\Delta\|_1$

LAST CASE

$\Delta_1 < 0, \quad \Delta_j < 0$

$$|\underset{\overset{\wedge}{0}}{\Delta_j} + \underset{\overset{\wedge}{0}}{(p_j - p_1)}| + |\underset{\overset{\wedge}{0}}{\Delta_1 + (p_1 - p_j)}|$$

$p_j + \Delta_j > p_1 + \Delta_1$

$\Rightarrow \Delta_1 + p_1 - p_j < \Delta_j < 0$

⇒ all negative

$$= -\Delta_j - \cancel{(p_j - p_1)} - \Delta_1 - \cancel{(p_1 - p_j)}$$

$$= |\Delta_1| + |\Delta_j|$$

∎

⇒ ① is done! so WLOG we may assume $i=1$ is the max$^{\underline{l}}$ elt of $p+\Delta$.

② Any minimizer has $\Delta_j \leq 0 \ \forall_{i>1}$

proof Assume not, and suppose $\Delta_j > 0$

$$\tilde{\Delta} = (\Delta_1 + \Delta_j, \Delta_2, \ldots, \Delta_{j-1}, 0, \Delta_{j+1}, \ldots, \Delta_N)$$

has $S(\tilde{\Delta}) < S(\Delta)$

since $\log(1) = 0$
and $\log(j) > 0$

so $\Delta_j \log(j) > \Delta_j \log(1) = 0$

■

③ The optimizer removes all mass from the tail of p.

find the last step i s.t.

$$\underbrace{\Delta_i < 0}_{\text{part deleted}}, \quad \underbrace{\Delta_{i+1} > -p_{i+1}}_{\text{not fully deleted}}$$

$$\underbrace{\Delta_j = -p_j \quad \forall_{j > i+1}}_{\text{all more tail are deleted.}}$$

like a simpler version of ①

CASE: $p_{i+1} + \Delta_{i+1} \leq p_i + \Delta_i$

then we may assume that the log term of the sorting is strictly larger for $p_{i+1} + \Delta_{i+1}$ than for $p_i + \Delta_i$,

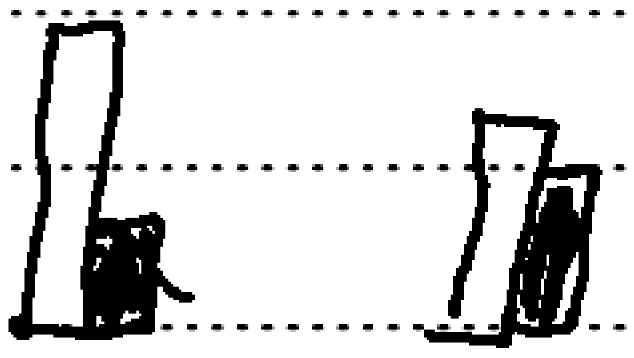
then

$$\tilde{\Delta} = (\Delta_1, \dots, \Delta_i + \delta, \Delta_{i+1} - \delta, \dots, \Delta_N)$$

has $S(\tilde{\Delta}) < S(\Delta)$ ⚡

CASE: $p_{i+1} + \Delta_{i+1} > p_i + \Delta_i$

since $p_{i+1} \leq p_i$,
$\Delta_{i+1}, \Delta_i \leq 0$



$\Rightarrow |\Delta_i| > |\Delta_{i+1}|$ $(\Delta_i < \Delta_{i+1})$
(you need to subtract off more to be valid)

We can just mirror the last proof

$$p_i \geq p_{i+1}$$

$$p_i + \Delta_i < p_{i+1} + \Delta_{i+1}$$

define to swap the order so

$$\tilde{\Delta}_j = \begin{cases} \Delta_{i+1} + (p_{i+1} - p_i) & j = i+1 \\ \Delta_i + (p_i - p_{i+1}) & j = i \\ \Delta_j & \text{otherwise} \end{cases}$$

This has the same

$S(\tilde{\Delta}) = S(\Delta)$, but maintaining the order. This reduces us to the previous case

checking needed conditions

$(\tilde{\Delta}_j \leq 0$,

$$p_{i+1} + \tilde{\Delta}_{i+1} < p_i + \tilde{\Delta}_i$$

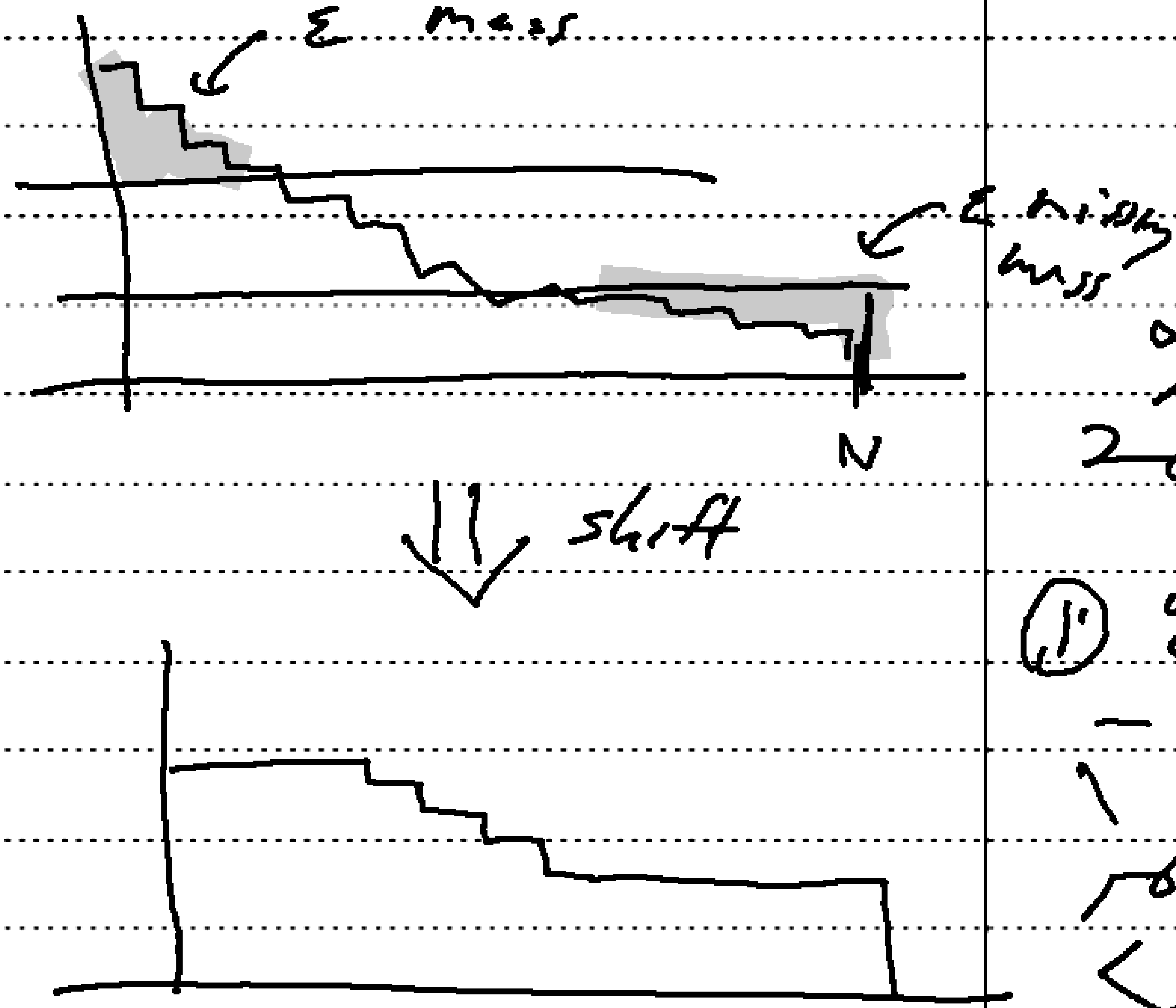
omitted since it is not interesting [redacted]

OK!

So this proves that we indeed decreased the $\underline{\text{max}}\ \ell$ rate by subtracting from the tail ℓ dumping on the head!

[illegible]:

① [illegible]



[illegible]

②

$$\log_3(2) + \log_3(3) + \log_3(4) - 3 < 0$$

[illegible]

$$\sum_{i=5}^{13} \log_3(i) - 9 \times 2 < 0$$

[illegible]

$$\tilde{S}(p) = \sum_{i=1}^{\infty} p_i \log(i+1)$$

[illegible]

①' [illegible] $S(p + \Delta^{(i)}) \to \infty$ [illegible] $i \to \infty$.

②* [illegible]?

[illegible]

[illegible]

Claim

For any distribution p, let

$$S_b(p) = \sum_{i=1}^{\infty} p_i \lfloor \log_b((b-1)i+1) \rfloor$$

Then there is an optimal b s.t.

$\log(b) S_b(p)$ is minimized. For some p, $b > 2$.

proof

[illegible] $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

$S_3(p) = 1$

$S_2(p) = \frac{2}{3} \cdot 1 \cdot \frac{1}{3} 2 = \frac{4}{3}$

$\log(3) \cdot 1 \approx 1.09$

$\vee$

$\log(2) \cdot \frac{4}{3} \approx 0.92$

$\log(6) \approx 1.79$

$\log(2) \cdot (\frac{2}{6} \cdot 1 + \frac{4}{6} \cdot 2) \approx 1.73$

[illegible] 2 [illegible] 2 [illegible]

[illegible] 2 [illegible]

No.

Claim

$S_b(p)$ [illegible] $b = 2$

proof

$\lfloor \log_b((b-1)k+1) \rfloor \cdot \log(b)$

$\wedge|$ $\vee|$

$\log((b-1)k+1)$ $\log(\frac{(b-1)k}{b})$

[illegible]

[illegible]

[illegible] 1 [illegible] . zipf

[illegible]:

$$\int_a^b \frac{1}{x}dx = \log(b)-\log(a)$$
$$= \log\left(\frac{b}{a}\right)$$

$$\int_a^b \frac{1}{x}\log(x)\,dx$$
$$= \frac{\log(b)^2}{2} - \frac{\log(a)^2}{2}$$

[illegible] $\mathrm{Zipf}(n) = \begin{cases} p_i = \frac{1}{i} & i \le n \\ 0 & \text{other} \end{cases}$

$$S(p) \approx \frac{1}{2}\frac{\log(n)^2 - \log(1)^2}{\log(n)-\log(1)}$$
$$= \frac{1}{2}\log(n)$$

[illegible]

[illegible]

L −

$$\left(\frac{\log(x)^2}{2\log(n)},\ \frac{\log\left(\frac{n}{x}\right)}{\log(n)}\right)$$

$$\frac{1}{\log(n)}\left(\frac{1}{2}\underbrace{\log(x)^2}_{t^2},\ \log(n) - \underbrace{\log(x)}_{t}\right)$$

$$\left(\frac{1}{2}\log(n)\,s^2,\ 1-s\right)$$

$$s = \frac{\log(x)}{\log(n)}$$

$$\underline{\underline{\left(S(p)\,s^2,\ 1-s\right)}}$$

[illegible] V [illegible]

$$\int_a^b dx = b - a$$

$$\int_a^b \log(x)\,dx = \begin{matrix} b\log(b)-b-a\times \\ \log(a)-a \end{matrix}$$

$$S(p) \approx \frac{n\log(n) - n - 1\cdot\log(1)+1}{n-1}$$
$$= \frac{n}{n-1}\log(n) - 1$$

V [illegible]:

$$\left(x\log(x)\qquad 1-\frac{x}{n}\right)$$

[illegible]

zipf

$$H(p) = -\sum_{i=1}^{N} \frac{Z}{i}\log\left(\frac{Z}{i}\right)$$

$$Z := \left(\sum_{i=1}^{N}\frac{1}{i}\right)^{-1}$$

$$= -Z\sum_{i=1}^{N}\frac{1}{i}\left[\log(Z) - \log(i)\right]$$

$$= S(p) - \log(Z)$$

$$Z \approx \left(\log(N)\right)^{-1}$$

$$\Rightarrow H(p) \approx S(p) + \log(\log(N))$$

[illegible] $S(p)$ [illegible] $H(p)$ [illegible]

[illegible] zipf [illegible]

2024/020

Zipf [illegible]

[illegible]

Claim

$$0 \le H(p) - S(p) \le \log(H_N)$$

when p is supported on N values, and H_N is the N^{th} Harmonic #.

proof

[illegible]

[illegible] p_i [illegible] $p_1 \ge p_2 \ge \dots$

$$H(p) - S(p) = -\sum_i p_i \left(\log(p_i) + \log(i)\right)$$

$$= -\sum_i p_i \log(i p_i)$$

[illegible]

$$F = -\sum_i p_i \log(i p_i) + \lambda\left(\sum_i p_i - 1\right)$$

$$\frac{dF}{dp_i} = -\frac{p_i i}{i p_i} + -\log(i p_i) + \lambda$$

$$= -1 + -\log(i p_i) + \lambda = 0$$

$$\frac{dF}{d\lambda} = \sum_i p_i - 1 = 0$$

[illegible]

$$-1 + \lambda + -\log(i p_i) = 0$$

$$\Rightarrow p_i = \frac{1}{i} e^{\lambda - 1}$$

[illegible] λ [illegible]

$$e^{\lambda - 1} = \frac{1}{H_N} = \frac{1}{\sum_{i=1}^{N} \frac{1}{i}}$$

$$\Rightarrow p_i = \frac{1}{i H_N} \Rightarrow \text{Zipf.}$$

[illegible]

[illegible] $p_j = 0$ [illegible]

[illegible]

$$\left(-\frac{K}{N}, -\frac{K}{N}, \dots, 1-\frac{K}{N}, \dots, -\frac{K}{N}\right)$$

[illegible] $p_j = 0 \Rightarrow 1-\frac{1}{N}$ [illegible] K [illegible]

… [illegible]

$$\exp(H-S) = \prod\left(\frac{1}{ip_i}\right)^{p_i}$$

$$\leq \sum p_i \frac{1}{ip_i} = H_N$$

[illegible] WEIGHTED AM-GM INEQUALITY

[illegible]

$\frac{1}{ip_i}$ [illegible]

$$\Rightarrow p_i = \frac{1}{iH_N} \qquad \square$$

[illegible]

$$\Rightarrow |H(p) - S(p)| = O(\log(\log(N)))$$