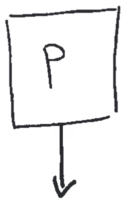
Lecture 13:

Testing Distributions

- Uniformity (cont.)
- Monotonicity

Turning to a new model: prob dists

Probability distributions - get samples of distribution



outputs iid samples

Domain D, $|D| = n$ ← known

$p_i = \Pr[p \text{ outputs } i]$ ← unknown

↖ this is all we can learn from

Examples:

Lottery data

Shopping choices

experimental outcomes

⋮

What do we want to know?

is it uniform? e.g. lottery

is it high entropy?

large support? (many distinct elements have >0 probability)

is p monotone increasing, k-modal, monotone hazard rate...?

how can we do it?

χ^2 test

plug in estimate

learn distribution, Maximum likelihood estimates

Goal: sample complexity SUBLINEAR in n

Testing Uniformity

The goal:

Uniform dist on D

- if $P = U_D$ then tester outputs PASS ← with prob $\geq 3/4$
- if $\underbrace{\text{dist}(P, U_D)}_{}$ $> \varepsilon$ then tester outputs FAIL

which measure of distance?

ℓ_1, ℓ_2, KL-divergence, Earth mover, Jensen-Shannon

today's focus

Distances

ℓ_1-distance : $\|p-q\|_1 = \sum_{i\in D} |p_i - q_i|$

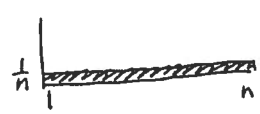
ℓ_2-distance : $\|p-q\|_2 = \sqrt{\sum_{i\in D} (p_i - q_i)^2}$

$$\|p-q\|_2 \le \|p-q\|_1 \le n^{1/2} \|p-q\|_2$$

examples:

1. $p = (1, 0, 0, \ldots 0)$

$q = (\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n})$



ℓ_1 distance:
$\|p-q\|_1 = \left(\frac{n-1}{n}\right) + (n-1)\cdot\frac{1}{n}$
≈ 2

ℓ_2-distance:
$\|p-q\|_2^2 = \left(1-\frac{1}{n}\right)^2 + (n-1)\left(\frac{1}{n}\right)^2$
≈ 1

2. $p = (\frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n}, 0, 0, \ldots 0)$

$q = (0, 0, \ldots 0, \frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n})$

ℓ_1 distance:
$\|p-q\|_1 = n \cdot \left(\frac{2}{n}\right) = 2$

ℓ_2-distance: $\|p-q\|_2^2 = n \cdot \left(\frac{2}{n}\right)^2 = \frac{4}{n}$

$\|p-q\|_2 = \frac{2}{\sqrt{n}}$

"Plug-in" Estimate:

Algorithm:

- take m samples from p
- estimate $p(x)$ $\forall x$ via

$$\hat{p}(x) = \frac{\#\text{ times } x \text{ occurs in sample}}{m}$$

- if $\sum_x \left|\hat{p}(x) - \frac{1}{n}\right| > \varepsilon$ reject

  else accept.

Analysis: (better analyses exist)

So, if $p = U_n$ then p passes

pick m s.t. $\forall x, \left|\hat{p}(x) - p(x)\right| < \frac{\varepsilon}{n} \Rightarrow \|\hat{p} - p\|_1 < \varepsilon$

by $\Delta\neq$, if $\|p - \hat{p}\|_1 < \varepsilon$ + $\|\hat{p} - U\|_1 < \varepsilon$

then $\|p - U\|_1 < 2\varepsilon$.

So, if $\|p - U_n\| \overset{?}{\geq} 2\varepsilon$ this test is likely to Fail

how many samples? $\Omega\left(\frac{n}{\varepsilon}\right)$ maybe even worse...

for each x, need to see it at least once in order to give non zero estimate.

$\Theta(n)$? Can we do better?

Better analysis:

Claim $E[\|\hat{p}-p\|_1] \le \sqrt{\frac{n}{m}}$

Pf

$$E[\|\hat{p}-p\|_1] = \sum_x E[|\hat{p}(x)-p(x)|]$$

note: $E[\hat{p}(x)] = \frac{1}{m}E[\sum 1_{i\text{th sample is } x}]$

$= \frac{1}{m}\sum_{i=1}^{m} E[1_{i\text{th sample is } x}]$

$= \frac{\cancel{m}\cdot p(x)}{\cancel{m}} = p(x)$

$$\le \sum_x \sqrt{E[(\hat{p}(x)-p(x))^2]}$$

Jensen's ≠

$$= \sum_x \sqrt{Var(\hat{p}(x))}$$

$Var(\hat{p}(x)) = \frac{1}{m^2} m\, p(x)(1-p(x))$

$\le \frac{p(x)}{m}$

$$\le \sum_x \sqrt{\frac{p(x)}{m}}$$

$$\le \frac{1}{\sqrt{m}}\cdot\sqrt{n}$$

since $\max_{p \in \text{prob dist over domain of size } n} \sum \sqrt{p(x)}$ is $\sqrt{n}$

So picking $m = \Omega\left(\frac{n}{\varepsilon^2}\right)$ gives

$$E[\|\hat{p}-p\|_1] \le \frac{\varepsilon}{2}$$

by Markov's ≠: with prob $1-\frac{1}{2}$, $\|\hat{p}-p\|_1 \le \varepsilon$

Note, this says can "learn" (approximate) any dist w.r.t. L_1 distance in $\Theta(n/\varepsilon^2)$ samples

L_2 - Distance (squared):

$$\|p - U_{[n]}\|_2^2 = \sum_{i \in [n]} \left(p_i - \tfrac{1}{n}\right)^2$$

$$= \sum p_i^2 - \tfrac{2}{n} \underbrace{\sum p_i}_{=1} + \underbrace{\sum \left(\tfrac{1}{n}\right)^2}_{=\frac{1}{n}}$$

$$= \underbrace{\sum p_i^2}_{} - \tfrac{1}{n}$$

Collision probability of p:

$$\|p\|_2^2 \equiv \Pr_{s,t \in p}[\, s = t \,] = \sum p_i^2$$

for $p = U$, $\|p\|_2^2 = \frac{1}{n}$

for $p \neq U$, $\|p\|_2^2 > \frac{1}{n}$

$$= \underbrace{\|p\|_2^2}_{\text{we can estimate this}} - \underbrace{\|U_{(n)}\|_2^2}_{\text{we know this since we know } n}$$

Algorithm

1. take s samples from p ← ① how many samples?
2. let $\hat{c} \leftarrow$ estimate of $\|p\|_2^2$ from sample ← ② how?
3. if $\hat{c} < \frac{1}{n} + \delta$ pass ← ③ what should δ be?

   else fail

First:

How to estimate $\|p\|_2^2$?

Naive idea:

take two new samples:

$$X_i \leftarrow \begin{cases} 1 & \text{if samples are equal} \\ 0 & \text{o.w} \end{cases}$$

"gives $\Theta(k)$ samples of collision probability from k samples of p"

Better idea: recycle - use all pairs in sample

"gives $\Theta(k^2)$ samples of collision probability from k samples of p"

Estimate by recycling:

- Take s samples from p: $X_1 \cdots X_s$
- for each $1 \le i < j \le s$

$$\sigma_{ij} \leftarrow \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{if } X_i \neq X_j \end{cases}$$

- Output $\hat{c} \leftarrow \dfrac{\sum_{i<j} \sigma_{ij}}{\binom{s}{2}}$

σ_{ij}'s not independent so can't use Chernoff

Analysis: $E[\hat{c}] = \frac{1}{\binom{s}{2}} \cdot \binom{s}{2} \cdot E[\sigma_{ij}]$

$= \|p\|_2^2$

How well do we need to estimate $\|p\|_2^2$?

Assumption ☆: $|\hat{c} - \|p\|_2^2| < \Delta$

will take enough samples so that this holds with prob $\geq 3/4$

this is our parameter that determines whether our approximation is good. Spoiler: will set $\Delta = \frac{\varepsilon^2}{2}$

What happens if ☆ holds with $\Delta = \frac{\varepsilon^2}{2}$?

Correct behavior!

- if $p = U_{[n]}$ then $\hat{c} \leq \|U_{[n]}\|_2^2 + \Delta = \frac{1}{n} + \frac{\varepsilon^2}{2}$

  so test will PASS

- if $\|p - U_{[n]}\|_2 > \varepsilon$ then $\|p - U_{[n]}\|_2^2 > \varepsilon^2$

  but $\|p\|_2^2 = \|p - U_{[n]}\|_2^2 + \frac{1}{n}$ ← see p.6

  $> \varepsilon^2 + \frac{1}{n}$

  & $\hat{c} > \|p\|_2^2 - \Delta$ ← ☆

  $\geq \varepsilon^2 + \frac{1}{n} - \Delta = \varepsilon^2 + \frac{1}{n} - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2} + \frac{1}{n}$

  so test will FAIL

Remaining Question:

How many samples do we need to estimate $\hat{c}$ to within Δ?

Analysis

$$E[6_{ij}] = Pr[6_{ij} = 1]$$
$$= \|p\|_2^2$$

recall: $Var[X] = E[(X - E[X])^2]$

$$E[\hat{c}] = \frac{1}{\binom{s}{2}} \binom{s}{2} E[6_{ij}] = \|p\|_2^2$$

$$Pr[\,|\hat{c} - \|p\|_2^2| > \rho\,] \le \frac{Var[\hat{c}]}{\rho^2}$$

Chebyshev ‡

Fact $Var[aX] = a^2 Var[X]$

So $$Var[\hat{c}] = Var\left[\frac{1}{\binom{s}{2}} \cdot \sum_{i<j} 6_{ij}\right]$$
$$= \frac{1}{\binom{s}{2}^2} Var\left[\sum_{i<j} 6_{ij}\right]$$

Fact $\Rightarrow$ $Var[\hat{c}] \le \frac{4 \cdot \left(\binom{s}{2} \|p\|_2^2\right)^{3/2}}{\binom{s}{2}^2} \le \theta(\|p\|_2^3 / s)$

Lemma $Var[\sum 6_{ij}] \le 4\left(\binom{s}{2} \|p\|_2^2\right)^{3/2}$

Why? (proof...)

def. $\bar{6}_{ij} = 6_{ij} - E[6_{ij}]$

So $E[\bar{6}_{ij}] = 0$

trick - will rewrite variance as $E[(\sum\bar{6}_{ij})^2]$.

why?

$$Var[\sum \bar{6}_{ij}] = E[(\sum\bar{6}_{ij} - \underbrace{E[\sum\bar{6}_{ij}]}_{=0})^2]$$
$$= E[(\sum 6_{ij} - E[\sum 6_{ij}])^2]$$
$$= Var[\sum 6_{ij}]$$

Also:
- $E[\bar{6}_{ij}\,\bar{6}_{kl}] \le E[6_{ij}\,6_{kl}]$

verify at home? (or trust...)
- $\left(\sum p(x)^3\right)^{1/3} \le \left(\sum p(x)^2\right)^{1/2}$
- $s^2 \le 3\binom{s}{2}$
- $\binom{s}{3} \le s^3/6$

e.g. $(a^3+b^3)^2 \le (a^2+b^2)^3$

$a^6 + 2a^3b^3 + b^6 \le a^6 + b^6 + 3a^4b^2 + 3a^2b^4$

So

$$Var\Big[\sum_{i<j}\sigma_{ij}\Big] = E\Big[\Big(\sum_{i<j}\sigma_{ij} - E\Big[\sum_{i<j}\sigma_{ij}\Big]\Big)^2\Big]$$

$$= E\Big[\Big(\sum_{i<j}\bar{\sigma}_{ij}\Big)^2\Big]$$

$$= E\Big[\underbrace{\sum_{i<j}\bar{\sigma}_{ij}^{\,2}}_{1} + \underbrace{\sum_{\substack{i<j\\k<l\\i,j,k,l\text{ distinct}}}\bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{2} + \underbrace{\sum_{\substack{i<j\\ \cancel{k<l}\\ i,j,l\\ \text{distinct}}}\bar{\sigma}_{ij}\bar{\sigma}_{il}}_{3} + \underbrace{\sum_{\substack{i<j\\k<j\\i,k,j\\ \text{distinct}}}\bar{\sigma}_{ij}\bar{\sigma}_{kj}}_{4}\Big]$$

(5) $+ \sum \bar{\sigma}_{ij}\bar{\sigma}_{jl}$

(6) $+ \sum \bar{\sigma}_{ij}\bar{\sigma}_{ki}$

(1) $E\Big[\sum_{i<j}\bar{\sigma}_{ij}^{\,2}\Big] \le E\Big[\sum \sigma_{ij}^2\Big] = \binom{s}{2}\|p\|_2^2$

$E[\sigma_{ij}] = E[\sigma_{ij}^2]$ since σ_{ij} is indicator var

(2) independent

$$E\Big[\sum_{\substack{i<j\\k<l\\\text{all 4 distinct}}}\bar{\sigma}_{ij}\bar{\sigma}_{kl}\Big] \le \sum E[\bar{\sigma}_{ij}]\,E[\bar{\sigma}_{kl}] = 0$$

Trick helps here: $\Longrightarrow$ gets rid of lots of terms

(3) $$E\Big[\sum \bar{\sigma}_{ij}\bar{\sigma}_{il}\Big] \le E\Big[\sum_{\substack{i,j,l\\\text{distinct}}}\sigma_{ij}\cdot\sigma_{il}\Big] = \sum_{\substack{i,j,l\\\text{distinct}}} \Pr[X_i = X_j = X_l]$$

$$\le \binom{s}{3}\sum_x p(x)^3$$

expected # 3-way collisions

$$\le \frac{s^3}{6}\Big(\sum_x p(x)^2\Big)^{3/2}$$

$$\le \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}\left(\|p\|_2^2\right)^{3/2}$$ by the facts

$$\frac{1}{6}\left(s^2\right)^{3/2} < \frac{\left(3\binom{s}{2}\right)^{3/2}}{6} = \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}$$

4 same as 3
5
6

In total:

$$\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] = \mathrm{Var}\left[\sum_{i<j} \bar{\sigma}_{ij}\right]$$

$$\leq \binom{s}{2}\|p\|_2^2 + 0 + 4\cdot\frac{\sqrt{3}}{2}\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$$

$$\leq 4\left[\binom{s}{2}\|p\|_2^2\right]^{3/2}$$

∎

Putting lemma into Chebyshev:

use $\rho = \frac{\varepsilon^2}{2}$

$$\Pr\left[\left|\frac{1}{c} - \|p\|_2^2\right| > \frac{\varepsilon^2}{2}\right] \le \frac{Var[\hat{c}]}{\varepsilon^4} \cdot 4$$

note $\frac{1}{\left(\frac{s}{2}\right)^{4/2}} \le \frac{1}{\sqrt{\frac{s^2}{2}}} \le \frac{2}{s}$

$$\le \frac{4\left[\left(\frac{s}{2}\right)\|p\|_2^2\right]^{3/2}}{\left(\frac{s}{2}\right)^2 \quad \varepsilon^4} \cdot 4 \le \underbrace{\frac{32}{\varepsilon^4} \cdot \frac{1}{s}}_{\text{also want this to be } \le 1} \cdot \underbrace{\|p\|_2^3}_{\le 1}$$

recall this comes from const in Proof of $\hat{c}$

So Pick $s \ge \Omega\left(\frac{1}{\varepsilon^4}\right)$

Note: Can get better bnd

1) Testing closeness to any known distribution — reduce to uniform case!

2) lower bound

How to estimate $\|p-U\|_1$?

1) $\|p-U\|_1 = 0 \Leftrightarrow \|p-U\|_2^2 = 0 \Leftrightarrow \|p\|_2^2 = \frac{1}{n}$

2) if $\|p-U\|_1 > \varepsilon \Rightarrow \|p-U\|_2 > \frac{\varepsilon}{\sqrt{n}}$

$$\Rightarrow \|p-U\|_2^2 > \frac{\varepsilon^2}{n}$$

$$\Rightarrow \|p\|_2^2 \geq \frac{1}{n} + \frac{\varepsilon^2}{n}$$

either additive estimate with error $\leq \frac{\varepsilon^2}{2n}$

or mult error $\leq \left(1 \pm \frac{\varepsilon^2}{3}\right)$

suffices

would have this if have additive error $\leq \frac{\varepsilon^2}{3n} \cdot \|p\|_2^2$

to get additive error $\leq \frac{\varepsilon^2}{3n} \|p\|_2^2$

suffices to have

$$s \geq \frac{\text{const} \cdot \sqrt{n}}{\varepsilon^2} \quad \text{samples}$$

since $\Pr\left[\,|\hat{c} - \|p\|_2^2| \geq \gamma \|p\|_2^2\,\right] \leq \frac{\overset{\text{const}}{k} \cdot \|p\|_2^3}{s \cdot \gamma^2 \left(\|p\|_2^2\right)^2} \leq \frac{k}{s \cdot \gamma^2 \cdot \|p\|_2}$

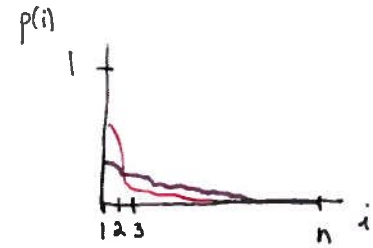
[note $\|p\|_2^2 > \frac{1}{n}$ so $\|p\|_2 > \frac{1}{\sqrt{n}}$ so $\frac{1}{\|p\|_2} < \sqrt{n}$]

$$\leq \frac{k \cdot \sqrt{n}}{s \cdot \gamma^2}$$

[note: we need $\gamma \approx \frac{\varepsilon^2}{3}$]

so picking $s \gg \frac{\sqrt{n}}{\varepsilon^4}$ will give small probability of error $\Longrightarrow \approx \frac{k \cdot \sqrt{n}}{s} \cdot \frac{1}{\varepsilon^4}$

Testing & Learning Monotone Distributions (over totally ordered domain)



Def. p over $[n]$ is "monotone decreasing"
if $\forall\, i \in [n-1]$ $\quad p(i) \geq p(i+1)$

Monotonicity Tester:

- if p monotone increasing, Pass with prob $\geq 3/4$
- if p ε-far in L_1 dist from mon increasing, Fail with prob $\geq 3/4$

Useful tool: "Birge Decomposition"

(note: this is a different decomposition than in homework (upcoming)
in particular, it is oblivious!)

decompose domain $1..n$ into $\ell = \Theta\left(\frac{\log \varepsilon n}{\varepsilon}\right) \approx \Theta\left(\frac{\log n}{\varepsilon}\right)$ intervals

$$I_1^\varepsilon, I_2^\varepsilon, \ldots, I_\ell^\varepsilon \quad \text{s.t.}$$

$$|I_{k+1}^\varepsilon| = \lfloor (1+\varepsilon)^i \rfloor$$

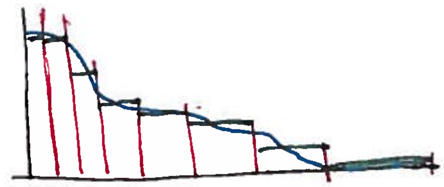
← will drop ε in notation once it is fixed

$|I_1^\varepsilon| = |I_2^\varepsilon| = \ldots = 1$

$|I_a^\varepsilon| = |I_{a+1}^\varepsilon| = \ldots = 2$

$\vdots$

but then at some point the sizes grow exponentially define "flattened distribution"



$\forall\ 1 \le j \le \ell$

$\forall\ i \in I_j$ $\qquad \tilde{q}_\varepsilon(i) = \dfrac{q(I_j)}{|I_j|}$ ← assign all elements in same interval the same probability

note: $q(I_j) = \tilde{q}_\varepsilon(I_j)$

Birge's Thm if q mon decreasing then $\|\tilde{q}_\varepsilon - q\|_1 < \varepsilon$

Corr if q ε-close to mon decreasing then $\|\tilde{q}_\varepsilon - q\|_1 < O(\varepsilon)$

Testing Algorithm:

how can we do this? $\tilde{q}$ isn't even if q monotone, exactly uniform. See problem from next hw set.

Take samples of q

do uniformity test for each partition (using samples that fell in it), fail if any partition fails (if not enough samples then pass)

$w_j \leftarrow$ # samples that fell in partition j

use LP to verify w close to monotone

↑ note this is LP on $O(\log n)$ vars

How many samples?

for each partition with enough weight, say $\frac{\varepsilon}{\log n}$, need $\frac{\sqrt{n}}{\varepsilon^2}$ samples

↑ need $\frac{\sqrt{n} \cdot \log n}{\varepsilon^3}$ for each one

$\approx O(\sqrt{n}\ \mathrm{polylog}\ n \cdot \mathrm{poly}\ \frac{1}{\varepsilon})$

need another $\log\log n$ for union bound

(note: this can be improved!!)

Last step:

Sampling error might make w_j's look non monotone

difficulty

purple is not monotone but is close

good thing: only $\frac{\log n}{\sqrt{\varepsilon}}$ variables!

Can be solved via brute force

LP (actually quite efficient)

⋮

So: monotone p likely to pass

ε-far ~~from~~ monotone p: either (1) non uniform in buckets or (2) w far from monotone

Slightly changing perspective...

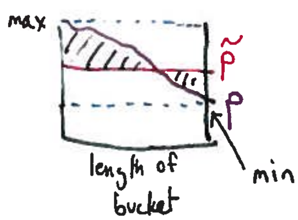
What if we know dist q is monotone, can we learn it?

Yes! use sampling to estimate $\tilde{q}_\varepsilon(I_j)$'s

Birge's Thm $\Rightarrow$ Can learn monotone distributions to w/in $\le \varepsilon$ L_1 error in $\Theta(\frac{1}{\varepsilon^3} \log n)$ samples.

Proof of Birge's Thm :

Error in bucket



gross upper bound on error:
$\leq$ (max−min) · bucket length

Partition of Intervals:

- Size 1 Intervals $|I_j| = 1$
- Short Intervals $|I_j| < 1/\varepsilon$
- Long Intervals $|I_j| \geq 1/\varepsilon$

← if we have any short intervals, there are $\Omega(1/\varepsilon)$ of these
if not, we can learn the distribution

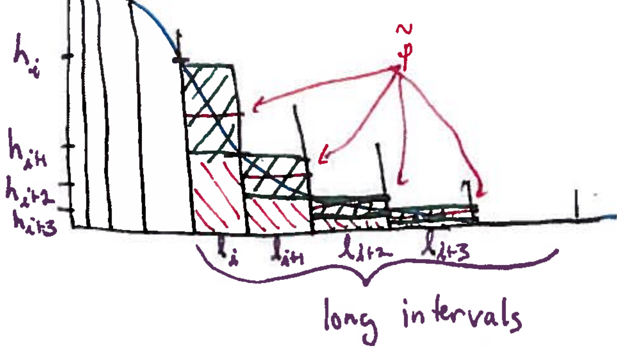
← if we have these then
max prob $\leq \varepsilon$ (since # size 1 intervals is $\Omega(1/\varepsilon)$)

↑ therefore min size 1 interval has prob $\leq \varepsilon$ which upper bounds later probabilities too since p is monotone

$$\text{total error} \leq \sum_{j=1}^{\ell} |I_j| \cdot (\text{max prob in } I_j - \text{min prob in } I_j)$$

$$= \underbrace{\sum_{\substack{\text{size 1} \\ \text{intervals}}} 1 \cdot 0}_{\substack{0 \\ \text{since no difference}}} + \underbrace{\sum_{\substack{\text{short} \\ \text{intervals}}} |I_j| (\text{max} - \text{min})}_{\text{omitted: idea is bound similarly to the long intervals but need to group together intervals of same size}} + \underbrace{\sum_{\substack{\text{long} \\ \text{intervals}}} |I_j| (\text{max} - \text{min})}_{\text{see below}}$$

Picture for long intervals:



green rectangles = upper bnd on error

$$\text{error} \leq (h_i - h_{i+1}) \cdot \ell_i + (h_{i+1} - h_{i+2}) \ell_{i+1} + (h_{i+2} - h_{i+3}) \ell_{i+2} + \ldots$$

$$= \underbrace{h_i \ell_i}_{\text{all } h_i\text{'s in this area are } < \varepsilon!} + h_{i+1}(\ell_{i+1} - \ell_i) + h_{i+2}\underbrace{(\ell_{i+2} - \ell_{i+1})}_{\text{positive, } \approx \varepsilon \cdot \ell_{i+1} \text{ by way that we partitioned}} + h_{i+3}(\ell_{i+3} - \ell_{i+2})$$

$$\leq \varepsilon \Big[\ell_i + \underbrace{\sum h_i \ell_{i-1}}_{\text{this is area of red rectangles, which is upper bounded by } p \text{ so sum is } \leq 1} \Big]$$

↑ get rid of this when bounding short intervals