COM 3240

REINFORCEMENT LEARNING

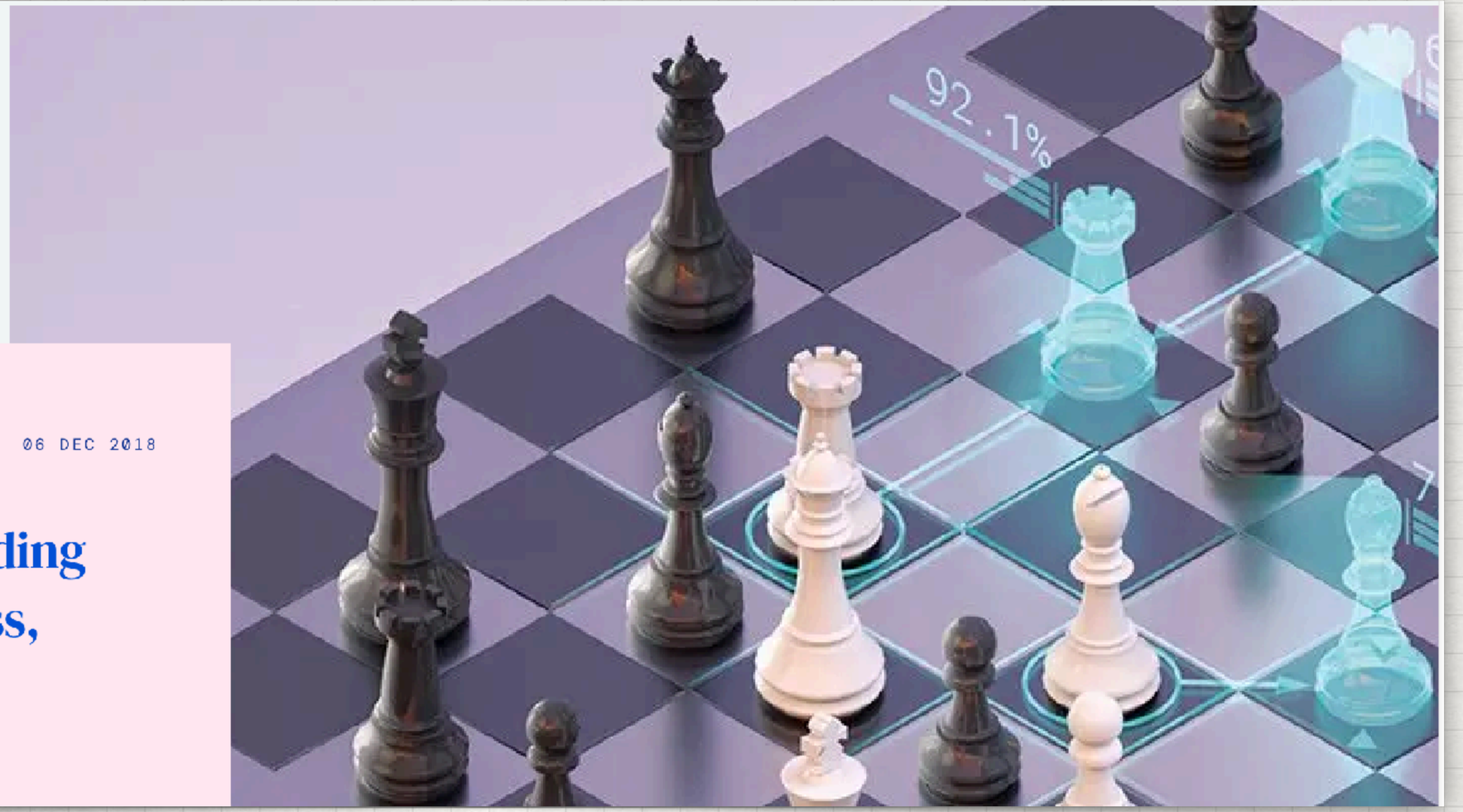
PHILOSOPHY OF THE CLASS

A POSTERIORI

- First principles, building upon A-level maths/further maths.
- Revision/introduction of all necessary knowledge.
- (Almost) every result is grounded.
- One key technique across (almost) every topic.
- It is not about memorising, but understanding.

ALPHA ZERO

BY GOOGLE DEEPMIND



BLOG POST
RESEARCH

06 DEC 2018

AlphaZero: Shedding new light on chess, shogi, and Go

https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go


DALL·E

OPTIMISATION

GRADIENT DESCENT

Desirable: A system that performs a specific task

The system has parameters that we need to select appropriately (optimise) for that task

$$f(x_1, x_2, \ldots, x_n)$$

parameters

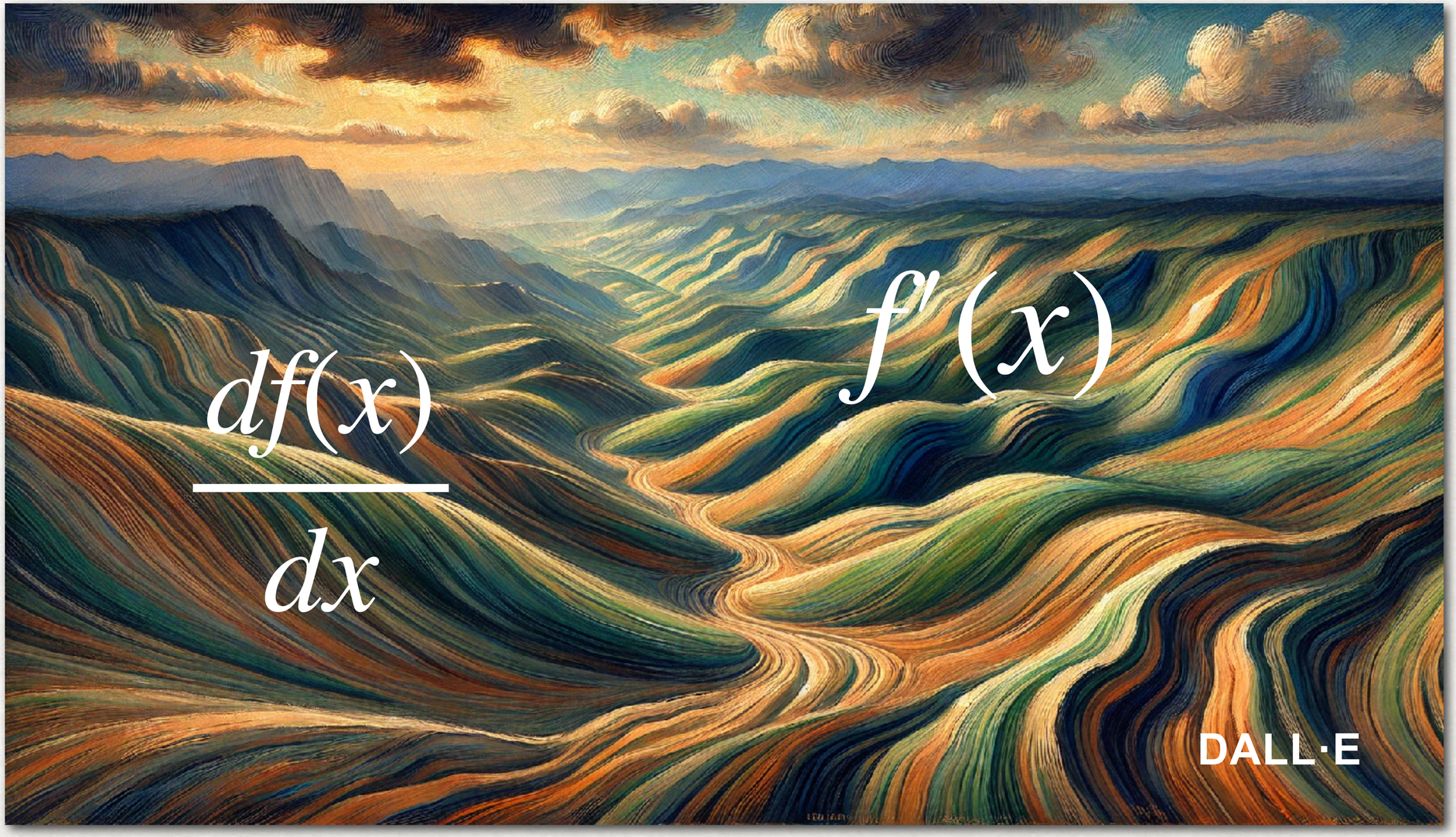
f

System's performance: maximise

System's error : minimise

DERIVATIVES

OPTIMISATION - GRADIENT METHODS



MOTIVATION FOR PARTIAL DERIVATIVES

LIGHTS AND BUTTONS PUZZLE

PROBABILITIES

PROPERTIES OF CONDITIONAL EXPECTATIONS

$$E_X[X \mid Y = y] = \sum_x x \cdot P(X = x \mid Y = y)$$

$$E[X] = E_Y[E_X[X \mid Y]]$$ **Law of Total Expectation**

$$E[aX + bY \mid Z] = aE[X \mid Z] + bE[Y \mid Z]$$ **Linearity**

$$E[X \mid Y] = E[X]$$ **Independence**

DISCOUNT FACTOR

IN REINFORCEMENT LEARNING

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots$$

$$0 \leq \gamma < 1$$

This is why I am impatient

FUTURE REWARDS

TOTAL RETURN

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

$$G_t = R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \dots \right)$$

$$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$ **Recursive!**

FUTURE REWARDS

ACTION-VALUE FUNCTIONS

$$G_t = R_{t+1} + \gamma G_{t+1}$$

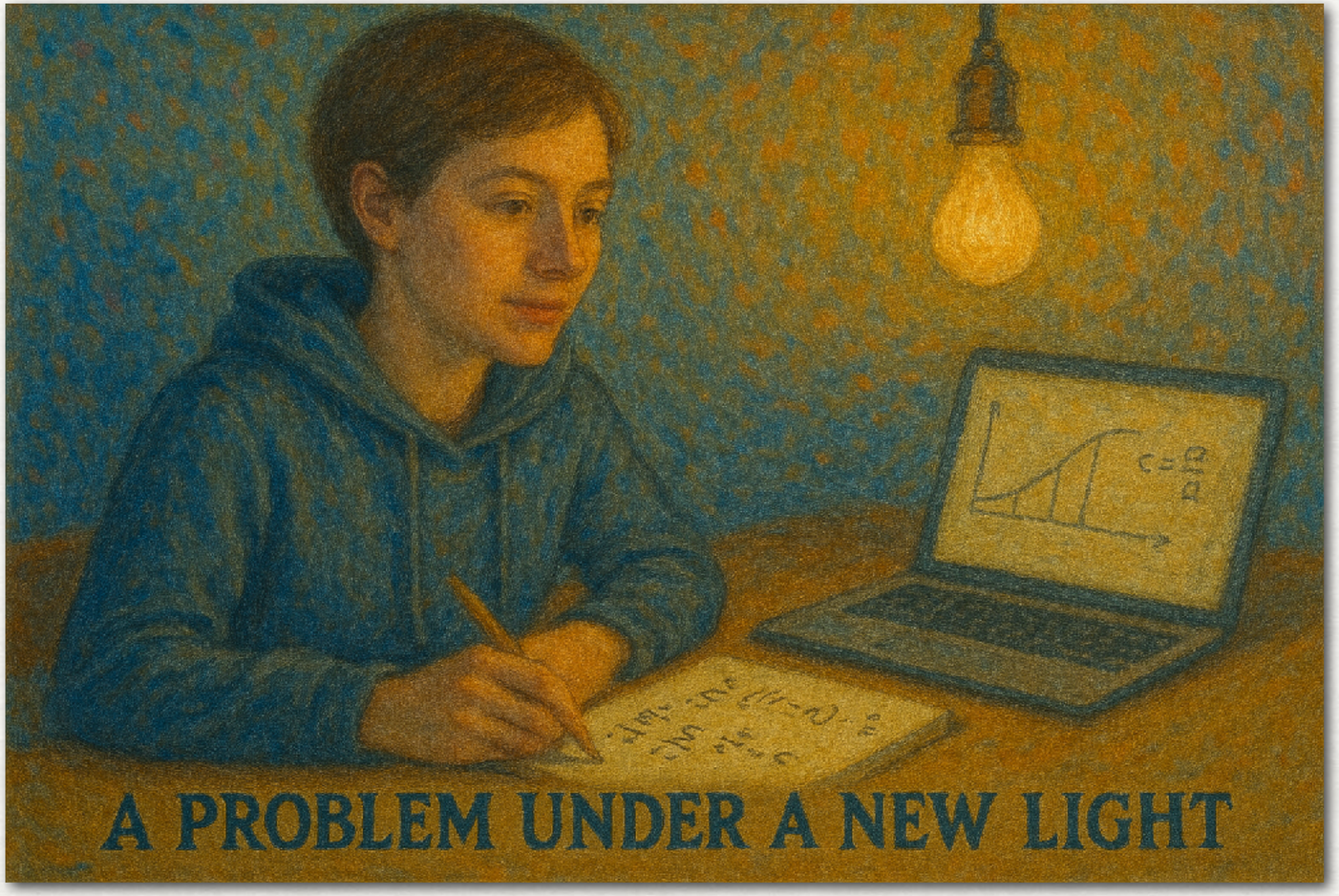
$$q^{\pi}(s, a) = \mathbb{E}_{\pi}[G_t \,|\, S_t = s, A_t = a]$$

$$q^{\pi}(s, a) = \mathbb{E}_{\pi}[R_{t+1} + \gamma \, q^{\pi}(s', a') \,|\, S_t = s, A_t = a]$$

The recursive form of total returns is also followed by the action value function.

RL AS PREDICTING THE CORRECT Q VALUES

A MINIMISATION PROBLEM



FUTURE REWARDS

SARSA

$$q^{\pi}(s,a) = E_{\pi}[r + \gamma\, q^{\pi}(s',a') \,|\, s,a]$$

$$E_{\pi}[r + \gamma\, q^{\pi}(s',a') - q^{\pi}(s,a) \,|\, s,a] = 0$$

$$\mathscr{L}(s,a) = \frac{1}{2N} \sum_{i=1}^{N} \left(Q(s,a) - [r^{(i)} + \gamma Q(s'^{(i)}, a'^{(i)})] \right)^2$$

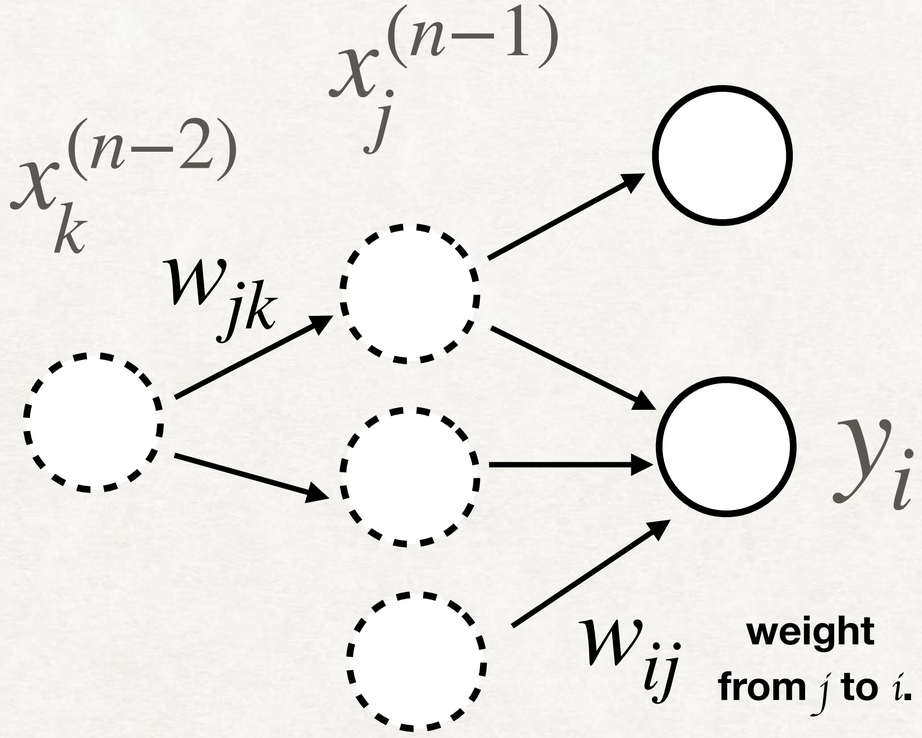
ARTIFICIAL NEURAL NETWORKS AS FUNCTION APPROXIMATIONS

Q-VALUES TABLE

$Q(s_1, a_1)$	$Q(s_2, a_1)$		
$Q(s_1, a_2)$			

ARTIFICIAL NEURAL NETWORKS AS FUNCTION APPROXIMATIONS

HIDDEN LAYER



$$L = \frac{1}{2}\sum_{\mathbf{x}}\sum_{i}\left(y_i^*(\mathbf{x}) - y_i(\mathbf{x};\mathbf{W})\right)^2 = \sum_{\mathbf{x}} L_o$$

$$\Delta w_{ij} = -\eta\frac{\partial L}{\partial w_{ij}} = -\eta\sum_{\mathbf{x}}\frac{\partial L_o}{\partial w_{ij}}$$

$$\Delta w_{jk} = -\eta\frac{\partial L}{\partial w_{jk}} = -\eta\sum_{\mathbf{x}}\frac{\partial L_o}{\partial w_{jk}}$$

Many layers

Input
$\mathbf{x}$

Output

Similarly for biases

Exam info:

Six questions-reply to all.

The guest lecture will not be examined.

The rest of the module is examinable.

Expect questions from the introductory material too!

The majority will be core RL topics.

I am not testing your memory!

Example:

Find the derivative of 7x using the definition of the derivative.

Replying 7 because you remember it is not a valid answer.

PROOF STRATEGY

STARTING POINTS

- You know the target expression you are asked to reach.
- You know the goal in words, but not the final form.

PROOF STRATEGY

WHEN YOU KNOW THE TARGET EXPRESSION

- Write down target statement.
- Decide where to begin your manipulations.
  - Starting from the complex form.
    - Often easier to do because it allows simplification.
  - Starting from the simpler form.
    - Usually requires adding terms, possibly more challenging.

PROOF STRATEGY

THE GENERAL PROOF PROCESS

1. Start with what you know (definitions, identities, known results).
2. Substitute and expand using correct definitions.
3. Simplify or build up expressions as needed:
   - Prefer simplifying complexity if possible.
   - Introduce new terms when necessary.
4. Continue until you reach the desired result.

Let’s see a few examples.

Definition of derivative.

Bayes Rule.

Bellmann Equations.

Deriving update rules from optimisation.

THANK YOU!