

Def. The expectation of a random vector

$\underline{X} = (X_1, \dots, X_n)$  is  $E(\underline{X}) \triangleq [E(X_1), \dots, E(X_n)]$

- (a)  $g$  convex,  $\underline{X}$  random vector with finite  $E(\underline{X}) \rightarrow E[g(\underline{X})] \geq g[E(\underline{X})]$ . Jensen's Inequality
- (b)  $g$  concave  $\rightarrow E[g(\underline{X})] \leq g[E(\underline{X})]$ .

(attributed to Johan Jensen (1859-1925), Danish mathematician & engineer)

(14 Nov 19)

Applications of (3)

Suppose that  $X_1, \dots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(p)$ .

Then  $E(X_i) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = p$  and

$E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i) = np = \text{mean of Binomial}(n, p)$

Expectation  
of a product  
when the  
 $X_i$  are  
independent

$X_1, \dots, X_n$  independent  $n$  rv, each with 181  
finite  $E(X_i) \rightarrow$

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$$

Contrast this with a sum:  $E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$

whether the  $X_i$  are independent or not;

$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$  only when the  $X_i$

are independent.

Example

You have

a (Brita) water filter that you use to  
improve the taste of Santa Cruz water.

How much better would the filter do

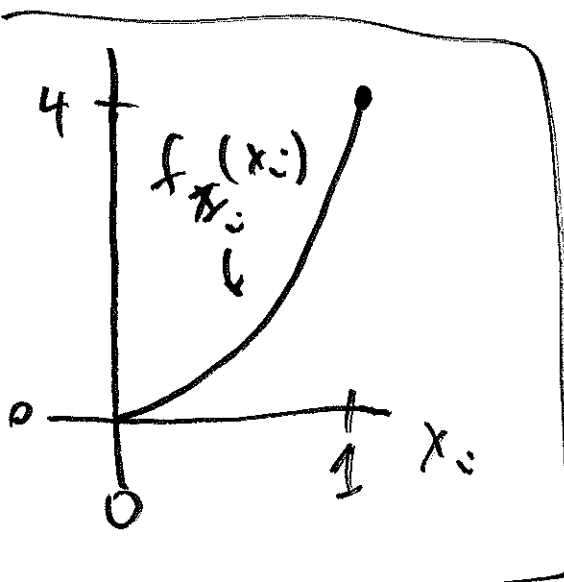
its job if you filtered the water twice  
instead of once?

$X_1$  = proportion of bad stuff removed in the 1<sup>st</sup> filtering (18%)

$X_2$  = proportion removed in 2<sup>nd</sup> filtering of what was left from 1<sup>st</sup> filtering

Reasonable to assume that  $X_1, X_2$  are independent; suppose they're IID with

common PDF  $f_{X_i}(x_i) = \begin{cases} 4x^3 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$



(sensible shape)

Let  $Y$  = proportion of original bad stuff remaining after 2 filtrations =  $(1-X_1)(1-X_2)$

Then  $E(Y) = E[(1-X_1)(1-X_2)] \stackrel{\text{independence}}{=} E(1-X_1) \cdot E(1-X_2)$

$X_1, X_2$  independent

$\leftrightarrow (1-X_1), (1-X_2)$

independent too

$E(1-X_1) \stackrel{\text{identical distribution}}{=} E(1-X_2) \triangleq \mu$ ;

then  $E(Y) = \mu^2$ .

$$\mu = E(1 - X_i) = \int_0^1 (1 - x_i) 4x_i^3 dx_i = 0.2, \quad (1F3)$$

so 80% of bad stuff expected to be removed in 1<sup>st</sup> filtering;  $E(I) = \mu^2 = 0.04$ , so expect only 4% of bad stuff to remain after 2 filterings.

(b) Suppose  
(9)

$X$  is a discrete rv with possible values  $0, 1, 2, \dots$ ; then  $E(X) = \sum_{x=0}^{\infty} P(X \geq x)$ .

(b) If  $X$  is a continuous rv with possible values  $(0, \infty)$ , then  $E(X) = \int_0^{\infty} [1 - F_X(x)] dx$ , and CDF  $F_X(x)$ ,

Example of b(9)

I throw a dart at a dartboard repeatedly, trying to get a bullseye (success).

$X = \#$  of throw on which I first succeed.

(Ex. throws FFS  $\rightarrow X=3$ ) Suppose that my 189

F = failure

S = success

success probability is constant across the throws and equals  $p$ , & throws are independent.

Then  $E(X)$  should be inversely related to  $p$ :

The worse I am, the longer I expect the

(1st duhrey)  
process to take;  $E(X) = ?$

Geometric distribution

At least 1 throw

always required so  $P(X \geq 1) = 1$ ; for  $n > 1$

(at least  $n$  throws required)  $\leftrightarrow$  (none of the first  $(n-1)$  throws succeeded)

so  $P(X \geq n) = (1-p)^{n-1}$  and Geometric series

$$E(X) = \sum_{n=1}^{\infty} (1-p)^{n-1} = 1 + (1-p) + (1-p)^2 + \dots$$

$$= \frac{1}{1 - (1-p)} = \frac{1}{p}$$

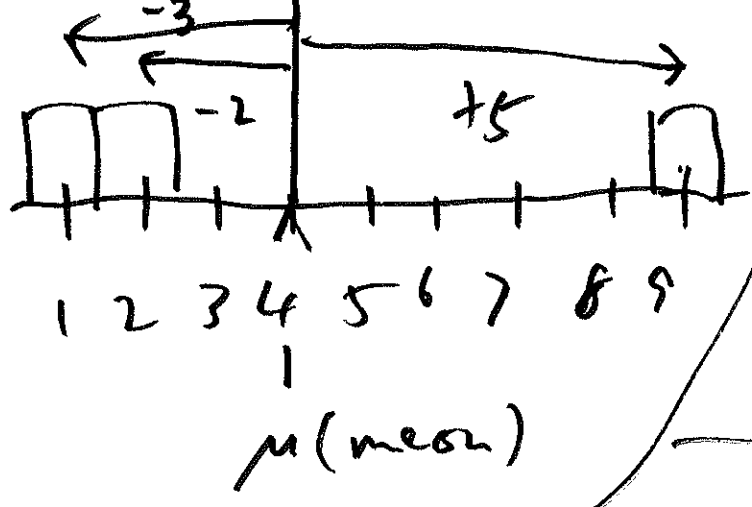
(inverse relation  $\checkmark$ )

If I'm terrible ( $p = .01$ )

I expect to succeed on

the  $\frac{1}{.01} = 100^{\text{th}}$  throw.

Variance and standard deviation



(185)  
 $\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$   
 mean  $4 = \mu$

$X$  discrete rv, uniform  $\{1, 2, 9\}$ ;  $E(X) = 4 = \mu$

d. How spread out is the dist. of  $X$  around its mean  $\mu$ ?  
 $(X - \mu) \sim \text{Uniform} \{-3, -2, +5\}$   
 ← deviation from  $\mu$

Could try calculating  $E(X - \mu)$ , but this is 0 for any rv  $X$ , because of cancellation of  $\oplus$  and  $\ominus$  deviations; two different

easy fixes:  $E|X - \mu| \stackrel{\Delta}{=} \text{mean average absolute deviation (AAD) (MAD)}$   
 Gauss (1800)  $\downarrow$   $E(X - \mu)^2 \stackrel{\Delta}{=} \text{variance of rv } X$   
 Laplace (1785)

AAD not used much; variance used constantly.

Def |  $X$  rv with finite mean  $E(X) = \mu$ ; (186)

variance of  $X = V(X) \triangleq E[(X - \mu)^2]$ .

DS we  
Var( $X$ )

If  $E(X) = \pm\infty$  or  $E(X)$  doesn't exist,  $V(X)$  doesn't exist.

One problem with variance

The units are messy: if  $X$  is in \$,  $V(X)$  is in \$<sup>2</sup>.

Easy fix: standard deviation  $\triangleq \sqrt{V(X)} \triangleq \underline{\underline{SD(X)}}$   
of  $X$  (K. Pearson [name])

Consequences of these definitions

$$\textcircled{1} V(X) = E[(X - \mu)^2]$$

$$= E(X^2 - 2\mu X + \mu^2)$$

$$= E(X^2) - 2\mu \underbrace{E(X)}_{\mu} + \mu^2$$

$$= E(X^2) - \mu^2 = E(X^2) - (E(X))^2$$

This is a different way to compute the variance

so  $V(X) = \left( \text{expectation of } X^2 \right) - \left( \text{square of expectation of } X \right)$  (187)

Toy example  $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$   $X \sim \text{Uniform}\{1, 2, 9\}$   
 mean  $\mu = 4$   $E(X - \mu)^2 = \frac{1}{3}(1-4)^2 + \frac{1}{3}(2-4)^2 + \frac{1}{3}(9-4)^2 = 12.7$   
 $\sqrt{}$   
 $SD(X) = \sqrt{12.7} = 3.6$   $= V(X)$

this is a reasonable summary of the lengths of the arrows

(2) For any rv  $X$ ,  $V(X) \geq 0$ ; if  $X$  is bounded,  $V(X)$  exists & is finite.

This is a consequence of Jensen's Inequality:  
 $f(x) = x^2$  is convex so  $E(X^2) \geq [E(X)]^2$ ,  
 ie.  $V(X) = E(X^2) - [E(X)]^2 \geq 0$ . concave



③  $V(X) = 0 \iff P(X=c) = 1$  for some constant  $c$  (this is a trivial rv)

Notation In the same way that, by

convention,  $E(X) = \mu_X$ ,  $V(X) \stackrel{\Delta}{=} \sigma_X^2$

and  $SD(X) \stackrel{\Delta}{=} \sigma_X$  (lower-case sigma)

④  $X$  rv,  $Y = aX + b$

$$\rightarrow V(Y) = a^2 V(X) = a^2 \sigma_X^2 \quad \text{and}$$

$$SD(Y) = |a| \sigma_X \quad (\text{for any constants } a, b)$$

Special cases  $a = 1$ :  $V(X+c) = V(X)$   
 $SD(X+c) = SD(X)$

$V(aX) = a^2 V(X)$   
 $(b=0) SD(aX) = |a| SD(X)$  ⑤ If  $X_1, \dots, X_n$  are independent rv with

finite means,  $V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i)$ .

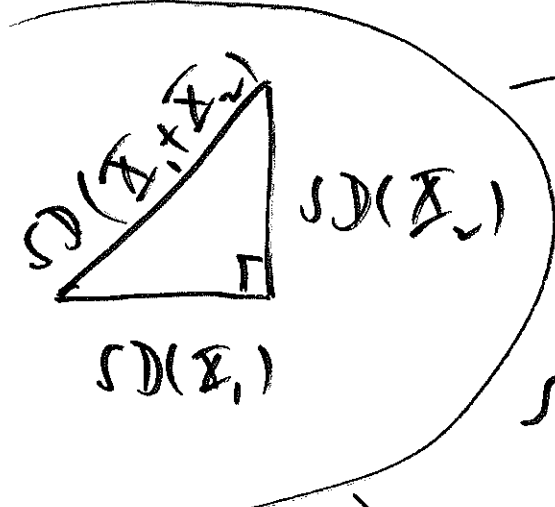
This is why the concept of variance (189) has endured even though the units of the variance are wrong: for

independent rvs, variance is additive

whereas  $\rightarrow$  SD is not correct units.

name of SD: Karl Pearson (1890)  
Special case of (5):

$X_1, X_2$  independent  $\rightarrow V(X_1 + X_2) = V(X_1) + V(X_2)$



so  
 $SD(X_1 + X_2) = \sqrt{[SD(X_1)]^2 + [SD(X_2)]^2}$

ie., SD grows like the hypotenuse of a right triangle.

Immediately,  $\max\{SD(X_1), SD(X_2)\} < SD(X_1 + X_2) < SD(X_1) + SD(X_2)$  (indep)

Consequence of (5)  $X_1, \dots, X_n$  independent r.v., (150)  
 $a_1, \dots, a_n, b$  constants  $\rightarrow$

$$V\left[\left(\sum_{i=1}^n a_i X_i\right) + b\right] = \sum_{i=1}^n a_i^2 V(X_i)$$

Example)  $X \sim \text{Binomial}(n, p)$ ; we already know that  $E(X) = np$ ; what about  $V(X)$  and  $SD(X)$ ?

Let  $S_i = \begin{cases} 1 & \text{if success on } i^{\text{th}} \text{ success} \\ 0 & \text{failure trial} \\ 0 & \text{else} \end{cases}$   
for  $(i=1, \dots, n)$  and suppose as usual that

$S_1, \dots, S_n$  are IID Bernoulli( $p$ ) —

then  $X = \sum_{i=1}^n S_i$  and we can work out its variance without difficulty.

$$V(\mathbf{X}) = V\left(\sum_{i=1}^n S_i\right) \stackrel{\text{independence}}{=} \sum_{i=1}^n V(S_i) \quad \text{so } \textcircled{19/}$$

we need to work out

the variance of a Bernoulli rv. we already know that  $E(S_j) = p$ , so if we use the formula  $V(S_j) = E(S_j^2) - [E(S_j)]^2$  we're halfway there.

Bernoulli rvs are funny:  $S_j = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1-p) \end{cases}$

so  $S_j^2 = \begin{cases} 1^2 = 1 & \text{with probability } p \\ 0^2 = 0 & \text{with probability } (1-p) \end{cases}$

so  $E(S_j^2) = E(S_j) = p$  and finally

$$V(S_j) = E(S_j^2) - [E(S_j)]^2 = p - p^2 = p(1-p)$$

and  $V(\bar{X}) = \sum_{i=1}^n V(S_i) = \sum_{i=1}^n p(1-p) = \boxed{np(1-p)}$  192

and  $SD(\bar{X}) = \sqrt{np(1-p)}$ . Example: T-S disease

$\bar{X} =$  (# T-S babies in family of  $n=5$ , both parents carriers so  $p = P(\text{T-S baby}) = \frac{1}{4}$ )

$\sim \text{Binomial}(n, p) = \text{Binomial}(5, \frac{1}{4})$

We already worked out that  $E(\bar{X}) = np = 1.25$

Now  $SD(\bar{X}) = \sqrt{np(1-p)} = \sqrt{5(\frac{1}{4})(\frac{3}{4})}$

It's useful to summarize  $\bar{X} = 0.97$   
 $\downarrow$   
 $= 1$

this by saying "The number of T-S babies

this couple will have will be around 1.25,  
 give or take about  $\pm \sigma_{\bar{X}}$  1.25  
↑  
 $\mu_{\bar{X}}$

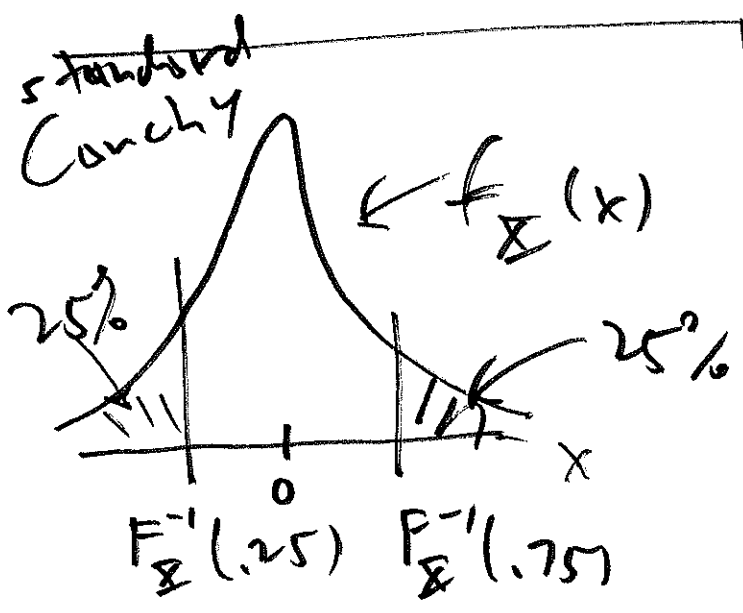
How do you measure the spread of a distribution if the variance doesn't exist?

Example  $X \sim$  (standard) Cauchy

$$f_X(x) = \left\{ \begin{array}{l} \frac{1}{\pi(1+x^2)} \\ \text{for all } -\infty < x < \infty \end{array} \right\}$$

Earlier we saw that  $E(X)$  doesn't exist, so clearly  $V(X)$  doesn't exist either.

But we can use the idea of quantiles on any dist., whether its variance exists or not.



Earlier we defined the interquartile range as

$$\underline{\text{IQR}} = F_X^{-1}(0.75) - F_X^{-1}(0.25)$$

standard  
Cauchy CDF is  $F_X(x) = \int_{-\infty}^x \frac{1}{\pi(1+t^2)} dt$  (194)

(arctangent)

Here  $\tan^{-1}(x)$  is

(calculator)  
 $= \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi}$

what's called the principal

inverse of  $\tan(x)$ , varying from  $-\frac{\pi}{2}$  to

$+\frac{\pi}{2}$  as  $-\infty < x < \infty$

Need to solve

$$F_X(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi} = p \text{ for } x;$$

$$\text{result is } x = F_X^{-1}(p) = \tan\left(\frac{p - \frac{1}{2}}{\frac{1}{\pi}}\right),$$

$$= -\cot(p\pi)$$

So the IdR  
standard  
for the Cauchy  
distribution is

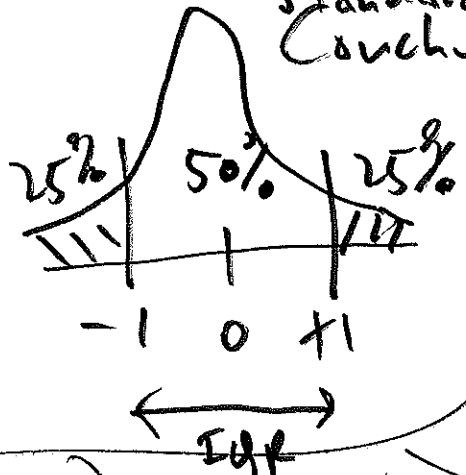
$$\text{IdR} = F_X^{-1}\left(\frac{3}{4}\right) - F_X^{-1}\left(\frac{1}{4}\right)$$

$$= \tan\left(\frac{\pi}{4}\right) - \tan\left(-\frac{\pi}{4}\right)$$

$$= 2.$$

standard  
Cauchy PDF

# Moments of a rv



$$E(X) = E(X^1)$$

$$V(X) = E(X^2) - [E(X^1)]^2$$

$$= E(X - \mu)^2$$

with <sup>2</sup> the

usual mathematical impulse to generalize:

Def.  $X$  rv,  $k$  integer  $\geq 1 \rightarrow$

$E(X^k) \triangleq$  the  $k^{\text{th}}$  moment of  $X$

of course  $E(X^k)$  may not exist, and if it does it may be infinite, but the idea is still useful.

You can show

that  $\left( \begin{array}{l} k^{\text{th}} \text{ moment} \\ \text{of } X \text{ exists} \end{array} \right) \iff E(|X|^k) < \infty$



Consequences  
of the  
moment  
definition

① IF  $E(|X|^k) < \infty$  for  $(196)$   
some integer  $k \geq 1$ , then  
 $E(|X|^j) < \infty$  for all integers  
 $j < k$ ;

in other words, if the  $k^{\text{th}}$   
moment of  $X$  exists, so do the  
 $(k-1)^{\text{st}}$ ,  $(k-2)^{\text{nd}}$ , ..., moments.

Definition

$X$  rv with expectation  $E(X) = \mu$ ,  $k$   
integer  $\geq 1 \rightarrow E[(X - \mu)^k]$  is called  
the  $k^{\text{th}}$  central moment of  $X$  or  
the  $k^{\text{th}}$  moment of  $X$  around its mean.

Clearly this idea generalizes the  
variance of  $X = E[(X - \mu)^2]$

②  $E[(X-\mu)^1] = E(X) - \mu = \mu - \mu = 0,$

i.e., every rv has 1<sup>st</sup> central moment 0.  
the dist. of

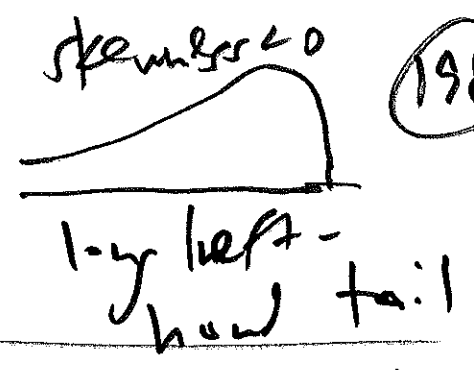
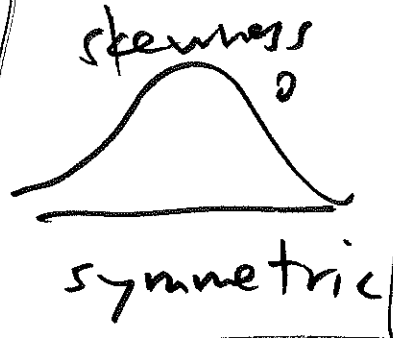
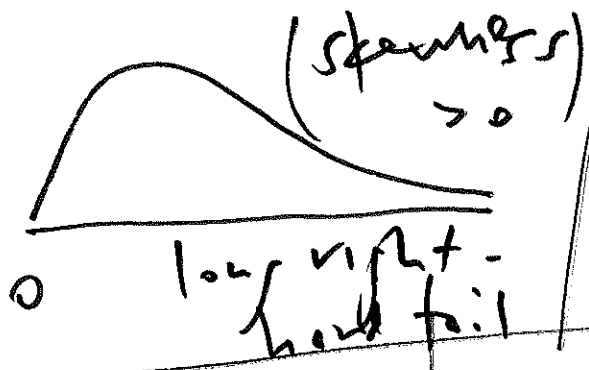
③ If  $X$  is symmetric around  $\mu_X$ ,

then  $E[(X-\mu)^k] = 0$  for all odd integers  $k$  for which  $E[(X-\mu)^k]$  exists

This motivates a new definition:

Def  $X$  rv with mean  $\mu_X$ , SD  $\sigma_X$ ;  
if the third moment of  $X$  exists and is finite, then skewness  $(X) \triangleq E \left[ \frac{X - \mu_X}{\sigma_X} \right]^3$

All symmetric distributions with finite 3<sup>rd</sup> moment have skewness 0.  
converting  $X$  to standard units



Moment generating functions

Def. If  $X$  is a real number  
 $\psi(t) \triangleq E(e^{tX})$  is called  
 the moment generating function of  $X$   
 (MGF)

The reason for this definition

Theorem If  $X$  is a real number with MGF  $\psi_X(t)$ , finite for all values of  $t$  in an open interval  $(-a, b)$  around 0 ( $a > 0$ ;  $b > 0$ );

then for all integers  $n > 0$ ,

$$E(X^n) = \left. \frac{d^n}{dt^n} \psi_X(t) \right|_{t=0}$$

$\leftarrow$   $n^{\text{th}}$  derivative of  $\psi_X$ , evaluated at  $t=0$ .

This is a handy theorem: if its premise is satisfied & the calculations are manageable, you get all the moments of  $X$  just by computing  $\psi_X(t)$  and differentiating it over & over.

(16 May 19)  
( $\lambda > 0$ )

Example

$X \sim \text{Exponential}(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & \text{else} \end{cases}$$

$$\psi_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \cdot \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{(t-\lambda)x} dx$$

Now this integral is finite only if  $t - \lambda < 0$ , is for  $t < \lambda$ , but this means (since  $\lambda > 0$ ) ~~finite~~  $-\lambda < 0 < \lambda$

that it's definitely finite in an open interval around 0 (eg.  $(-\lambda, \lambda)$ ).