Interactive Online Learning

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

Announcements (all on Ed!)

1. Assignment 0 (survey released)

2. Lecture 1 slides + notes up on website

3. Office hours available:

Sanjiban (Tue/Thurs 11-12pm, Gates 413B)
Dhruv (Mon/Wed 11-12 pm, Rhodes 400)

Learning

Today!

Robot
Decision
Making

Reinforcement Learning

Imitation Learning

Interactive Online Learning

Meta Learning

Model Predictive Control

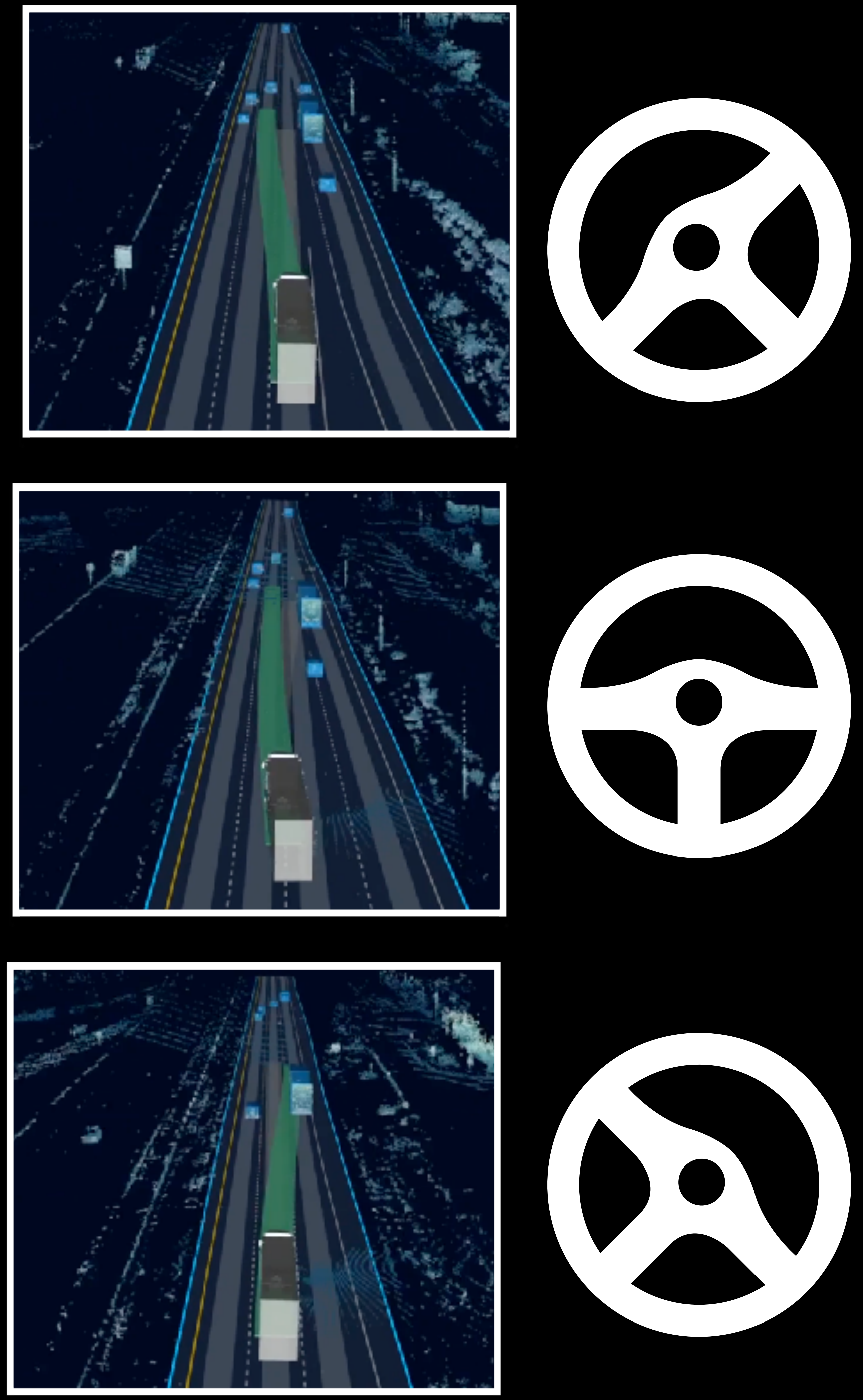
Anytime Planning

How humans learn …

Can't we collect a LOT of data and train robots *offline*?

SUPERVISED LEARNING

#1 Get Data



Input (s) Output (a)

#2
Train
Policy

$$\pi : s \rightarrow a$$

#3 Deploy!


Train ≠ Test

Activity!

Think-Pair-Share!

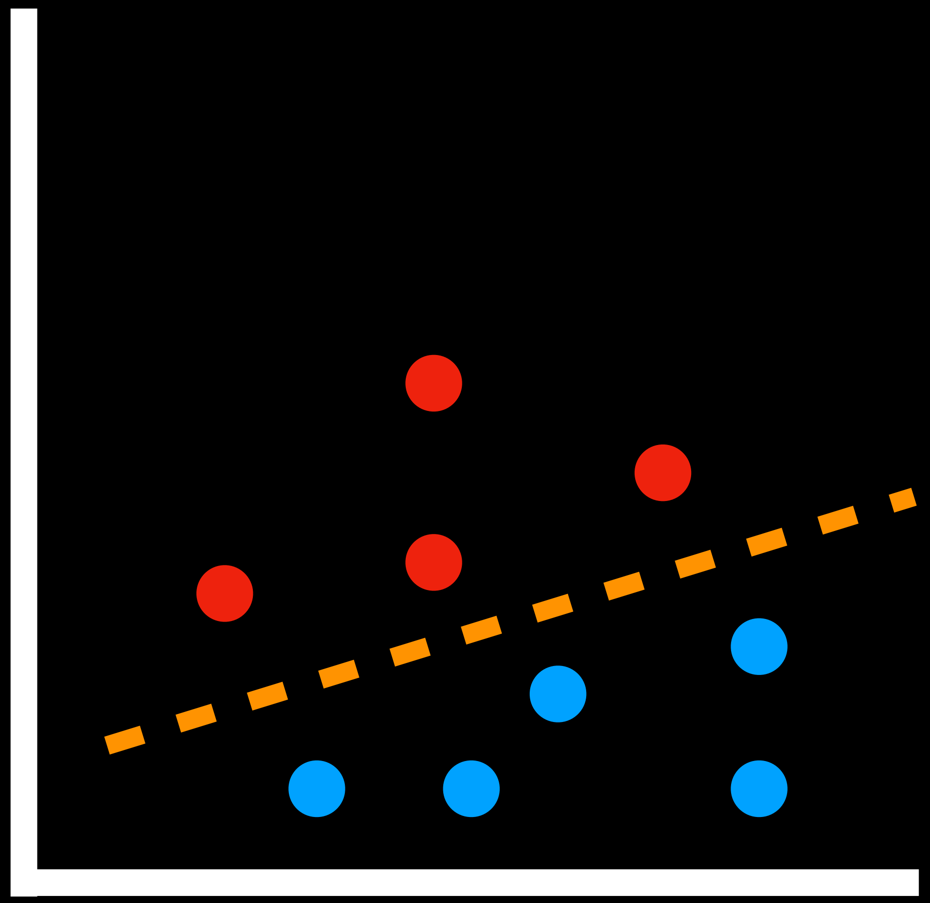
Think (30 sec): What are different sources of train-test mismatch?

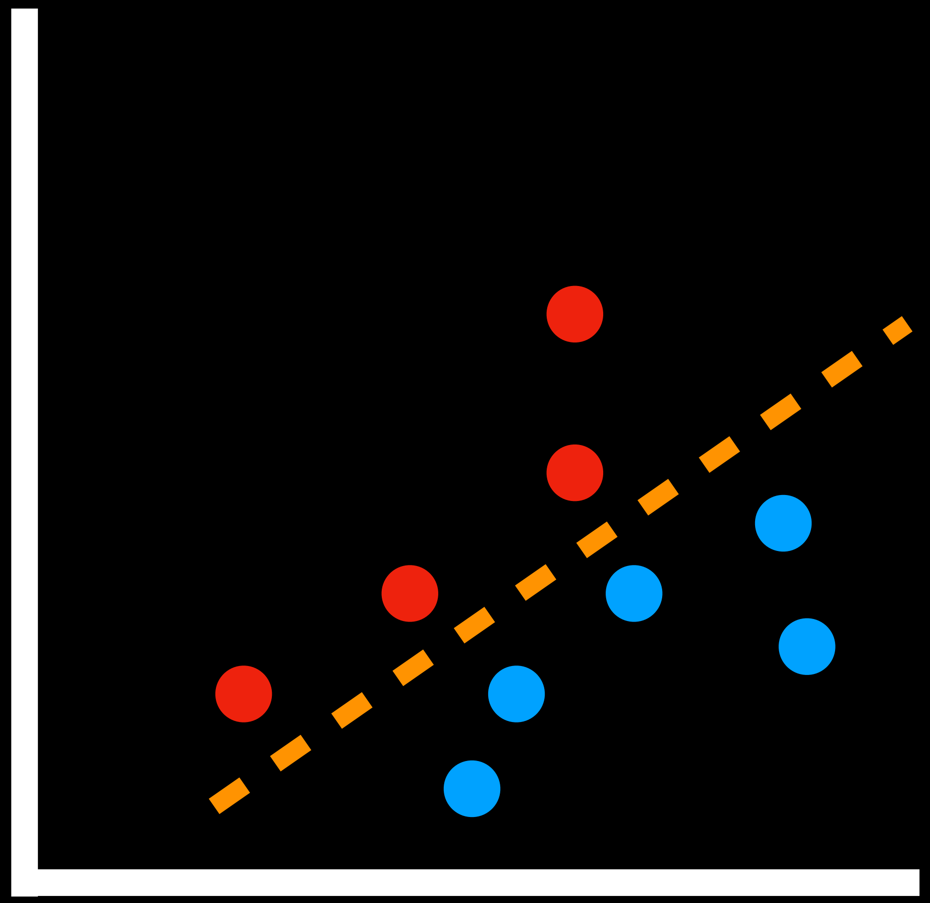
Pair: Find a partner

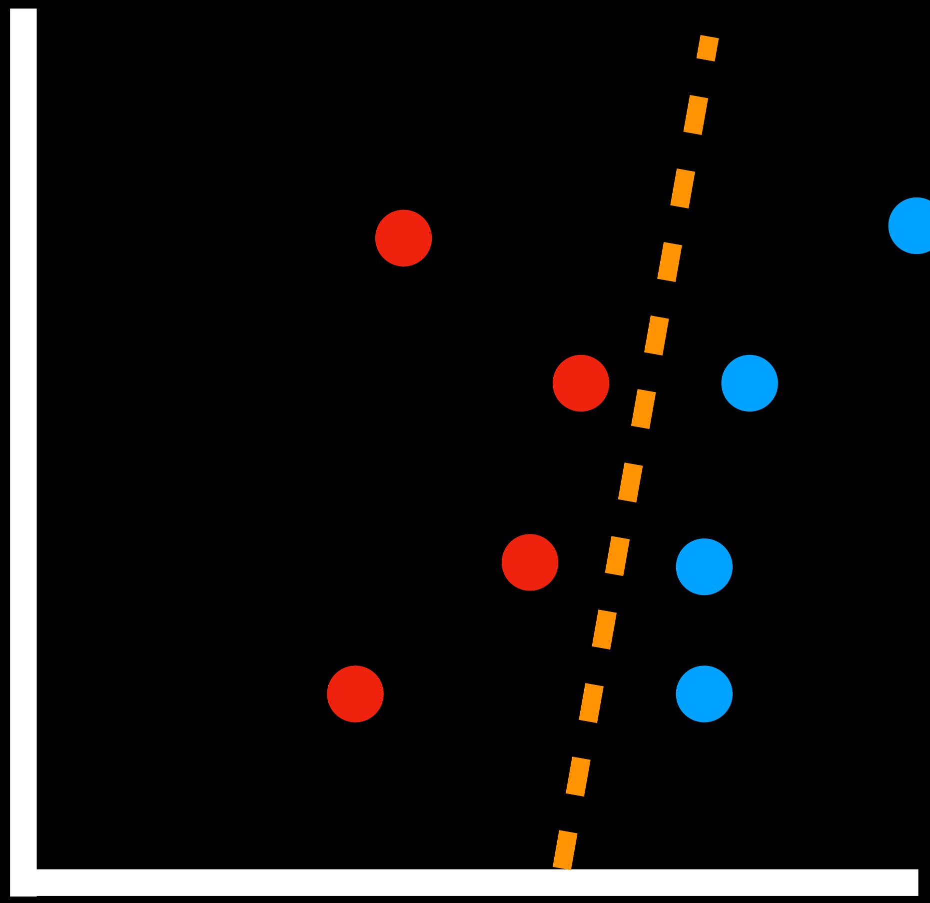
Share (45 sec): Partners exchange ideas

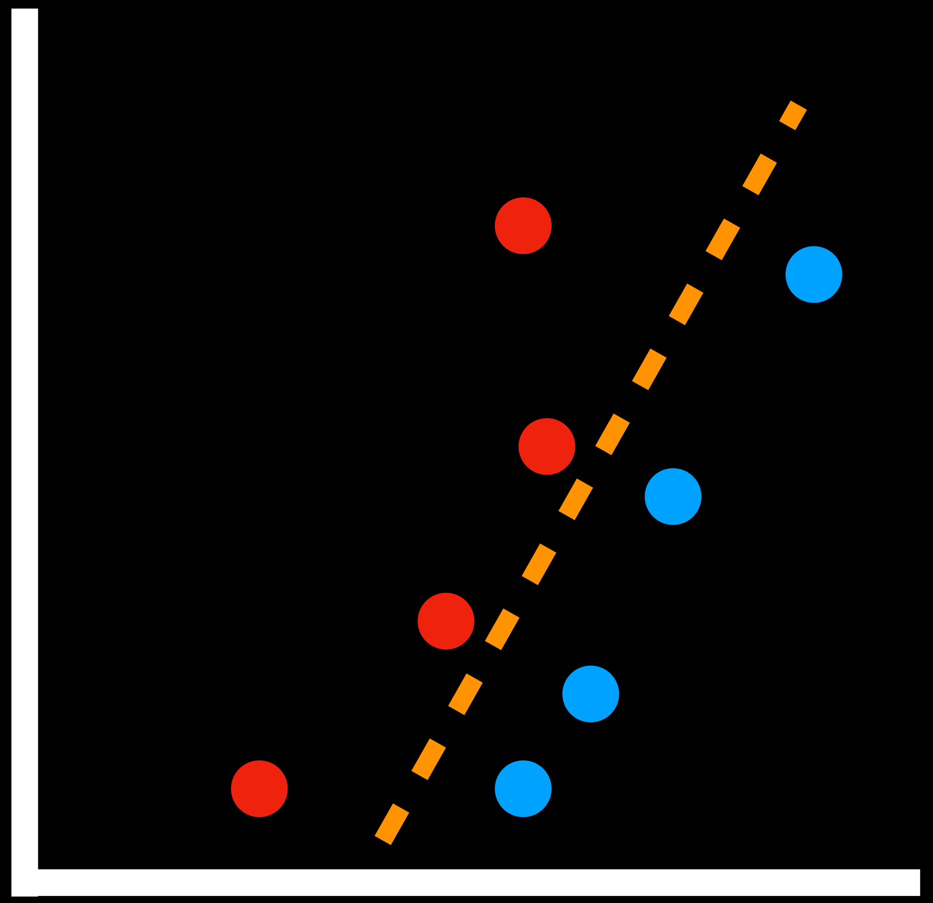
Train $\neq$ Test

Case 1: Data changes over time



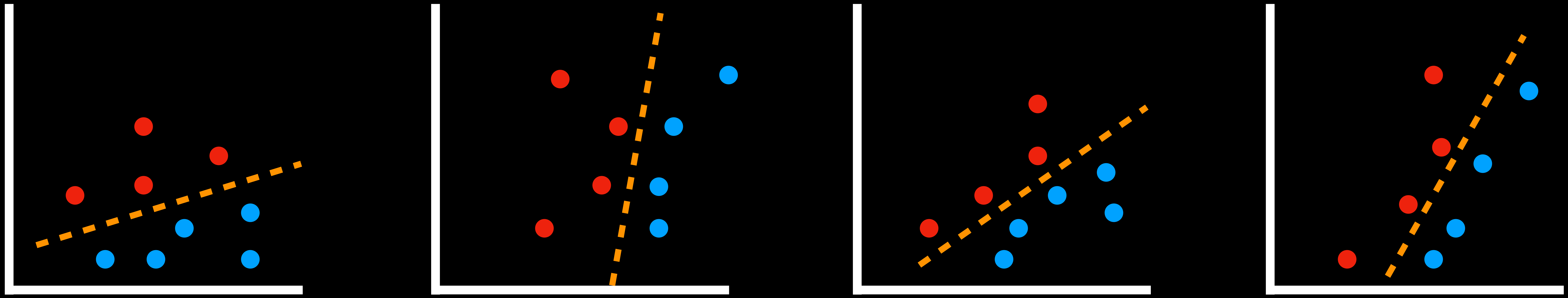






Case 2: Data changes with robot behavior

Case 3: Data changes adversarially (game)



Challenge:

Don’t know the test distribution upfront

Learner $\nleftarrow$ Collect Data

Interactive Learning

Interactive Learning

Learner		Adversary
Initialize policy	π_1 [policy] $\rightarrow$	
	$\leftarrow$ $l_1(\,.\,)$ [loss]	Chooses loss
Update policy	π_2 $\rightarrow$	
	$\leftarrow$ $l_2(\,.\,)$	Chooses loss
⋮		⋮

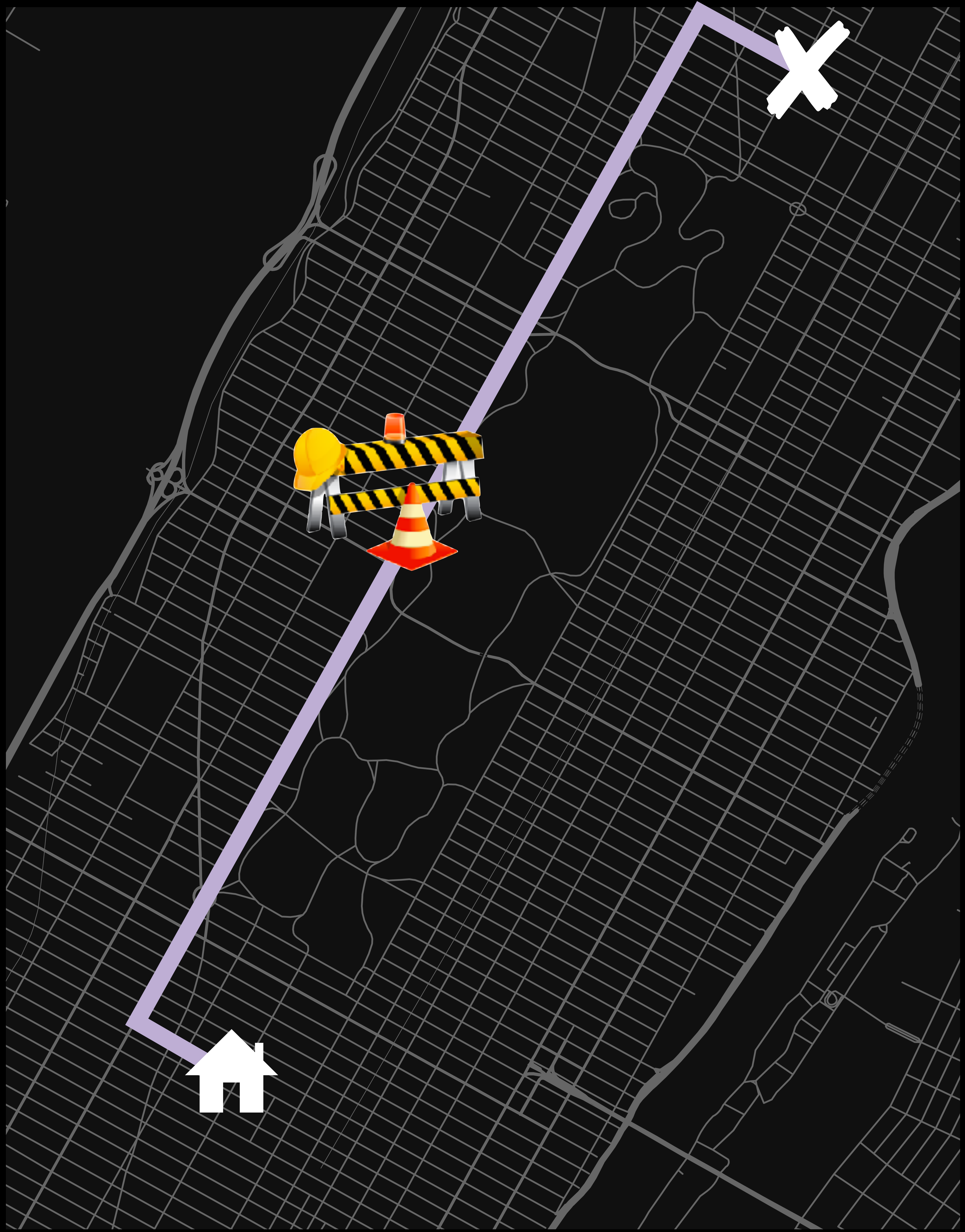
Prediction with Expert Advice

Expert 2

Expert 1

Expert 3

Expert 2



Loss = 1.0

Let's formalize!

$$\text{Regret} = \sum_{t=1}^{T} l_t(\pi_t) - \min_{\pi^*} \sum_{t=1}^{t} l_t(\pi^*)$$

(Learner) (Best in hindsight)

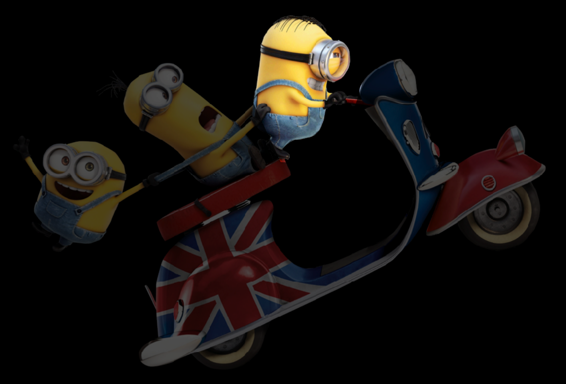
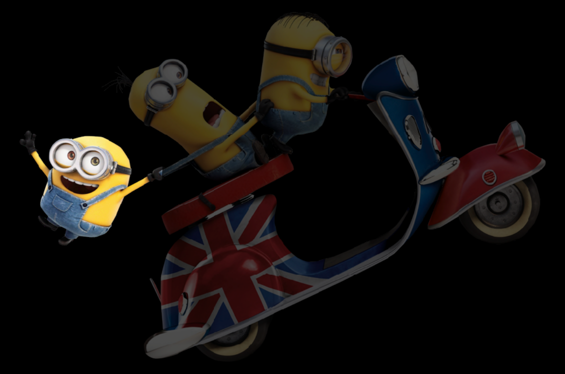
How do we design algorithms that are no-regret?

$$\text{Regret} = \sum_{t=1}^{T} l_t(\pi_t) - \min_{\pi^*} \sum_{t=1}^{t} l_t(\pi^*)$$

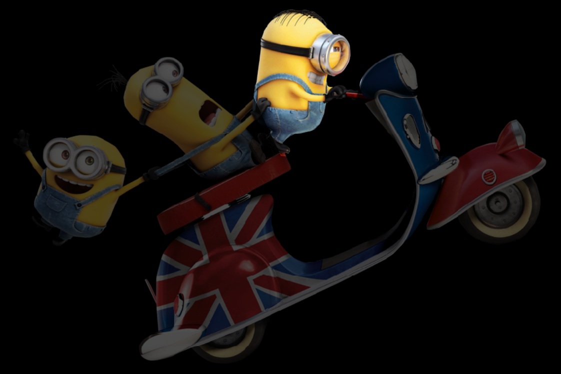
At every round t, choose the best expert in hindsight

$$\pi_t = \arg\min_{\pi} \sum_{i=1}^{t-1} l_i(\pi)$$

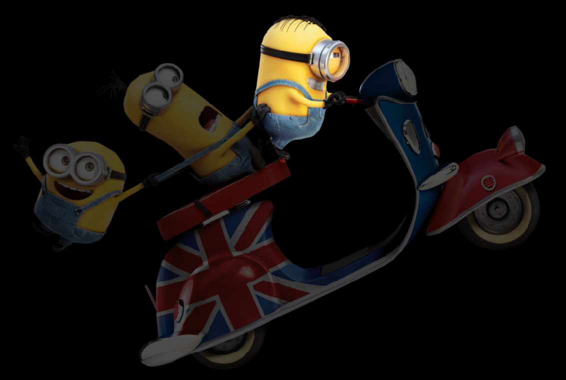
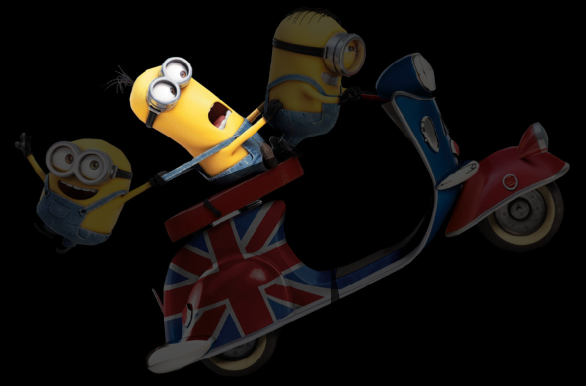
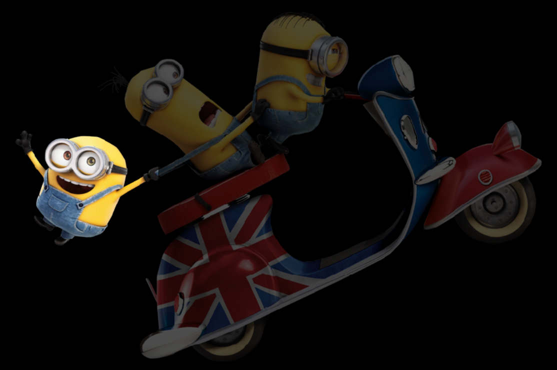
(lowest total loss)

$\sum l_t$		l_1
- -	Expert 1 	$\boxed{1.0}$
- -	Expert 2 	0.2
- -	Expert 3 	0.5

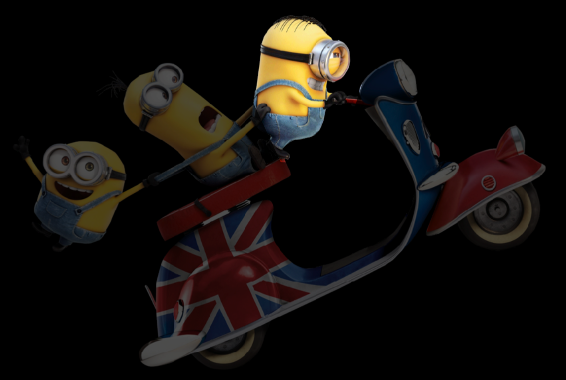
Avg. Regret: - -

$\sum l_t$		l_1	l_2
1.0	Expert 1 	$\boxed{1.0}$	0.5
0.2	Expert 2 	0.2	$\boxed{0.5}$
0.5	Expert 3 	0.5	0.2

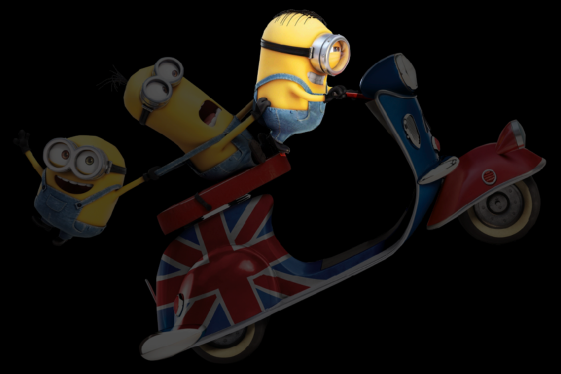
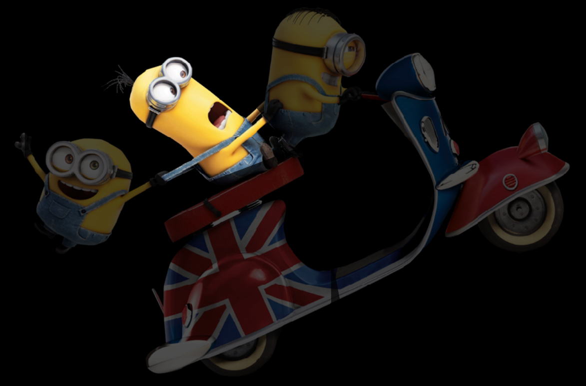
Avg. Regret: 0.80

$\sum l_t$		l_1	l_2	l_3
1.5	Expert 1 	$\boxed{1.0}$	0.5	0.5
0.7	Expert 2 	0.2	$\boxed{0.5}$	$\boxed{1.0}$
0.7	Expert 3 	0.5	0.2	0.2

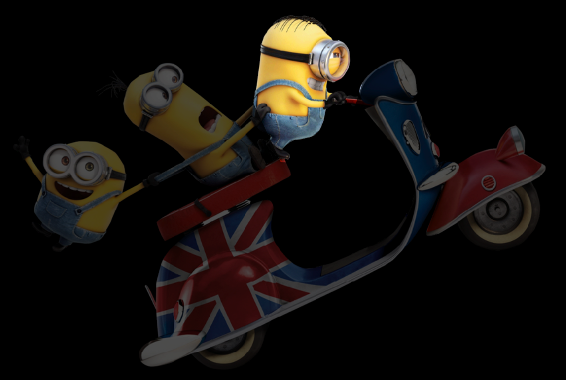
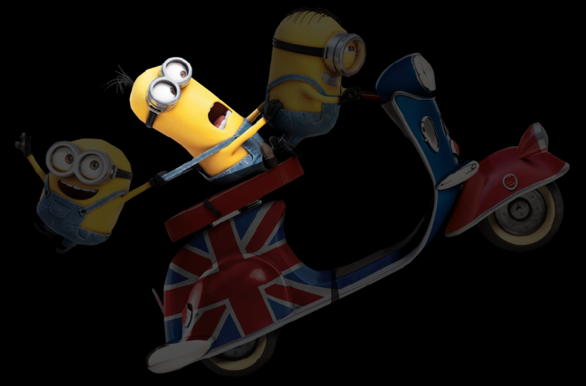
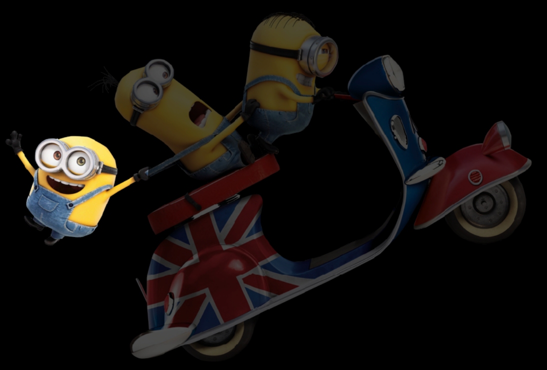
Avg. Regret: 0.40

$\sum l_t$		l_1	l_2	l_3	l_4
2.0	Expert 1 	$\boxed{1.0}$	0.5	0.5	1.0
$\boxed{1.7}$	$\boxed{\text{Expert 2}}$ 	0.2	$\boxed{0.5}$	$\boxed{1.0}$	0.2
0.9	Expert 3 	0.5	0.2	0.2	$\boxed{0.5}$

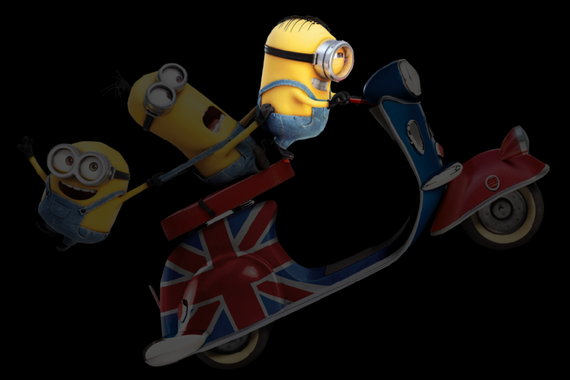
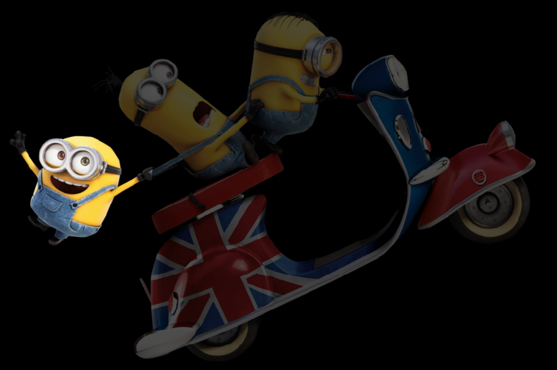
Avg. Regret: 0.53

$\sum l_t$		l_1	l_2	l_3	l_4	l_5
3.0	Expert 1 	$\boxed{1.0}$	0.5	0.5	1.0	0.5
1.9	Expert 2 	0.2	$\boxed{0.5}$	$\boxed{1.0}$	0.2	1.0
$\boxed{1.4}$	$\boxed{\text{Expert 3}}$ 	0.5	0.2	0.2	$\boxed{0.5}$	$\boxed{0.2}$

Avg. Regret: **0.40**

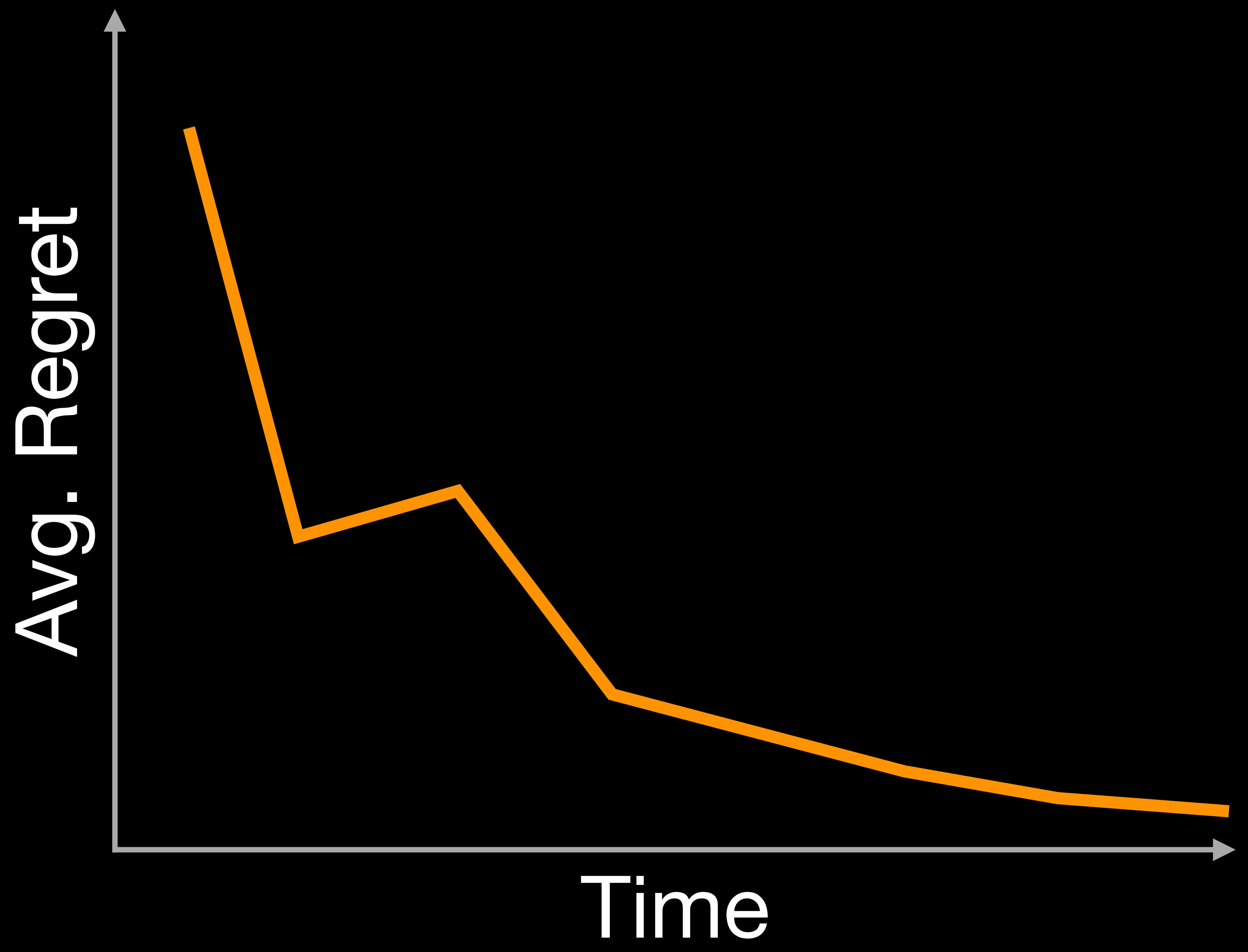
$\sum l_t$		l_1	l_2	l_3	l_4	l_5	l_6
3.5	Expert 1 	$\boxed{1.0}$	0.5	0.5	1.0	0.5	1.0
2.9	Expert 2 	0.2	$\boxed{0.5}$	$\boxed{1.0}$	0.2	1.0	0.5
1.6	Expert 3 	0.5	0.2	0.2	$\boxed{0.5}$	$\boxed{0.2}$	$\boxed{0.2}$

Avg. Regret: 0.32

$\sum l_t$		l_1	l_2	l_3	l_4	l_5	l_6
4.5	Expert 1 	$\boxed{1.0}$	0.5	0.5	1.0	0.5	1.0
3.4	Expert 2 	0.2	$\boxed{0.5}$	$\boxed{1.0}$	0.2	1.0	0.5
1.8	Expert 3 	0.5	0.2	0.2	$\boxed{0.5}$	$\boxed{0.2}$	$\boxed{0.2}$

Avg. Regret: **0.26**

FTL appears to be no regret …

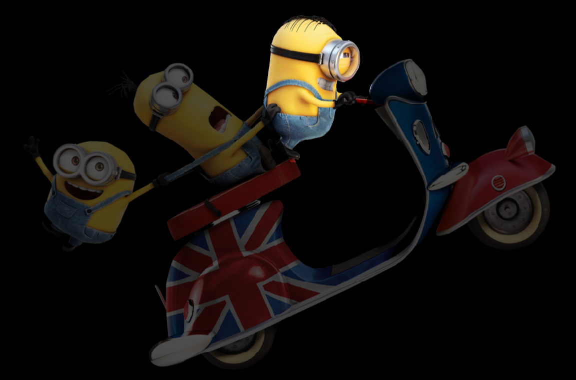


Let's prove it!

Can you make FTL have high regret?

$\sum l_t$

Expert 1

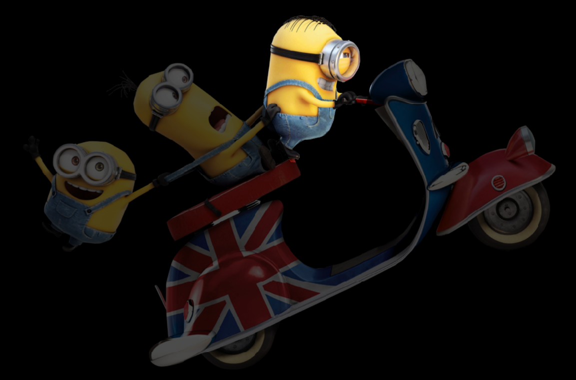
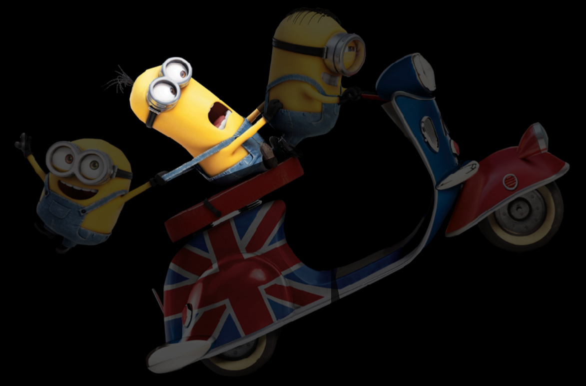
-- 

Expert 2

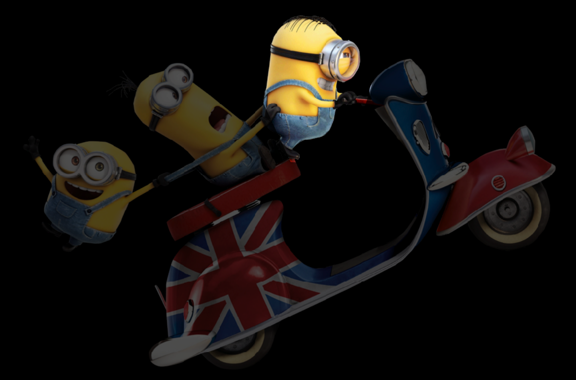
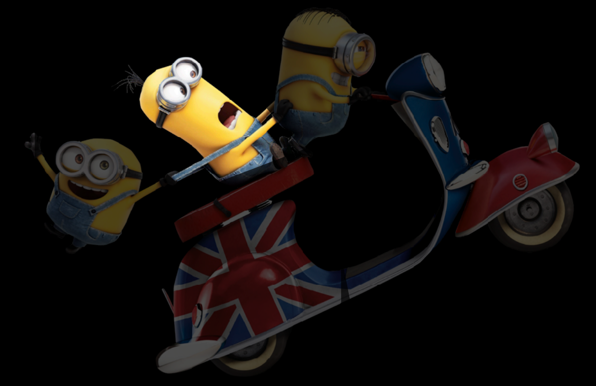
-- 

Avg. Regret: --

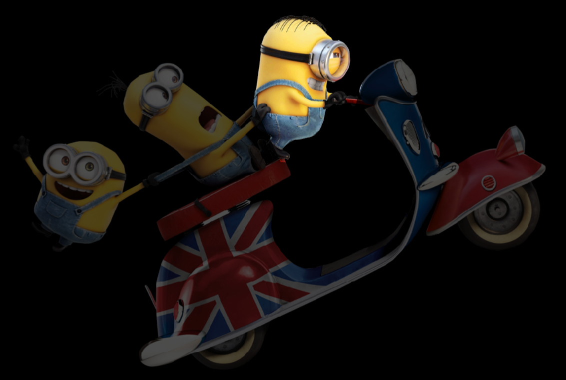
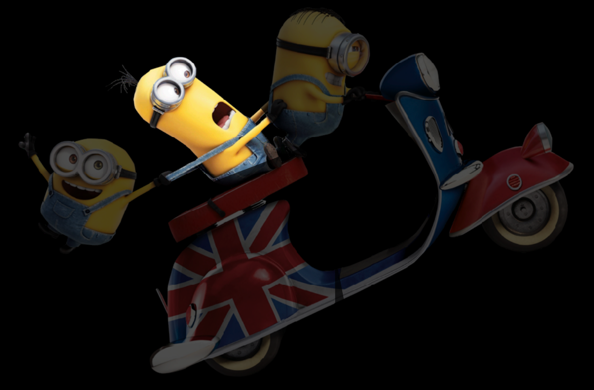
$\sum l_t$ l_1

- -	Expert 1 	$\boxed{1.0}$
- -	Expert 2 	0.0

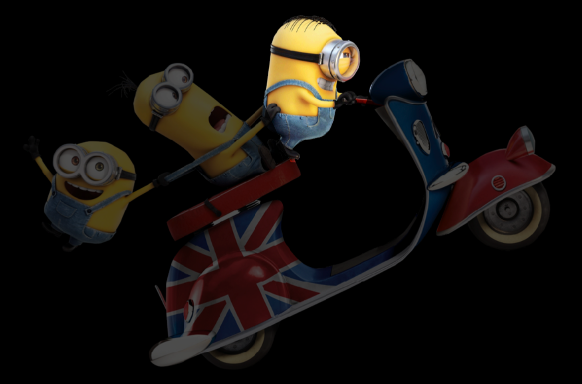
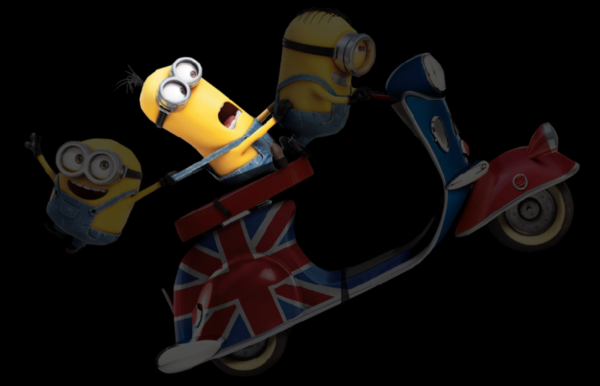
Avg. Regret: - -

$\sum l_t$		l_1	l_2
1.0	Expert 1 	$\boxed{1.0}$	0.0
0.0	Expert 2 	0.0	$\boxed{1.0}$

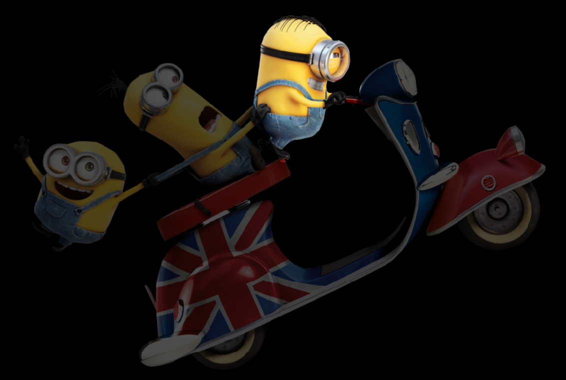
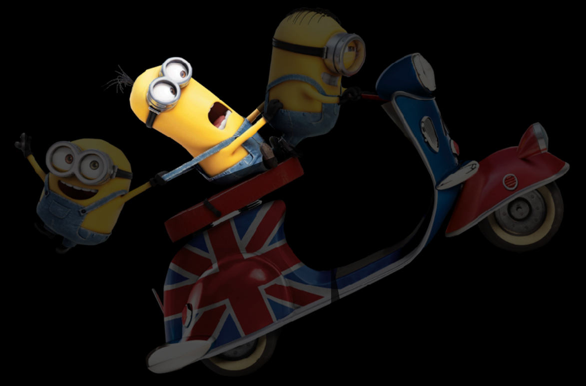
Avg. Regret: **1.00**

$\sum l_t$		l_1	l_2	l_3
1.0	Expert 1 	$\boxed{1.0}$	0.0	$\boxed{1.0}$
1.0	Expert 2 	0.0	$\boxed{1.0}$	0.0

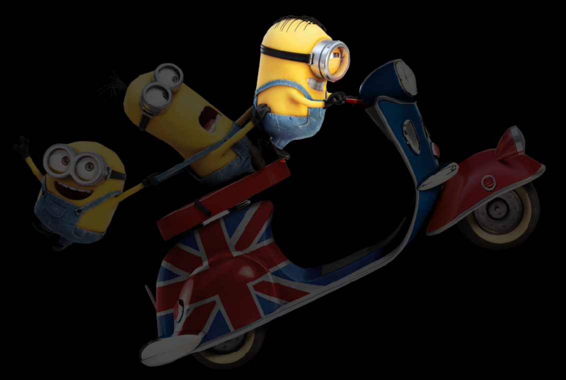
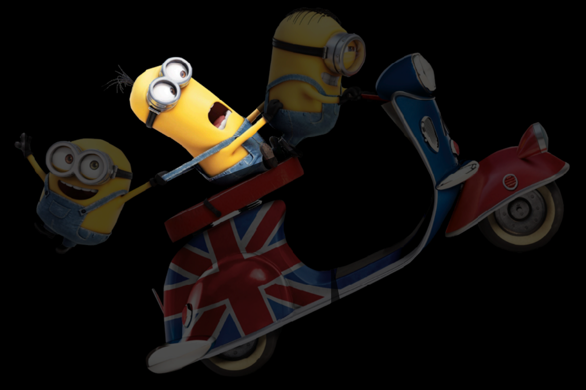
Avg. Regret: 0.50

$\sum l_t$		l_1	l_2	l_3	l_4
2.0	Expert 1 	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0
1.0	Expert 2 	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$

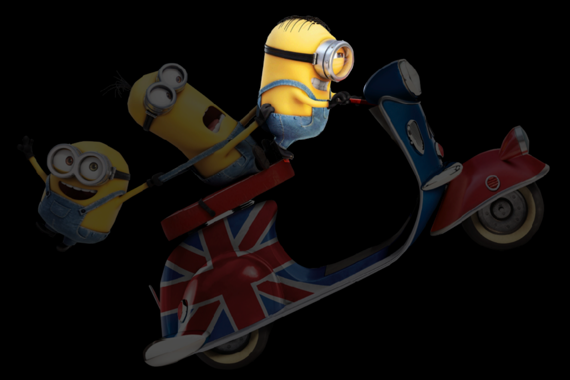
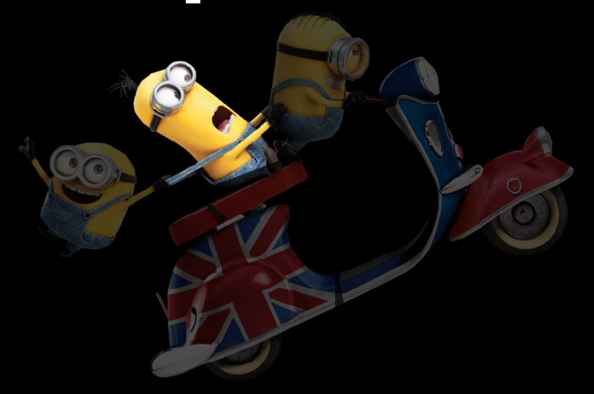
Avg. Regret: 0.67

$\sum l_t$		l_1	l_2	l_3	l_4	l_5
2.0	Expert 1 	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$
2.0	Expert 2 	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0

Avg. Regret: 0.50

$\sum l_t$		l_1	l_2	l_3	l_4	l_5	l_6
3.0	Expert 1 	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0
2.0	Expert 2 	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$

Avg. Regret: **0.60**

$\sum l_t$		l_1	l_2	l_3	l_4	l_5	l_6
3.0	Expert 1 	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0
3.0	Expert 2 	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$	0.0	$\boxed{1.0}$

Predictions not stable → High regret!

Avg. Regret: **0.50**

Cover's Impossibility Result

"A powerful enough adversary
can drive the Regret of
any deterministic online algorithm
to $O(T)$
by anticipating its prediction
and setting maximal loss"

How can we curb the power of the adversary?

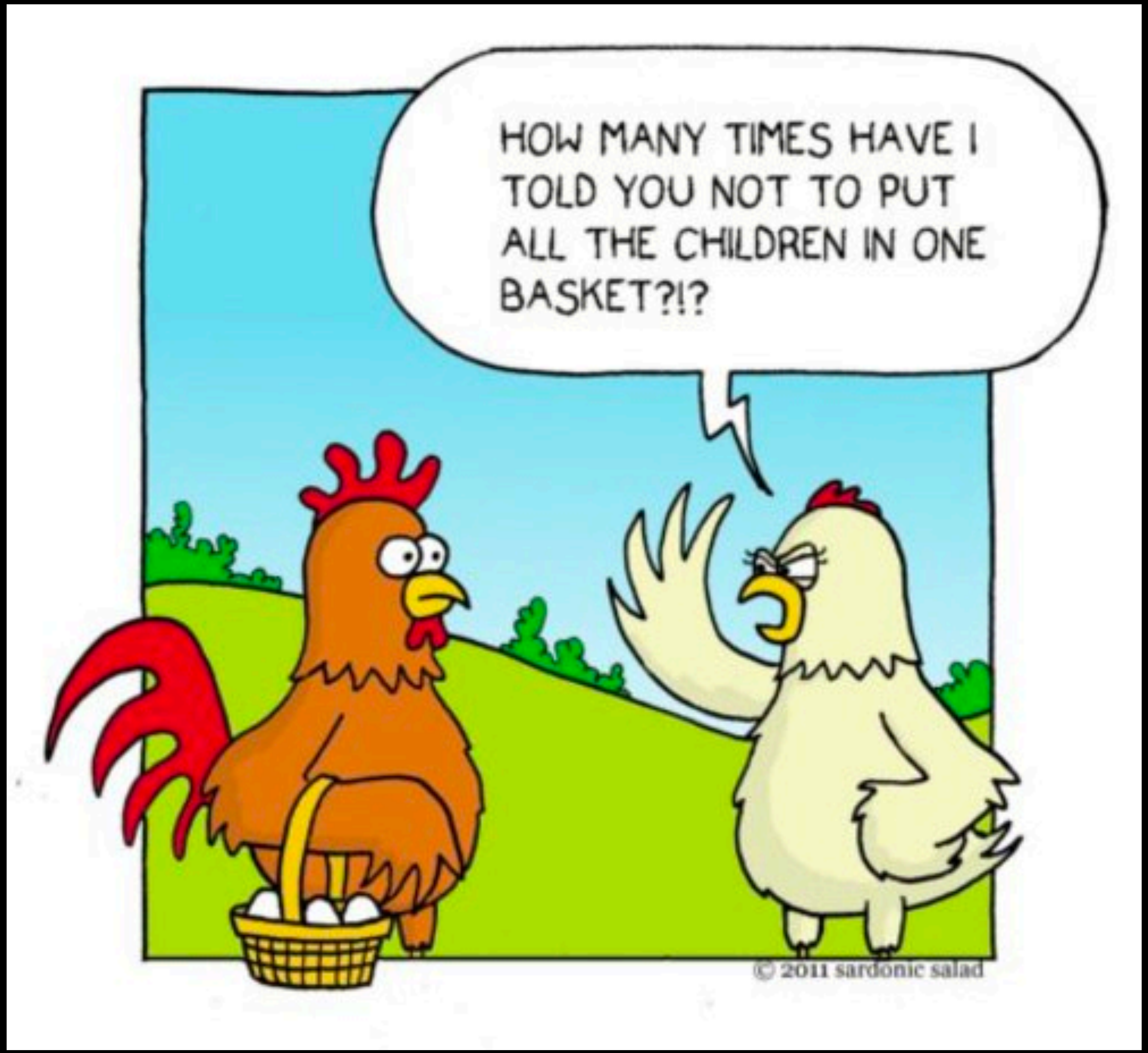
π_t

$l_t(\pi_t)$

$$\pi_t = \arg\min_{\pi} \sum_{i=1}^{t-1} l_i(\pi)$$

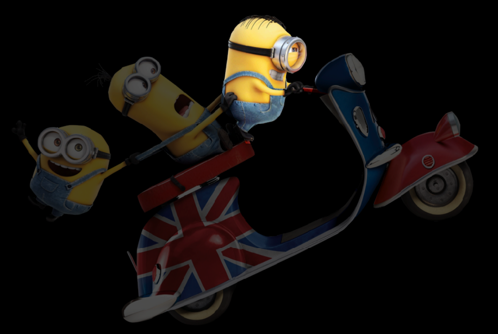
Adversary breaks *any* determinism

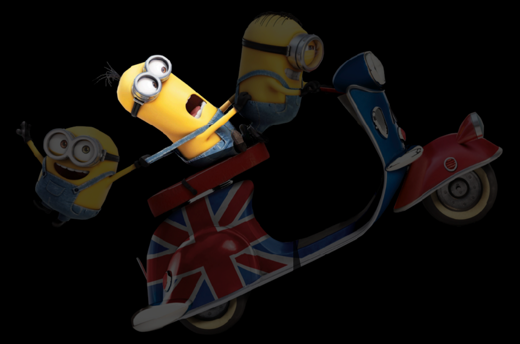
The virtue of hedging

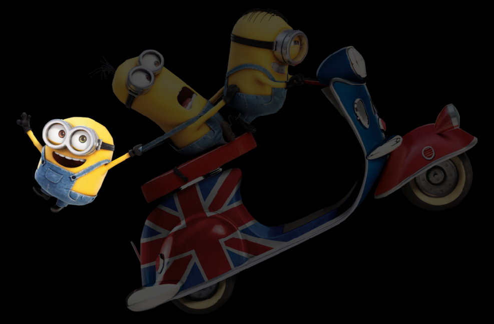


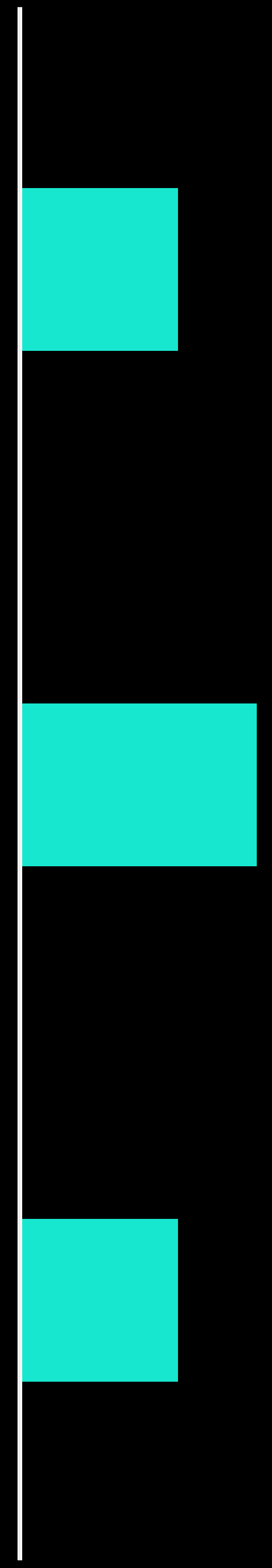
Choose probability over experts

$$p = \frac{w^i}{\sum_i w^i}$$

Expert 1  $w^1 = 1.0$

Expert 2  $w^2 = 2.0$

Expert 3  $w^3 = 1.0$



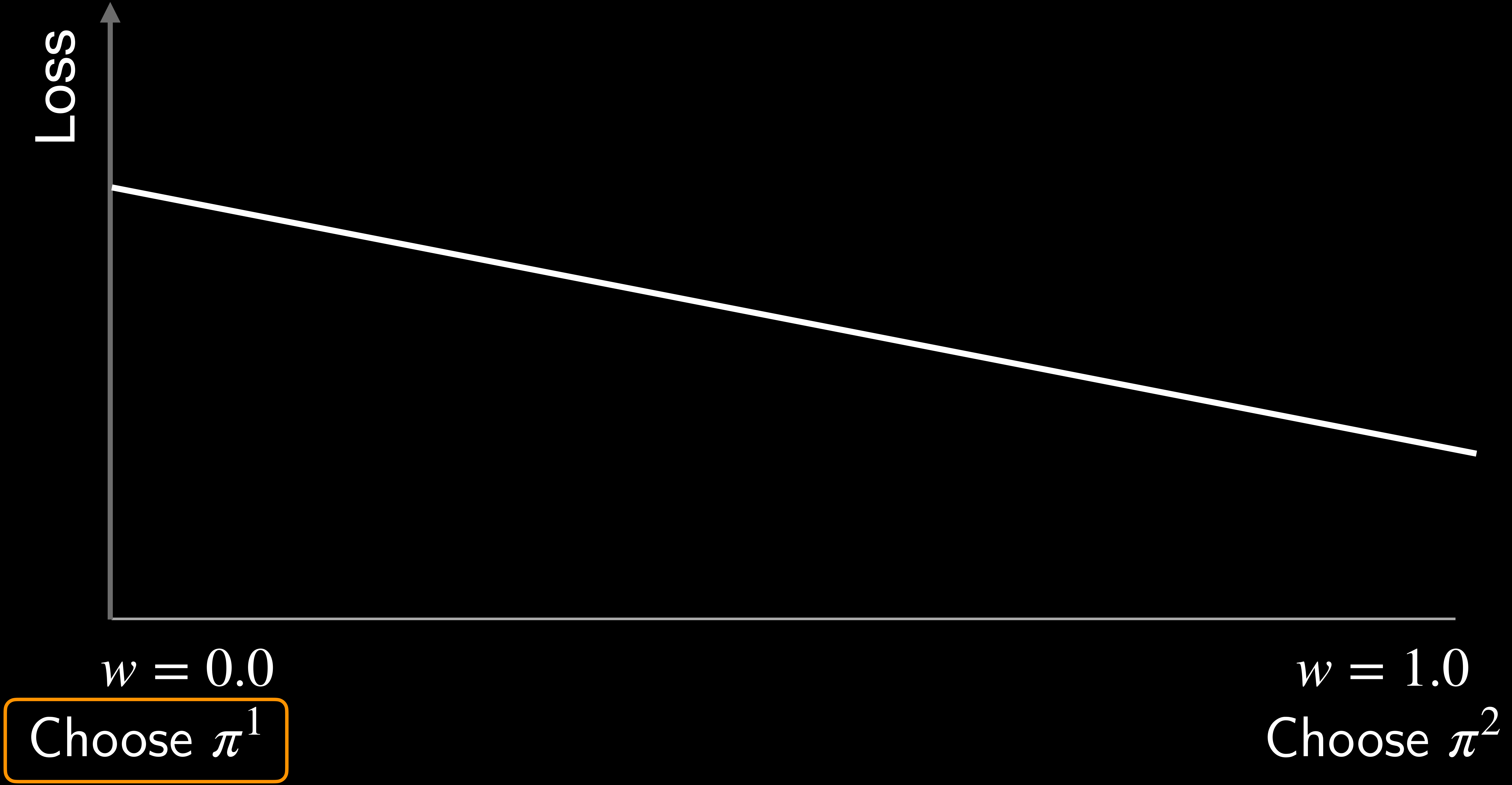
Let's formalize!

Let's apply FTL again (but on the space of weights)

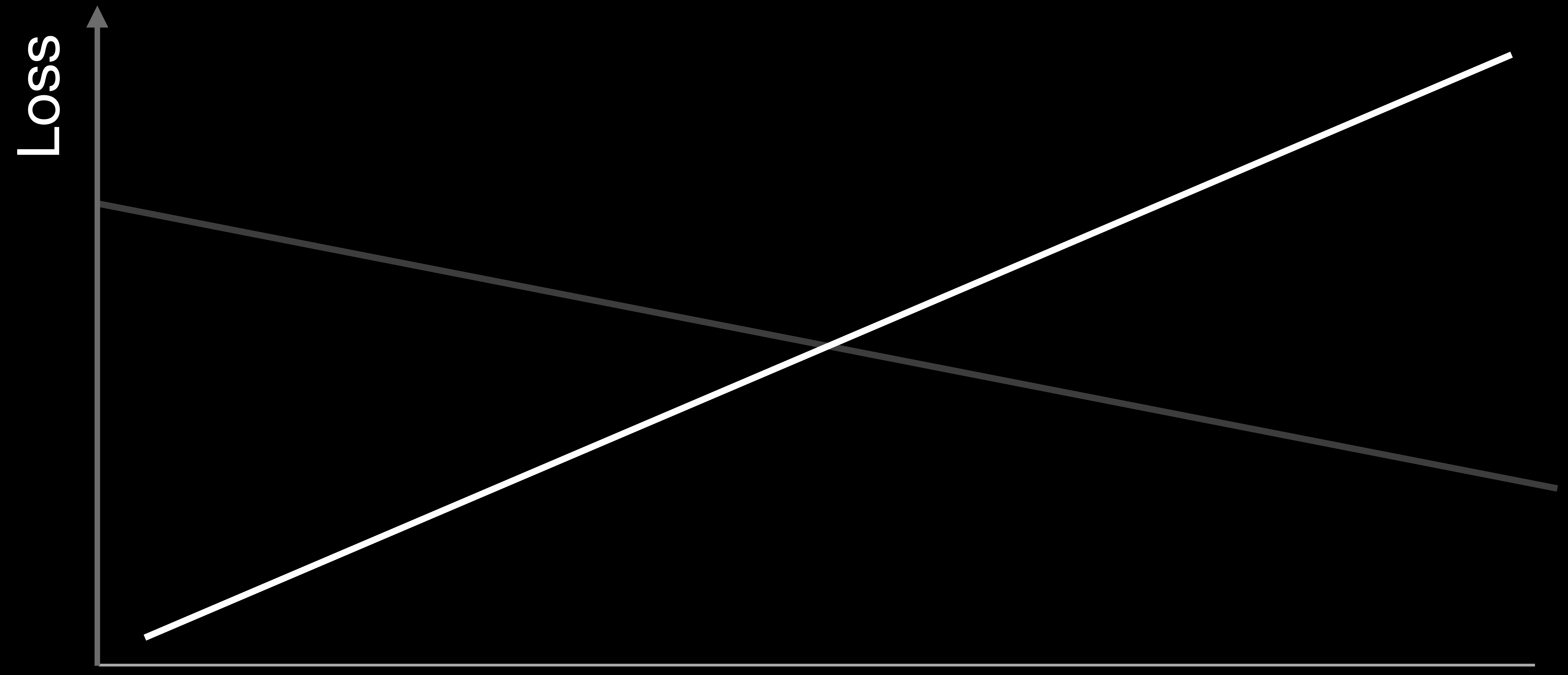
At every round t, choose the best weights in hindsight

$$w_t = \arg\min_{w} \sum_{i=1}^{t-1} l_i(w)$$

Loss = 0.75 Avg. Regret = 0.5



Loss = 1.0 Avg. Regret = 0.5



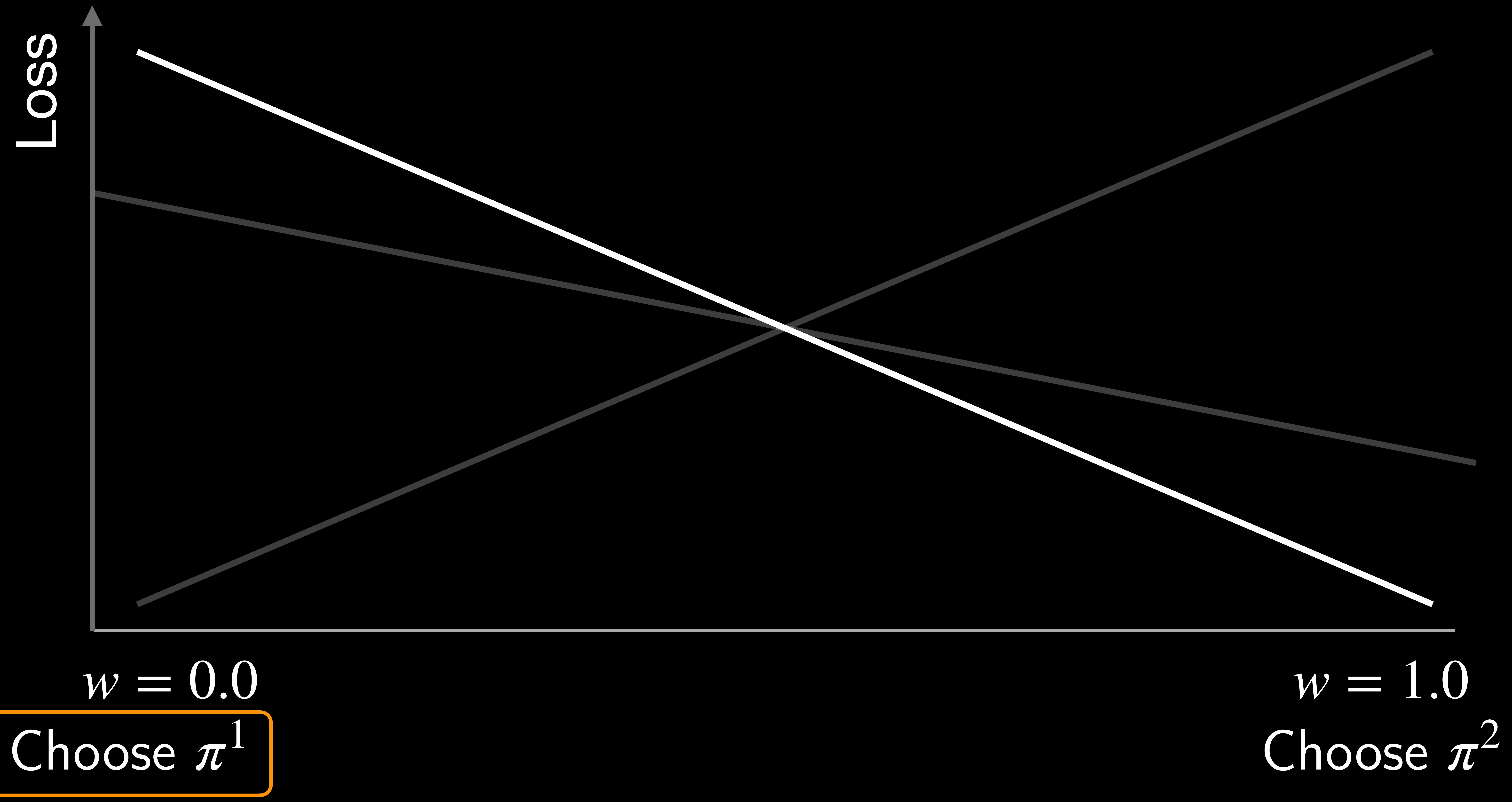
$w = 0.0$

Choose π^1

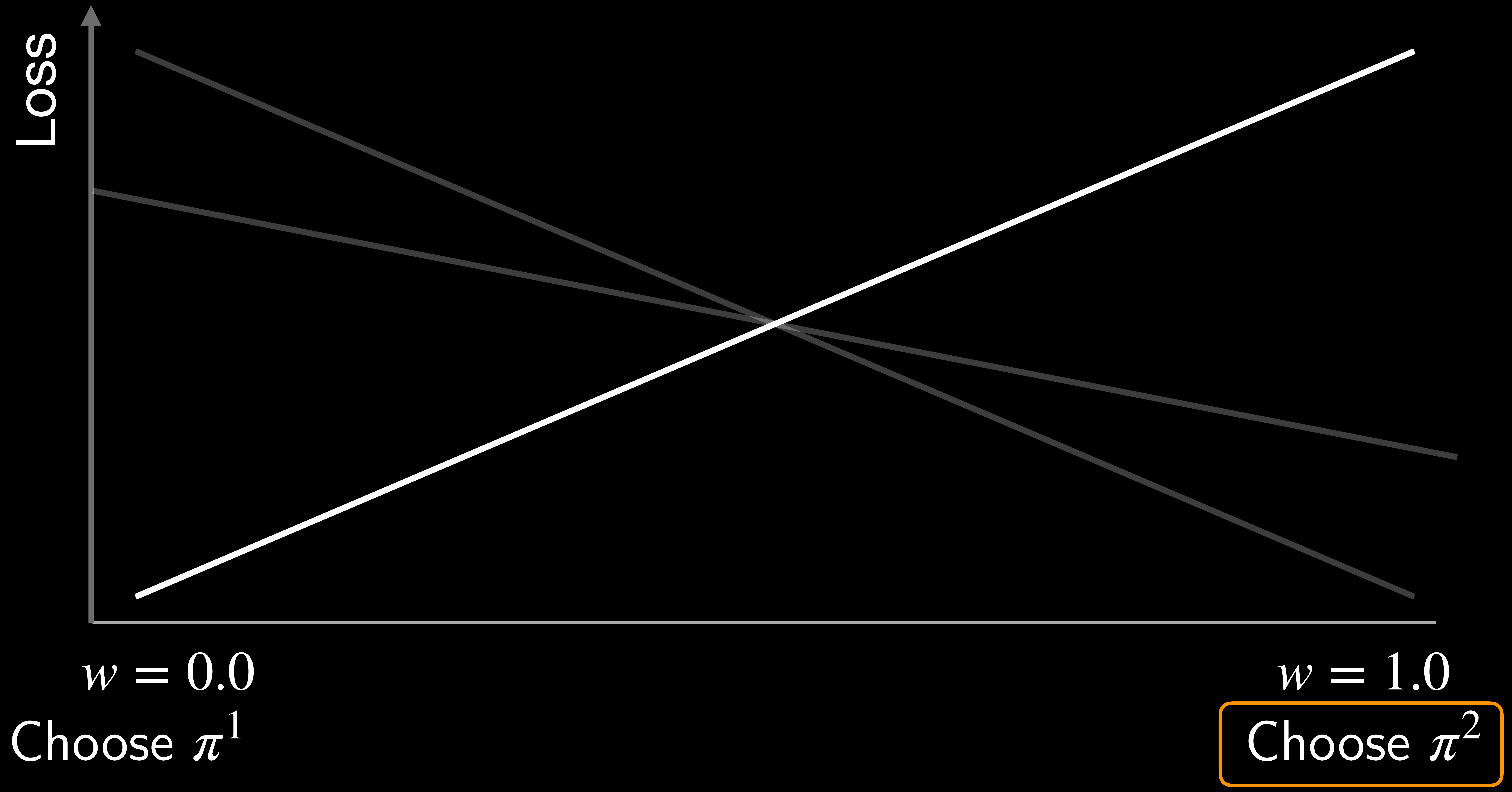
$w = 1.0$

$\boxed{\text{Choose } \pi^2}$

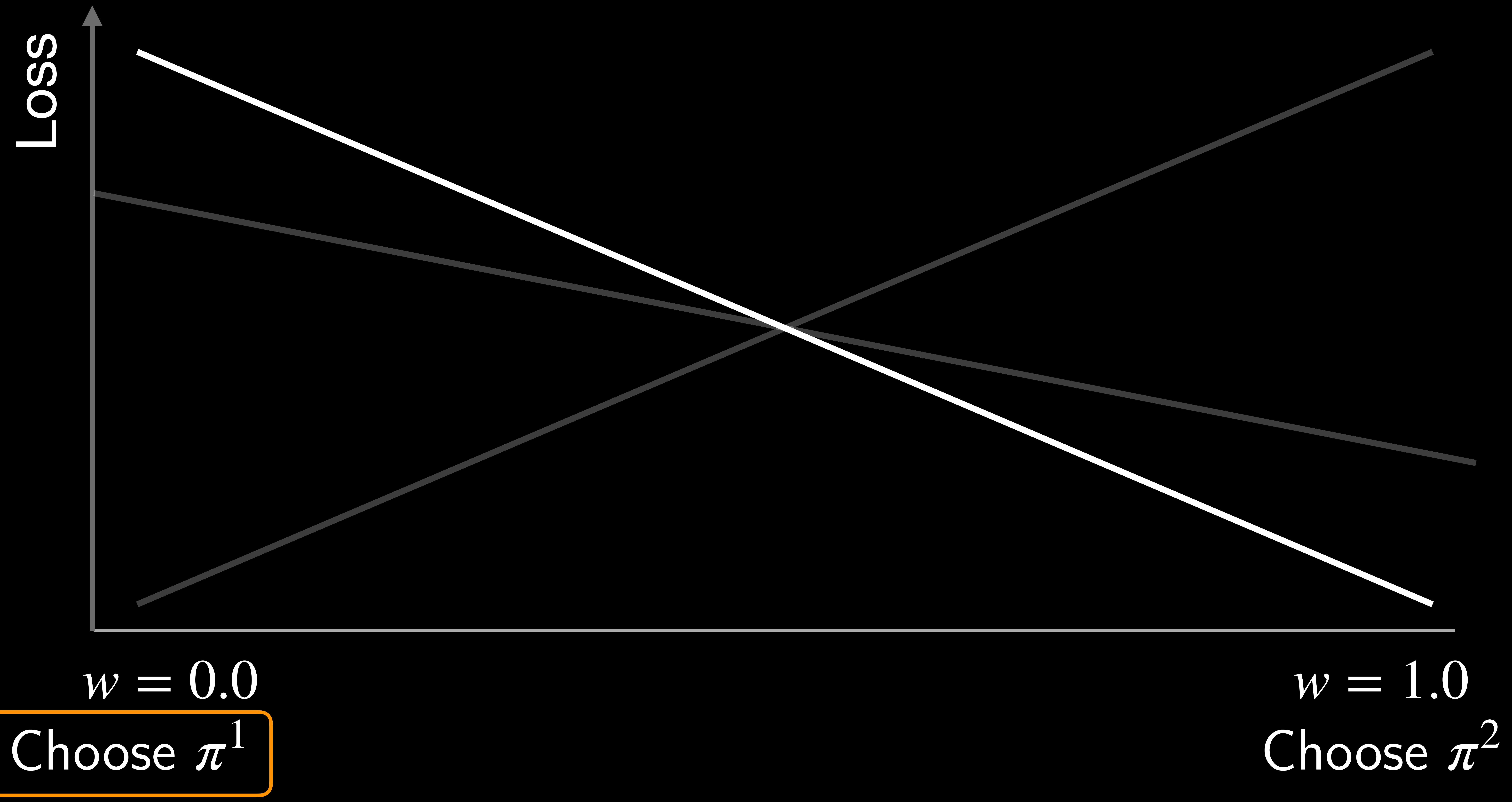
Loss = 1.0 Avg. Regret = 0.5



Loss = 1.0 Avg. Regret = 0.5



Loss = 1.0 Avg. Regret = 0.5



Follow the leader is too aggressive ...

Both in discrete and continuous settings!

Stability is the key problem!

$$w = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}$$

w_t

$l_t(w_t)$

$$w_t = \arg\min_w \sum_{i=1}^{t-1} l_i(w)$$

Unstable predictions!

Be stable

Slowly change predictions

Follow the Regularized Leader

$$w_t = \arg\min_w \sum_{i=1}^{t-1} l_i(w) + \eta_t R(w)$$

Strong regularization!

What are some choices for regularization?

Generalized Weighted Majority

Episode IV

A NEW HOPE

GENERALIZED WEIGHTED MAJORITY

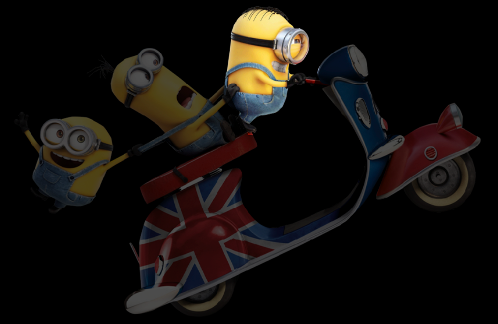
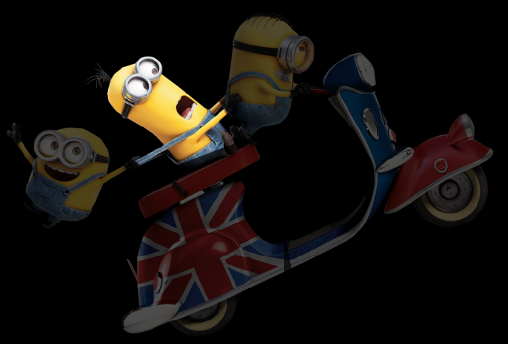
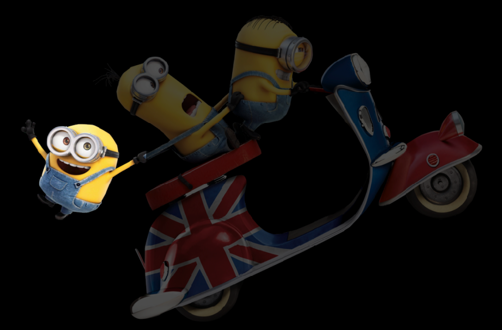
1. At t=1, set weight for expert i as $w_1^i = 1$

2. At time t, choose expert i with probability $\frac{w_t^i}{\sum_i w_t^i}$

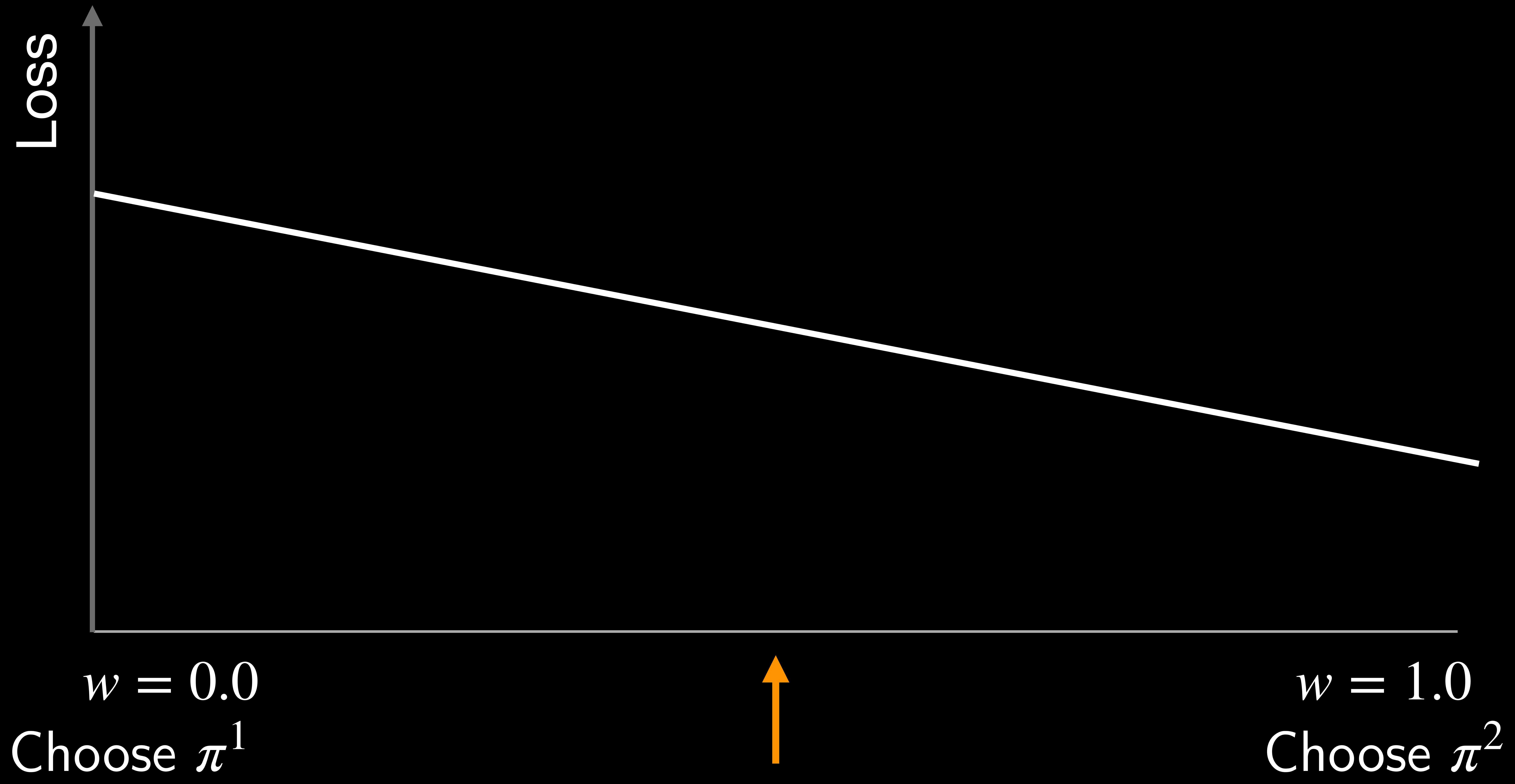
3. Update weight for expert i (Bump down if loss is high)

$$w_{t+1}^i = w_t^i \exp(-\eta l_t(\pi^i))$$

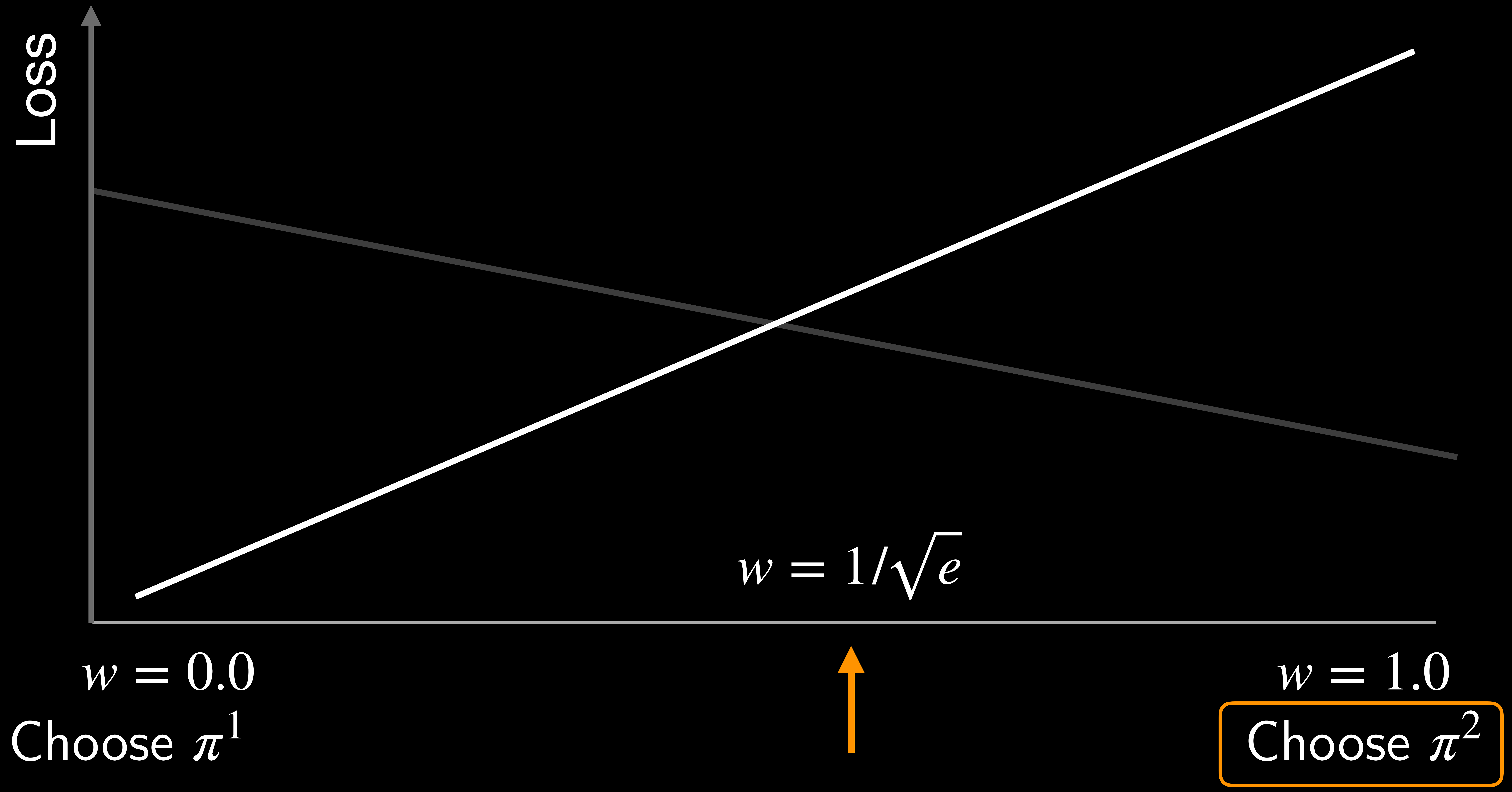
Generalized Weighted Majority

	$\frac{w_1^i}{\sum_i w_1^i}$	l_1	$\frac{w_2^i}{\sum_i w_2^i}$	l_2	$\frac{w_3^i}{\sum_i w_3^i}$	l_3	$\frac{w_4^i}{\sum_i w_4^i}$	l_4
		$\boxed{1.0}$		0.5		0.1		1.0
		0.2		0.1		$\boxed{0.5}$		$\boxed{0.1}$
		0.5		$\boxed{1.0}$		0.1		1.0

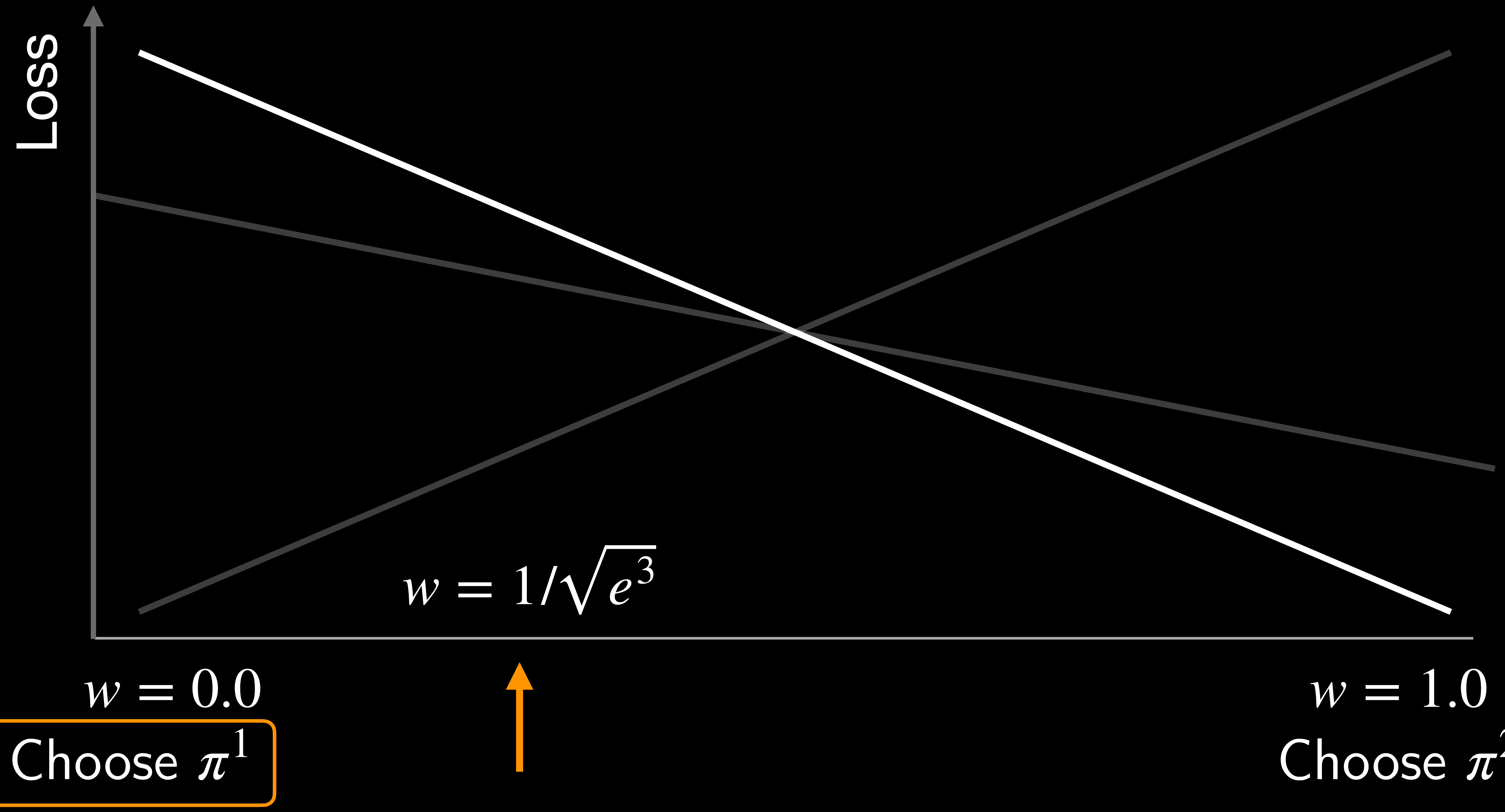
Loss = 0.5 Avg. Regret = 0.25



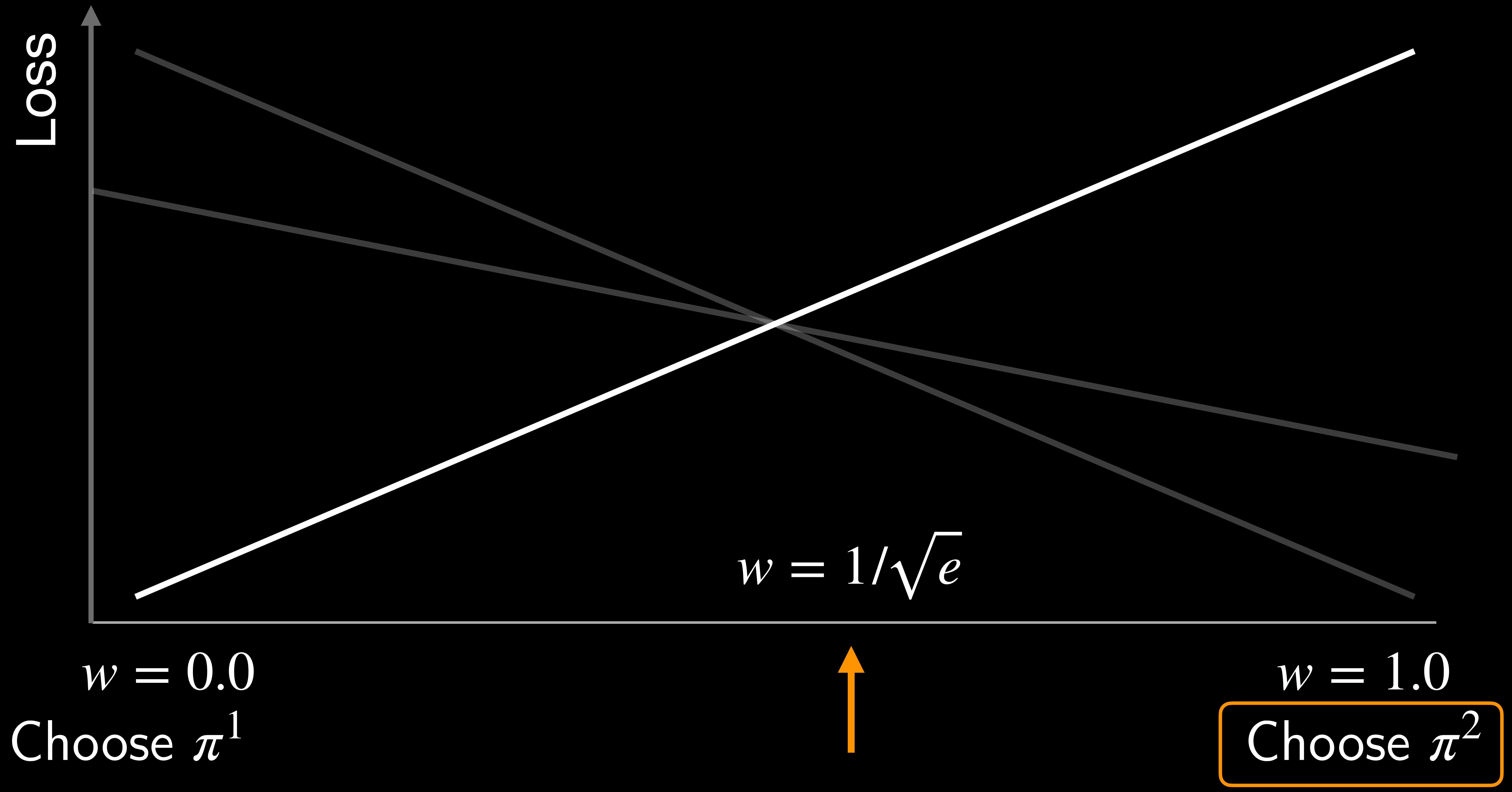
Loss = 0.6 Avg. Regret = 0.17



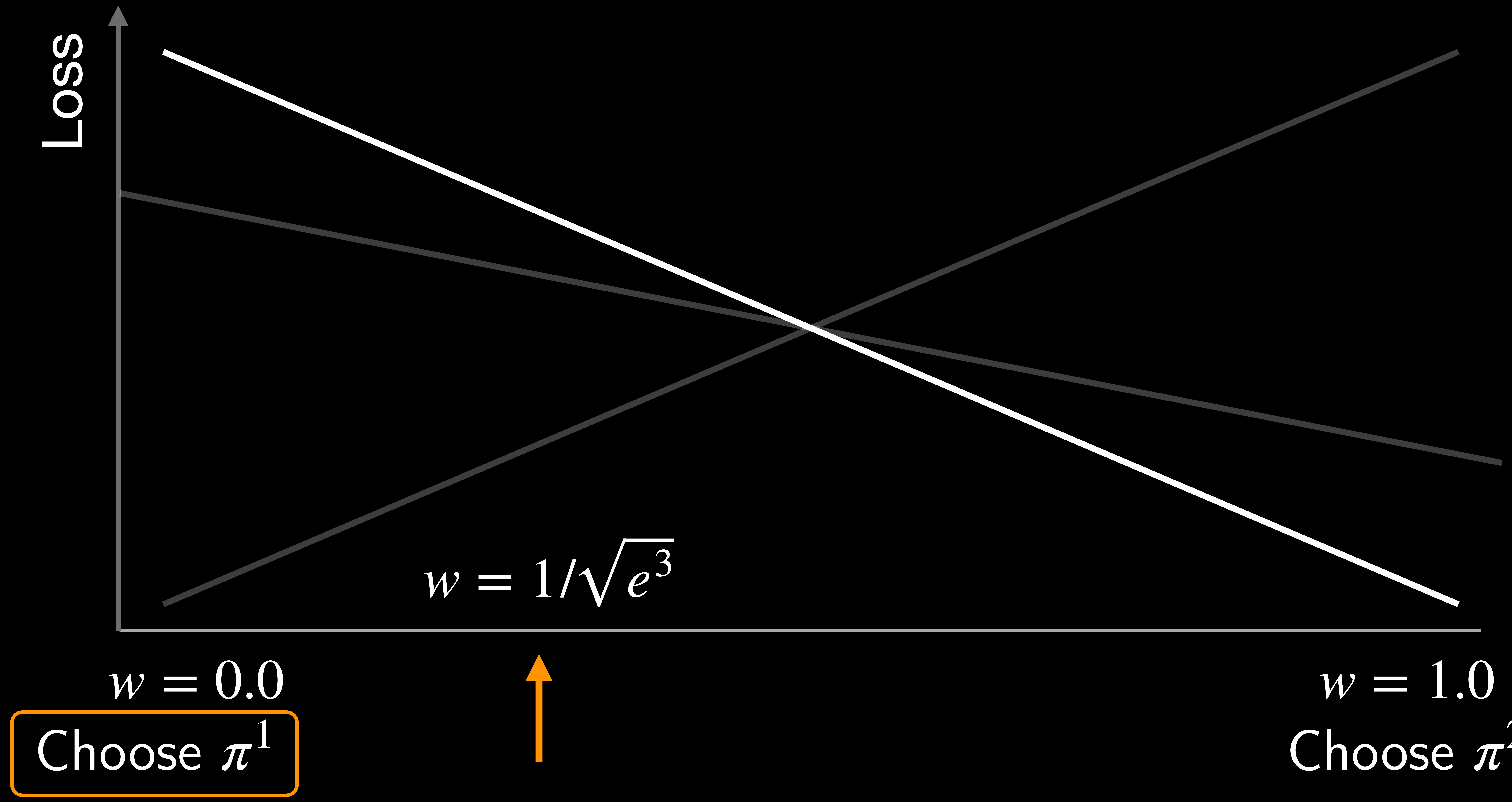
Loss = 0.78 Avg. Regret = 0.21



Loss = 0.6 Avg. Regret = 0.18



Loss = 0.78 Avg. Regret = 0.2




Linear Programming
Boosting
Games
Soft-RL

Three Challenges

C1: Derive GWM from
Follow the Regularized Leader

C2: Show that GWM is No-Regret

C3: Show that FTRL is No-Regret

(Share on Ed!)

$w = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}$

w_t

$l_t(w_t)$

$$w_t = \arg\min_w \sum_{i=1}^{t-1} l_i(w)$$

Unstable predictions!

Regularization

⇒No Regret!

$$w_t = \arg\min_w \sum_{i=1}^{t-1} l_i(w) + \eta_t R(w)$$