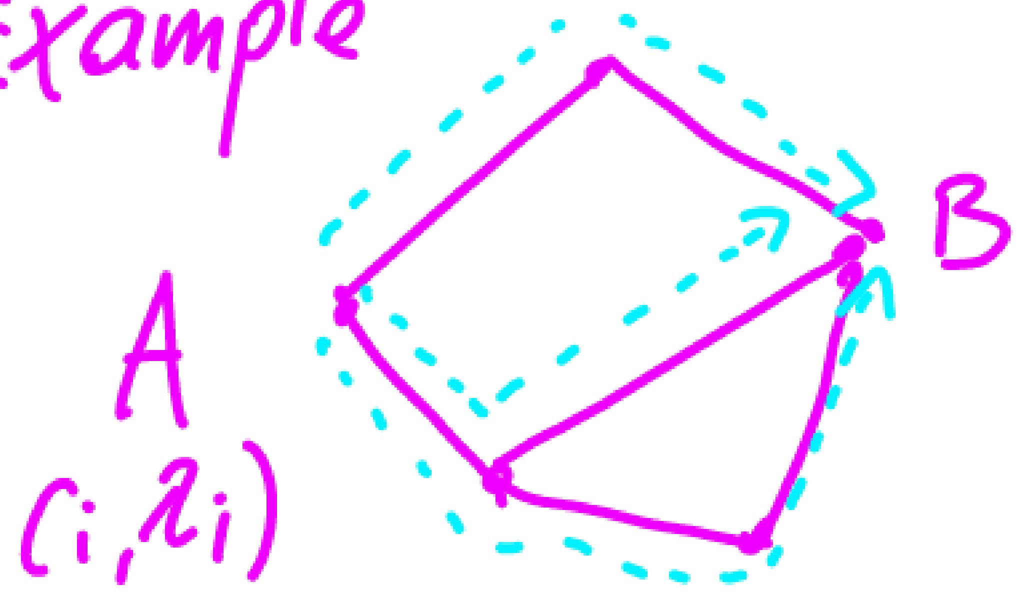
Approximate summation, operations on data sketches

Example



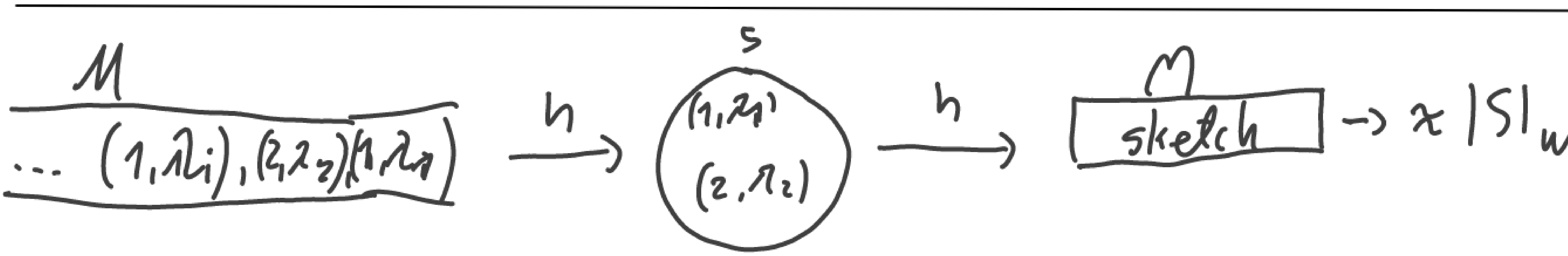
one node (B) can observe one packet many times.

- each node observes a stream $\mathcal{M} = (S, m)$, $m : S \to \mathbb{N}_+$
- $S = \{(1, \lambda_1), (2, \lambda_2), \dots, (n, \lambda_n)\}$, where

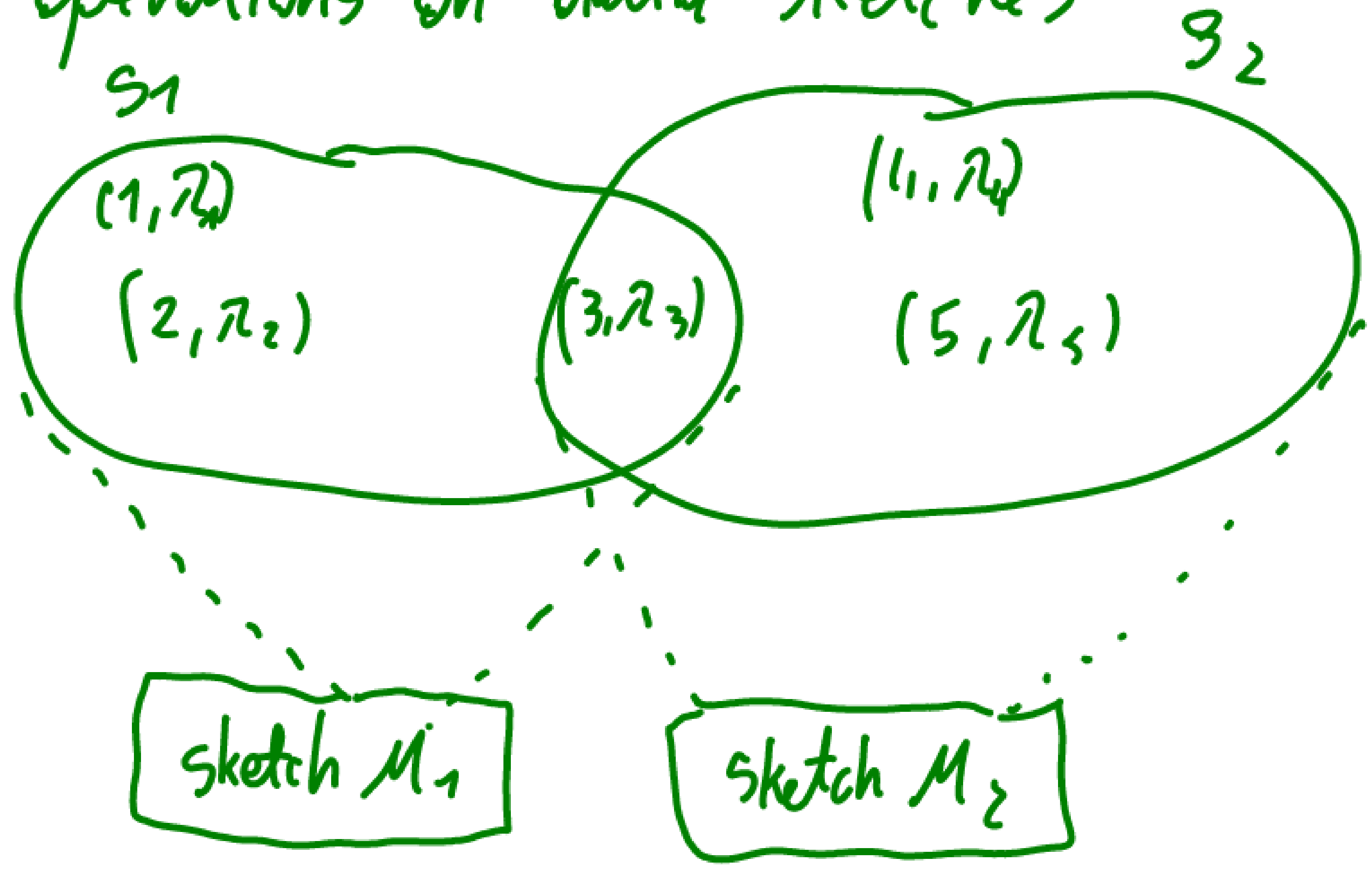
  for (i, λ_i), i - denotes a unique identifier of a packet
  $\lambda_i \in \mathbb{R}_+$ denotes size of a packet.

Aim 1: for each node we try to estimate

$$|S|_w := \sum_{(i,\lambda_i) \in S} \lambda_i$$



Aim 2: operations on data sketches



• Use sketches $\mathcal{M}_1$ and $\mathcal{M}_2$ to estimate the value of
$|S_1 \cap S_2|_w$, $|S_1 \cup S_2|_w$, $|S_1 \setminus S_2|_w$ very nice

---

ExpSketch($\mathcal{M}$, m, h):

$M = (\infty, \infty, \dots \infty)$

for each $(i, \lambda_i) \in \mathcal{M}$

  for each $k \in \{1, 2, \dots, m\}$

    $u \leftarrow h(i \| k)$ // concat of binary repr. we treat $h(i\|k) \sim U(0,1)$

    $M_k \leftarrow \min\left\{M_k, \frac{\ln(u)}{-\lambda_i}\right\}$

return M

$S_i^{(k)} \sim Exp(\lambda_i)$

<u>Lemma 1</u> Let $F(x)$ be CDF of some distribution and let
$F^{-1}(u) = \inf\{x : F(x) \geq u\}$, $0 \leq u < 1$.

smallest x so that $F(x) \geq u$

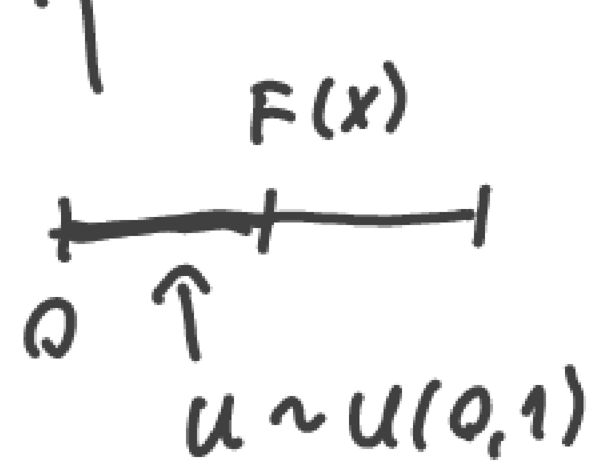
CDF
$F_{\mathcal{X}}(x) = \Pr[\mathcal{X} \leq x]$

then for $U \sim U(0,1)$ we have $F^{-1}(U) \sim F(x)$

Proof:

$\Pr[F^{-1}(U) \leq x] = \Pr[U \leq F(x)] = F(x)$

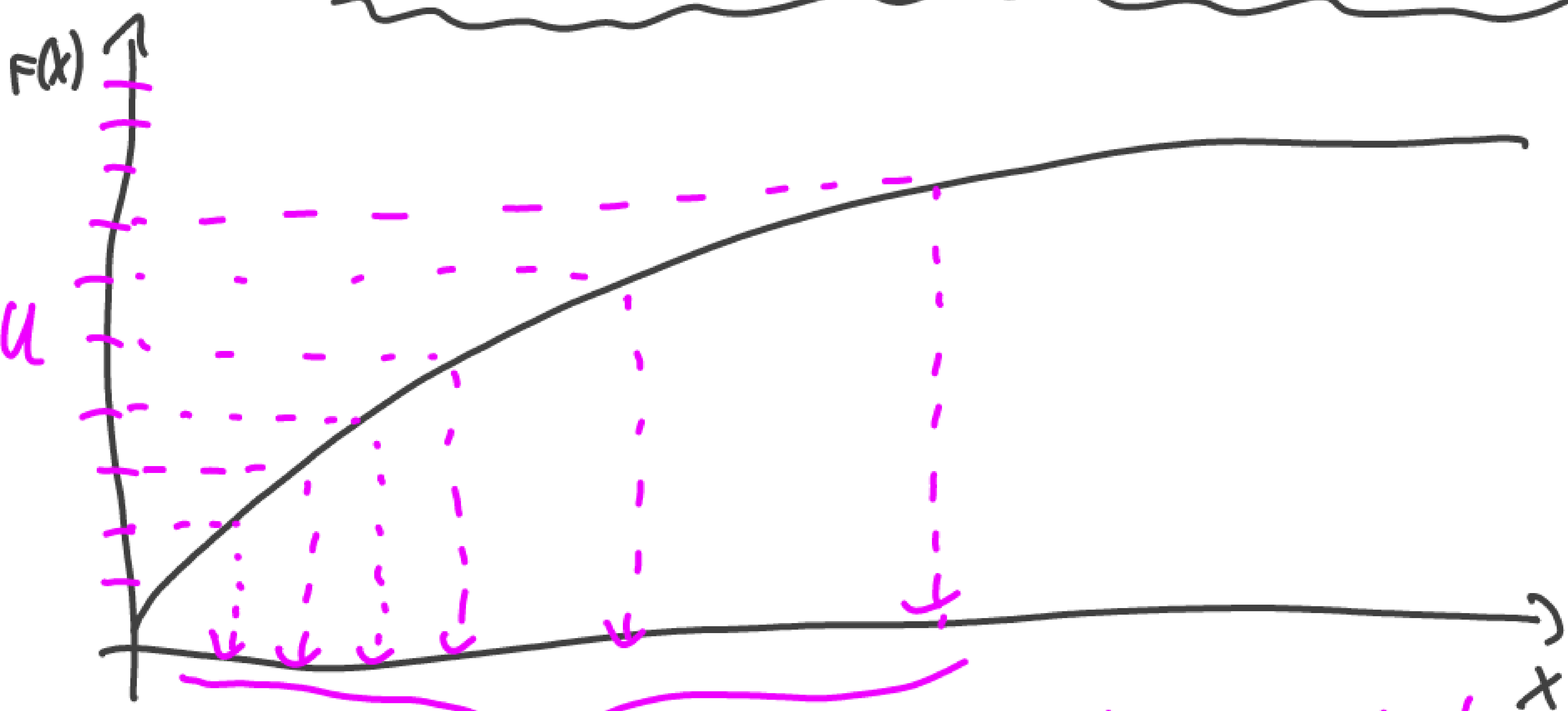
F is non-decreasing



we can use this to obtain any distribution by obtaining elements from Uniform distribution

Examples (why we need infimum in definition of $F^{-1}(x)$)

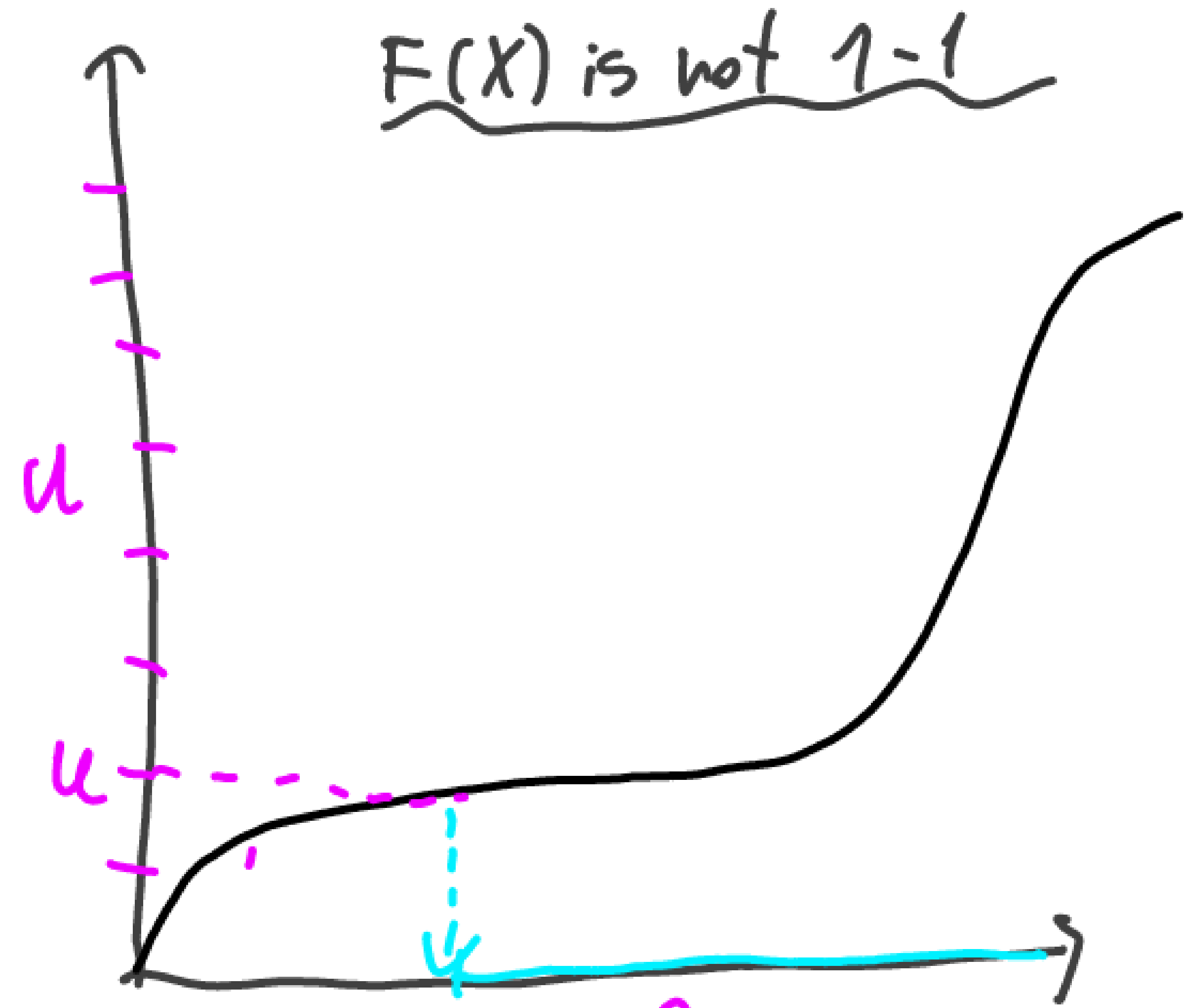
we assume that $F(x)$ is 1-1 and continuus



this is no longer u dist. it's F(x) dist

$F^{-1}(u) \sim F(x)$

F(X) is not 1-1

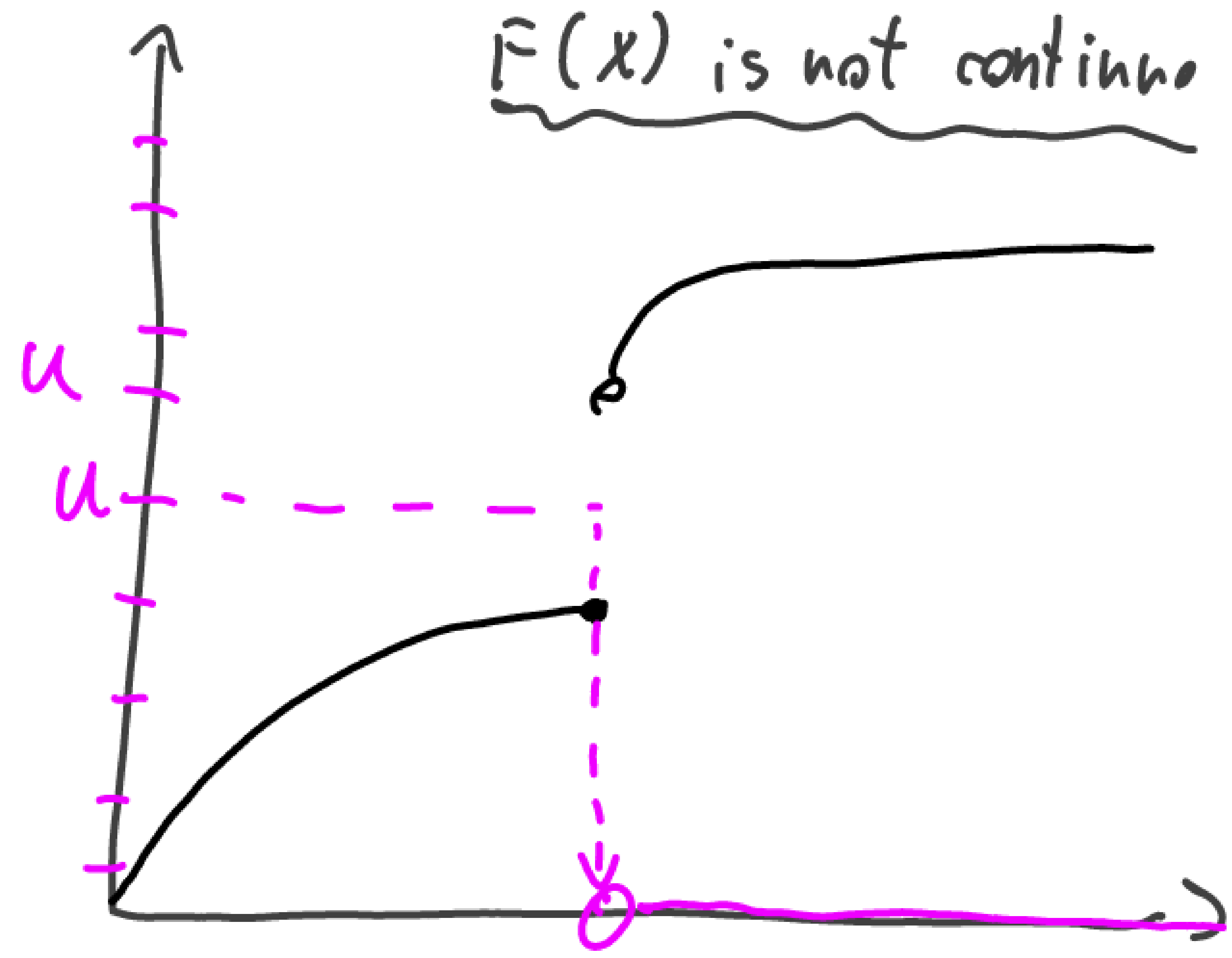


we can choose any point from here.

But from definition we choose x minimum

$\{x : F(x) \geq u\}$

F(X) is not continuo



X infimum for $\{x : F(x) \geq u\}$

Generating values from the exponential distribution

- $S \sim Exp(\lambda)$, $E[S] = \frac{1}{\lambda}$, $f(x) = \lambda e^{-\lambda x}$

CDF: $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, $\lambda > 0$

we look for inverse of F.

- $u = 1 - e^{-\lambda x} \rightarrow x = \frac{\ln(1-u)}{-\lambda} \rightarrow F^{-1}(u) = \frac{\ln(1-u)}{-\lambda}$

1-... doesn't change dist

- Lemma 1: $U \sim U(0,1) \rightarrow F^{-1}(U) \sim \frac{\ln(1-U)}{-\lambda} \sim \frac{\ln(U)}{-\lambda}$

$$\frac{\ln(U)}{-\lambda} \sim Exp(\lambda)$$

The next step is to derive the estimator

Estimator definition

$M = [\ |\ |\ |\ \dots\ |\ |\ |\ ]$

we think about $M[i]$ as results of indep. experiments

$M_1 = \min\{S_1^{(1)}, S_2^{(1)}, \dots S_n^{(1)}\}$

$M_m = \min\{S_1^{(m)}, S_2^{(m)}, \dots S_n^{(m)}\}$

$M_k = \min\{S_1^{(k)}, S_2^{(k)}, \dots S_n^{(k)}\}$, $S_i^{(k)} \sim Exp(\lambda_i)$

Lemma 2:

$|S|_w = \lambda_1 + \lambda_2 + \dots + \lambda_n = \Lambda \rightsquigarrow M_k \sim Exp(\Lambda)$

proof

$$F_{M_k}(x) = Pr[\min\{S_1, \dots, S_n\} \leq x] = 1 - Pr[S_1 > x \wedge S_2 > x \wedge \dots \wedge S_n > x] =$$

$$= 1 - \prod_{i=1}^{n} \underbrace{Pr[S_i > x]}_{= 1 - F(x)} = 1 - \prod_{i=1}^{n} e^{-\lambda_i x} = 1 - e^{-x \sum_{i=1}^{n} \lambda_i} = 1 - e^{-x\Lambda}$$

so for each element of sketch we know what's the distribution.

Conclusion

$$M_k \approx E[M_k]$$

$M_k \sim Exp(\lambda) \rightarrow E[M_k] = \frac{1}{\lambda} \longrightarrow \lambda \approx \frac{1}{M_k}$

from Hyperloglog we know that harmonic mean has better properties

$\cdot\ \bar{\lambda} = \dfrac{m}{\frac{1}{\left(\frac{1}{M_1}\right)} + \dots + \frac{1}{\left(\frac{1}{M_m}\right)}} = \dfrac{m}{\sum_{k=1}^{m} M_k}$ ← this estimator with the lowest variance.

Lemma 3

Let $G = \sum_{k=1}^{m} M_k \quad , M_k \sim Exp(\lambda)$

then $G \sim Gamma(m, \lambda)$, $m \in \mathbb{N}_+$, $\lambda \in \mathbb{R}_+$, $x \geq 0$

Proof (ex 25)

- $f_G(x) = \lambda e^{-\lambda x} \dfrac{(\lambda x)^{m-1}}{\Gamma(m)} \quad , \Gamma(m) = (m-1)!$ for $m \in \mathbb{N}$

- it is ok for $m=1$: $f_G(x) = \lambda e^{-\lambda x} \sim Exp(\lambda)$
- use induction to show that $\underbrace{G_m}_{} = \underbrace{G_{m-1}}_{Gamma(m-1, \lambda)} \overset{\text{convolution}}{+} M_1$ has distribution $Gamma(m, \lambda)$

tip: take integral

Lemma 4

$m \geq 2$, $\bar{\lambda} = \frac{m}{\sum_{k=1}^{m} M_k}$

we show that $E[\bar{\lambda}] = \frac{m}{m-1}\lambda$

so if we replace m with $m-1$ we get unbiased estimator

proof (ex. 26)

$$G = \sum_{k=1}^{m} M_k \sim \text{Gamma}(m, \lambda) \rightarrow$$

$$\rightarrow E[\bar{\lambda}] = E\left[\frac{m}{G}\right] \overset{*}{=} \int_0^\infty \frac{m}{x} f_G(x)\,dx = \ldots = \frac{m}{m-1}\lambda$$

$$* \; E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

Conclusion

$$\hat{\lambda} := \frac{m-1}{\sum_{k=1}^{m} M_k} \xrightarrow{\text{lemma 4}} E[\hat{\lambda}] = \lambda$$

best estimator

Lemma 5 (ex 27)

$m \geq 3$, $\tilde{\lambda} := \frac{m-1}{\sum M_k} \rightarrow SE[\tilde{\lambda}] = \frac{1}{\sqrt{m-2}}$

proof

- $SE[\tilde{\lambda}] = \sqrt{Var\left[\frac{\tilde{\lambda}}{\lambda}\right]}$
- $Var[X] = E[X^2] - E[X]^2$
- $E\left[\left(\frac{m-1}{G}\right)^2\right] \overset{*}{=} \ldots$

what we see as