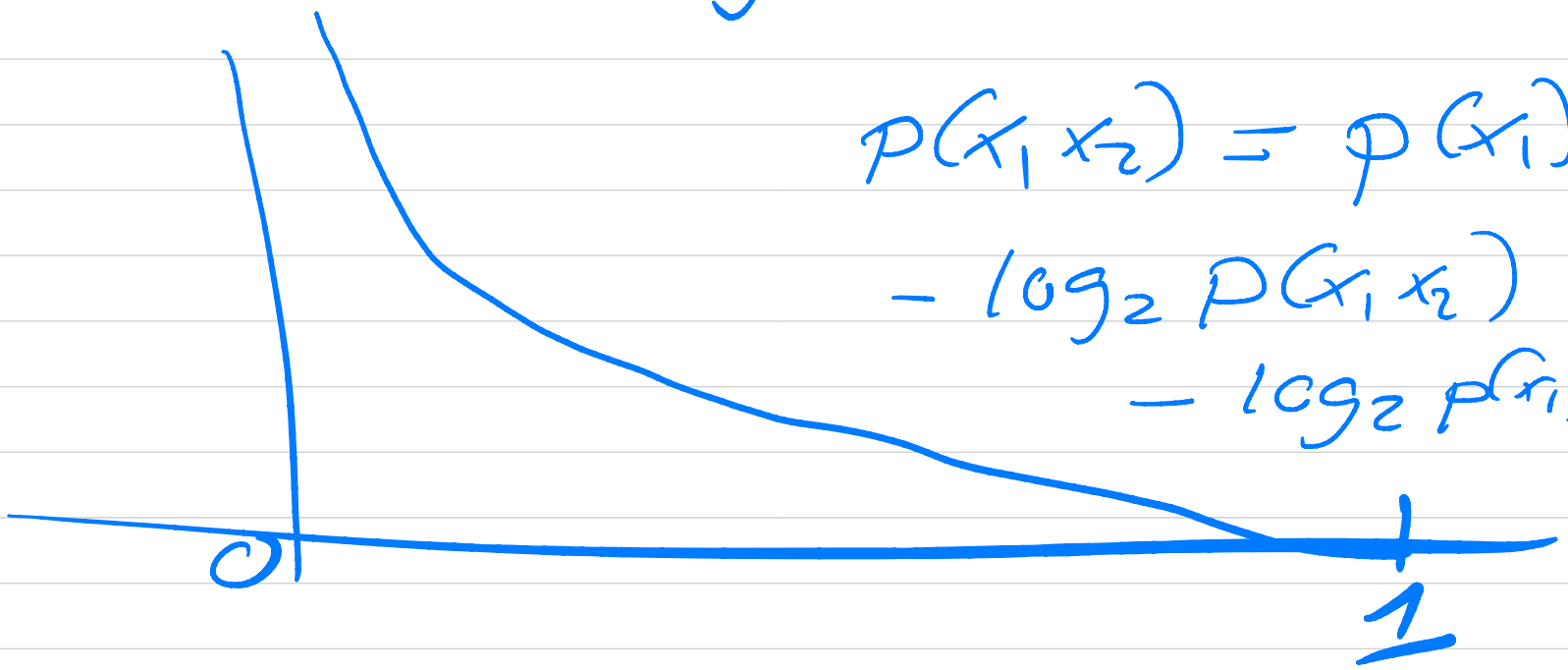
FYS-STK3155/4155 lecture, November 16, 2023

measure of "purity", 0 pure instance, 1 random selection

- additive
- continuous
- high for unlikely cases
- non-negative

$$S(x) = \mp \log_2 p(x) \qquad -\log_2 p(x)$$



$$p(x_1 x_2) = p(x_1)\, p(x_2)$$

$$-\log_2 p(x_1 x_2) =$$

$$-\log_2 p(x_1) - \log_2 p(x_2)$$

$$S = -\sum_i p_i \log_2 p_i$$

$$p_i \in [0, 1]$$

Example, two classes

p = True

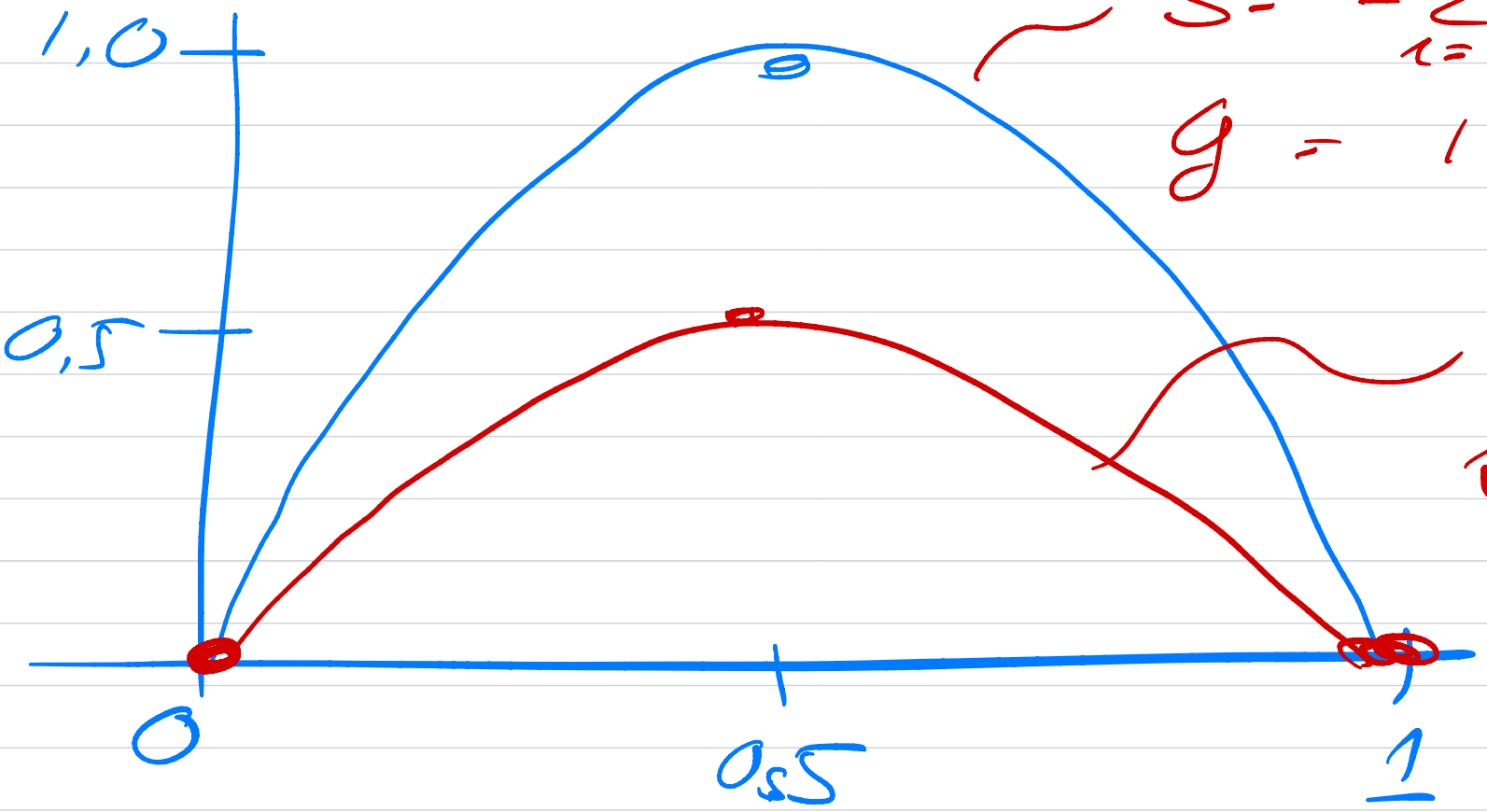
$1-p$ = False

$$S = -p \log_2 p - (1-p) \log_2 (1-p)$$

$p = 0 \Rightarrow S = -\log_2(1) = 0$

$p = 1 \Rightarrow S = -1 \log_2(1) = 0$

$p = 0.5 \Rightarrow S = \underline{1}$



$$S = -\sum_{i=1}^{c} p_i \log_2 p_i$$

$$g = 1 - \sum_{i=1}^{c} p_i^2$$

Simi index for binary case

Gini factor

$$g = \sum_{i=1}^{c} p_i (1 - p_i) \quad \Big| \quad \sum_{i=1}^{c} p_i = 1$$

$$= 1 - \sum_{i=1}^{c} p_i^2$$

$$g = 2p(1-p) \leftarrow c=2,\ p \text{ or } 1-p$$

Grade trend	Hours sleep	Hours studied	Grade
Above (1)	Low (0)	High (1)	1
Below (0)	High (1)	Low (0)	0
1	0	1	1
1	1	1	1
0	0	1	0
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	0
1	1	1	1

input features/attributes | output

Grade trend

$P(\text{trend} = \text{above}) = 6/10$

$P(\text{trend} = \text{below}) = 4/10$

if (past trend = above & grade = above)

$p = 4/6$

if (past trend = above, grade = below)

$p = 2/6$

gini index $= 1 - (4/6)^2 - (2/6)^2$

$= 0.45$

if (past trend negative, grade above)

$p = 0$

if (past trend negative, grade below)

$p = 4/4 = \underline{1} \Rightarrow g = 0$

weighted sum : $6/10 \times 0.45 +$

$4/10 \cdot 0 = 0.27$

Repeat for hours slept

$g = 0.47$

Hours studied

$g = 0.34$

Lowest is the link between grade and grade Trend. => root node is grade Trend

Grade Trend
$g = 0.27$
$IO = (6, 4)$

Above (A) 6

4 Below

T	Sleep	Studied	Grade
A	L (0)	H (1)	A
A	L (0)	H	A
A	H (1)	H	A
A	L (0)	L	B
A	L (0)	L	B
A	H (1)	H	A

$P(\text{sleep} = H) = 2/6$

$P(\text{sleep} = L) = 4/6$

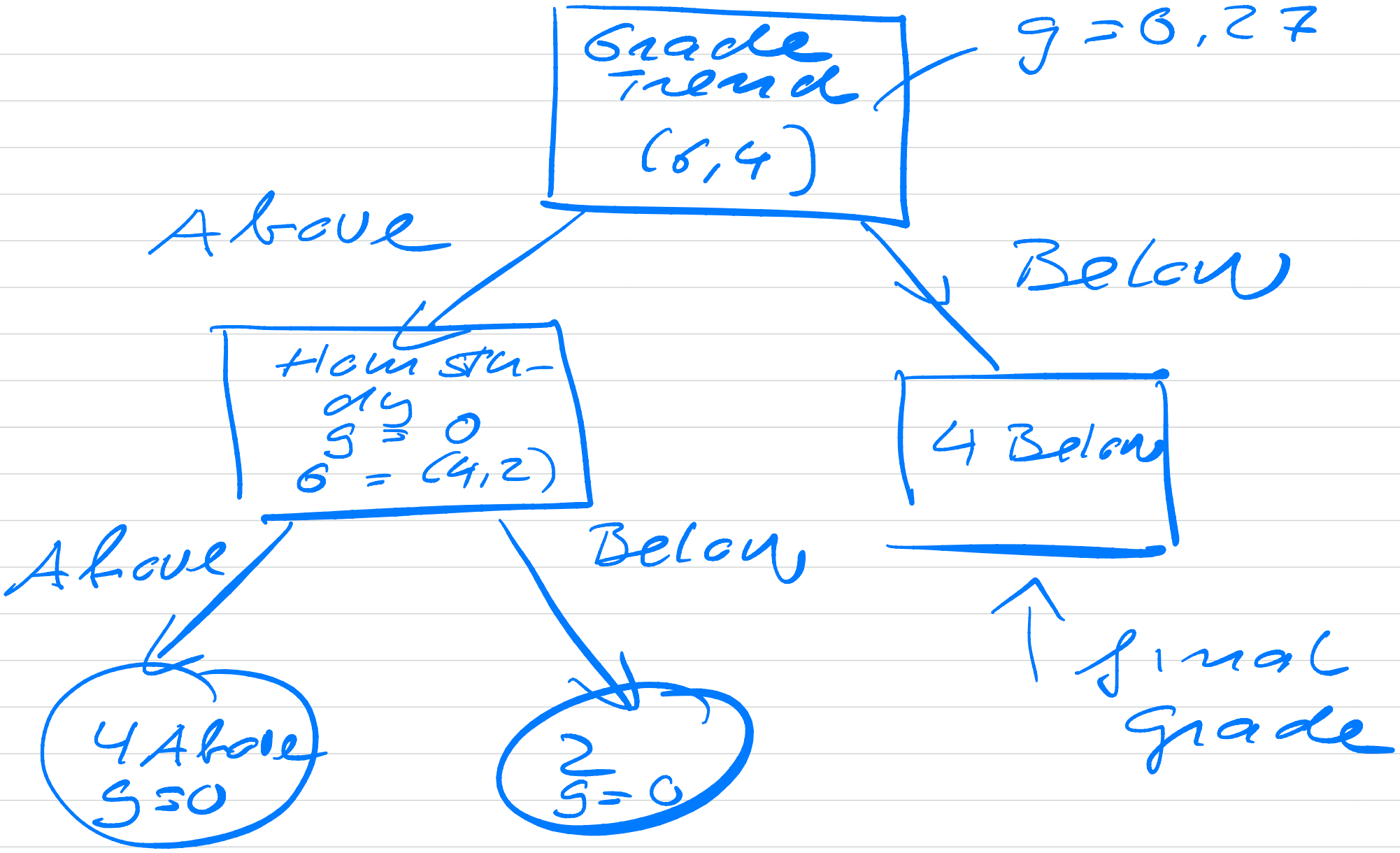
gini $1 - (1^2 + 0) = 0$

Hours studied
$g = 0$

hours slept $g = 0.33$

Trend	slept	studied	grade
B	H	L	B
B	L	H	B
B	H	H	B
B	L	H	B

$g_{slept} = 0$ $g_{studied} = 0$


Grade Trend
(6,4)
g = 6,27
Above
Below
Hom Stu-
dy
S = 0
6 = (4,2)
4 Below
Above
Below
4 Above
S=0
2
S=0
final
grade

Regression Tree