Continual Learning: Overcoming catastrophic forgetting in neural networks

Peeyush Singhal | Data Science Guild

Agenda

Overview & Introduction to Continual Learning via E.W.C.

Overview of the paper – "Overcoming catastrophic forgetting in neural networks"

Core Idea and some math

Toy Example

Toy Example / Implementing the paper

Relevance in Mapmaking

Discussion on projects

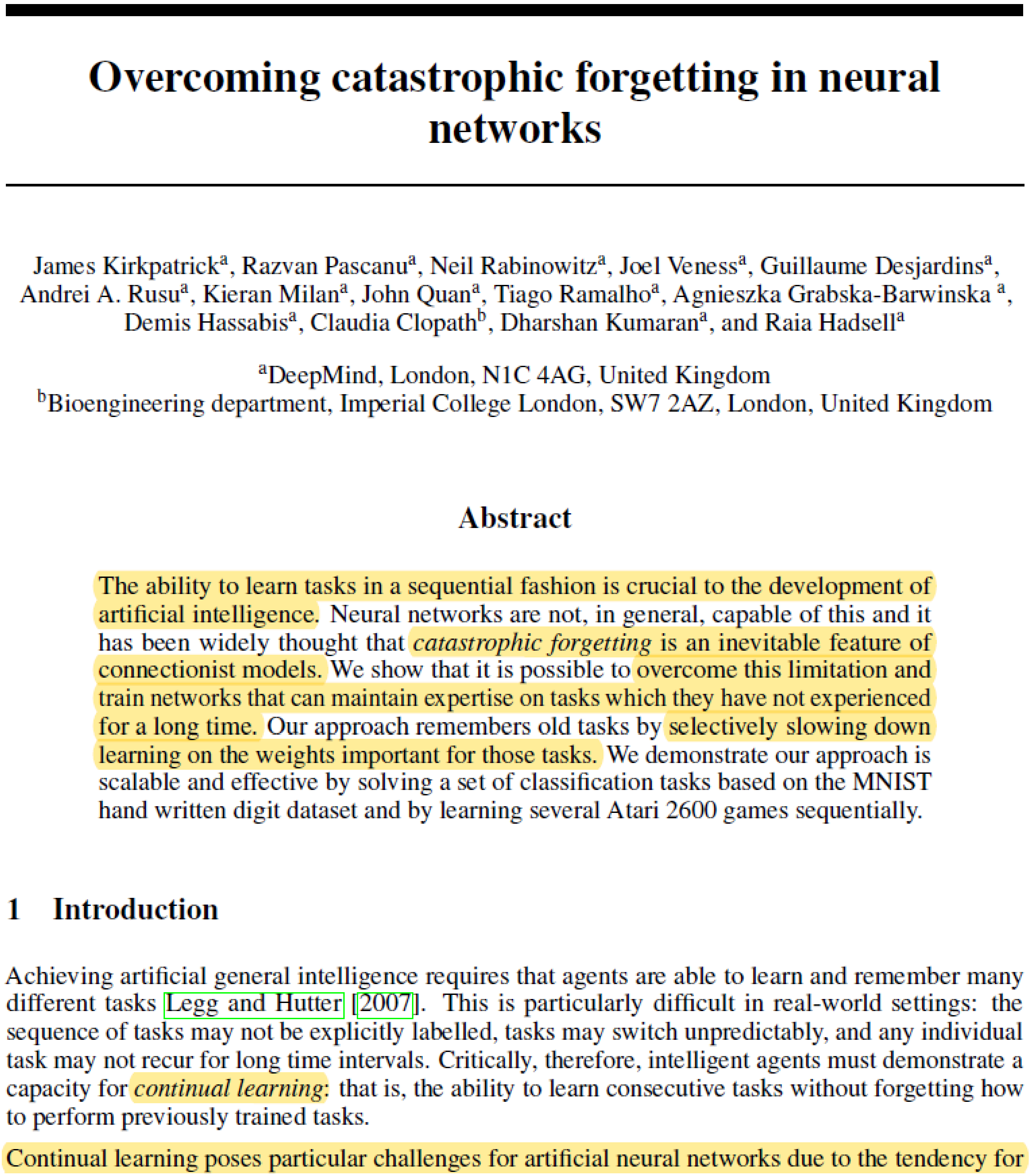
Fisher Information – intuitions (if time permits)

Motivation, Overview of the paper

…

- Have over 4000 citations
- Written by 14 people, mostly DeepMind
- Tutorial in 2022 NeurIPS on "Lifelong Learning Machines"
- One of the authors - Dharshan Kumaran – is a grand master
- One of the authors - Razvan Pascanu – wrote about exploding gradients, with Yoshua Bengio

… in all, the paper has all ingredients for an awesome paper



Overcoming catastrophic forgetting in neural networks

James Kirkpatrick[a], Razvan Pascanu[a], Neil Rabinowitz[a], Joel Veness[a], Guillaume Desjardins[a], Andrei A. Rusu[a], Kieran Milan[a], John Quan[a], Tiago Ramalho[a], Agnieszka Grabska-Barwinska [a], Demis Hassabis[a], Claudia Clopath[b], Dharshan Kumaran[a], and Raia Hadsell[a]

[a]DeepMind, London, N1C 4AG, United Kingdom
[b]Bioengineering department, Imperial College London, SW7 2AZ, London, United Kingdom

Abstract

The ability to learn tasks in a sequential fashion is crucial to the development of artificial intelligence. Neural networks are not, in general, capable of this and it has been widely thought that *catastrophic forgetting* is an inevitable feature of connectionist models. We show that it is possible to overcome this limitation and train networks that can maintain expertise on tasks which they have not experienced for a long time. Our approach remembers old tasks by selectively slowing down learning on the weights important for those tasks. We demonstrate our approach is scalable and effective by solving a set of classification tasks based on the MNIST hand written digit dataset and by learning several Atari 2600 games sequentially.

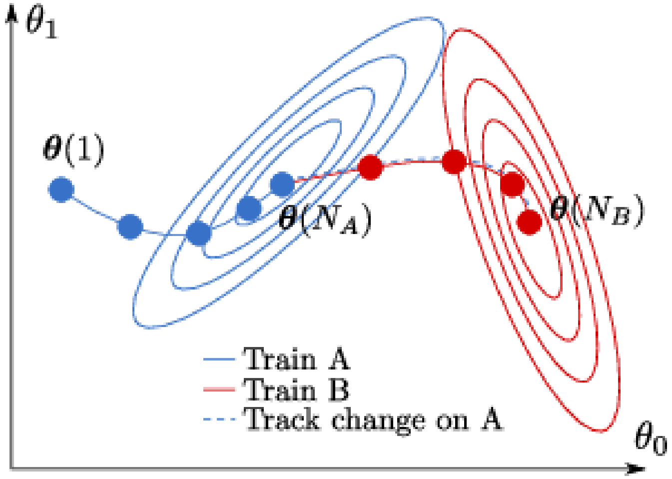
1 Introduction

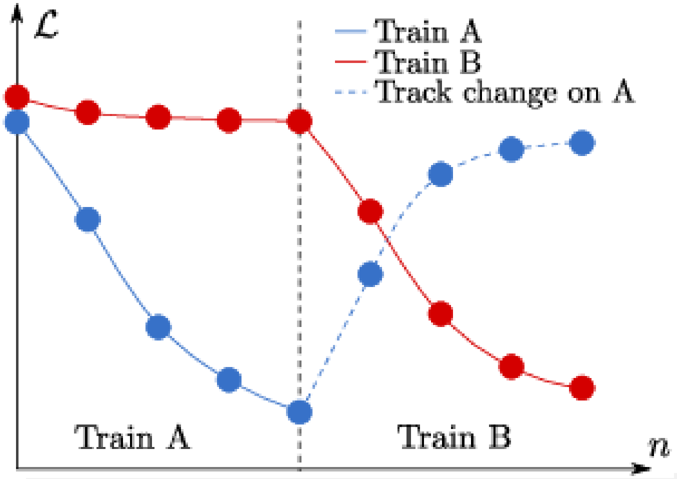
Achieving artificial general intelligence requires that agents are able to learn and remember many different tasks Legg and Hutter [2007]. This is particularly difficult in real-world settings: the sequence of tasks may not be explicitly labelled, tasks may switch unpredictably, and any individual task may not recur for long time intervals. Critically, therefore, intelligent agents must demonstrate a capacity for *continual learning*: that is, the ability to learn consecutive tasks without forgetting how to perform previously trained tasks.

Continual learning poses particular challenges for artificial neural networks due to the tendency for

Catastrophic Forgetting

a.k.a. Catastrophic Interference





- We see that standard backpropagation network can generalize to unseen inputs, but they are very sensitive to new information.
- The main cause of catastrophic interference seems to be overlap in the representations at the hidden layer of distributed neural networks.

How humans / animals deal with Catastrophic Forgetting:

- The mammalian brain may avoid catastrophic forgetting by protecting the previously-acquired knowledge in neocortical circuits *[Cichon and Gan, 2015]* . The dendrites (spine) persists swollen / enlarged despite the subsequent learning of other tasks, accounting for retention of performance *[Yang et al., 2009]*

Source : https://www.youtube.com/watch?v=RIbe-qW9nqY , Wikipedia, Image : Localizing Catastrophic Forgetting in Neural Networks (Wiewel & Yang, 2019)

Continual Learning v Catastrophic forgetting

$D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_n$: Sequence of data shown to the model

- **Continual Learning:**

$$p(y_n \mid x, D_1 \rightarrow \cdots \rightarrow D_n)$$

- **Catastrophic Forgetting:**

$$p(y_1 \mid x, D_1 \rightarrow \cdots \rightarrow D_n), \text{or mostly}$$

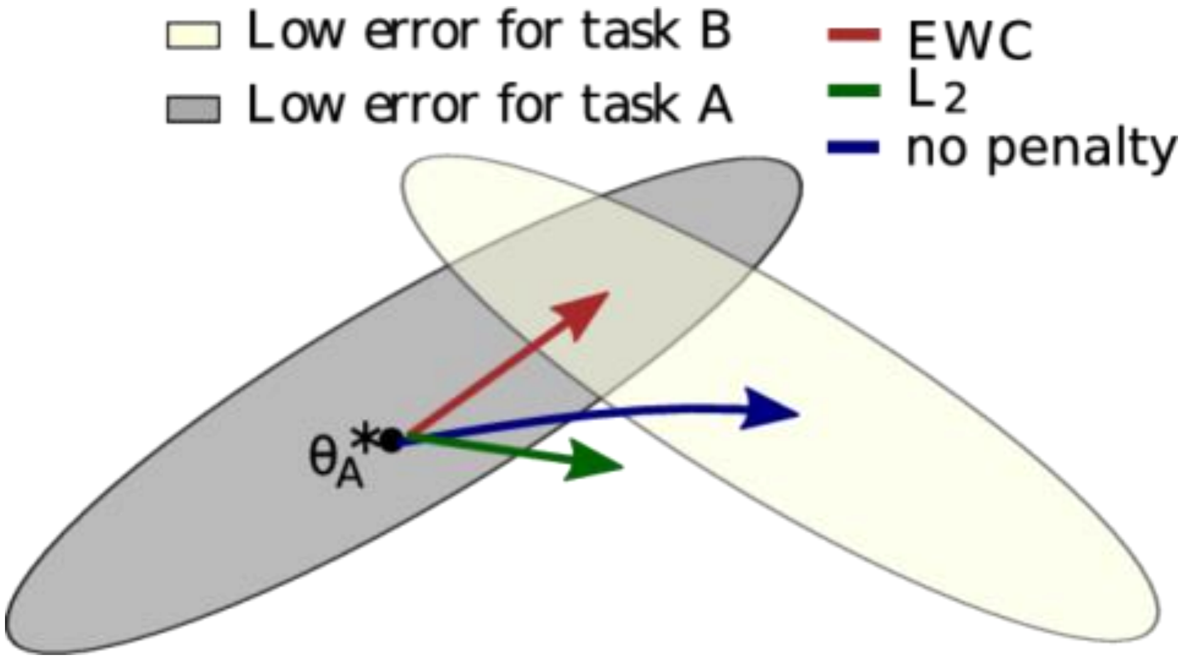
$$p(y_1 \mid x, D_1 \rightarrow D_2)$$

- Concerned about Model's ability to learn n^{th} task given $n - 1$ task

- Concerned about Model's ability to remember $n - 1$ tasks given training on new n^{th} task.
- Interference due to new task, the old task is forgotten

Core Idea in a picture

Penalize, but softly and choose whom to penalize



θ_A^* *are the optimum parameters (solution) for* $Task_A$

- For 'no penalty', we don't do bad, at least we are good for $Task_B$

- For L_2 penalty, we neither do good on $Task_A$ nor on $Task_B$
- This is worse than 'no penalty'. Too restrictive.

- For EWC penalty, we do good both on $Task_A$ and on $Task_B$. A softer way to penalize.
- The new optimum lies in the low error planes for both $Task_A$ and $Task_B$

- We typically look for low error (plane / zone..) for the parameters.
- When I'm learning new task, I would like my parameters to be close to original task's parameters θ_A^*
- There are a lot of parameters to play with, so we can choose whom to modify / penalize modification

Core Idea in math

If θ is weights of model and D is data distribution, then we are concerned about what is the best θ that would fit the data D

$$p(\theta|D) = \frac{p(D|\theta).p(\theta)}{p(D)}$$

or, $\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log p(D)$

$\log p(D|\theta)$: *best data distribution for θ, loss term* $(-\mathcal{L}(\theta))$

$\log p(\theta), \log p(D)$: *priors of θ (initialization) and Data*

Extending this to scenarios where we have one Data after another

$$\log p(\theta|D) = \log p(D_B|\theta) + \log p(\theta|D_A) - \log p(D_B)$$

D : *Entire Data* $- D_A + D_B$

$\log p(\theta|D)$: *represents the overall loss*, $\mathcal{L}(\theta)$,

$\log p(D_B|\theta)$: *represents the loss for the task B*, $\mathcal{L}_B(\theta)$, *Likelihood*

$\log p(\theta|D_A)$: *This should capture essence of first task*

$\log p(\theta|D_A)$: *an assumption is made to find this*

$$\log p(\theta|D_A) \approx \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

λ : *Weightage of old task*

F_i : *Fisher Information Matrix (diagonal entry)*

This can be thought of a matrix which gives importance to each weight

$$F_i = \mathbb{E}_x (\partial_{\theta_i} \log p_\theta(x))^2$$

We can now say that

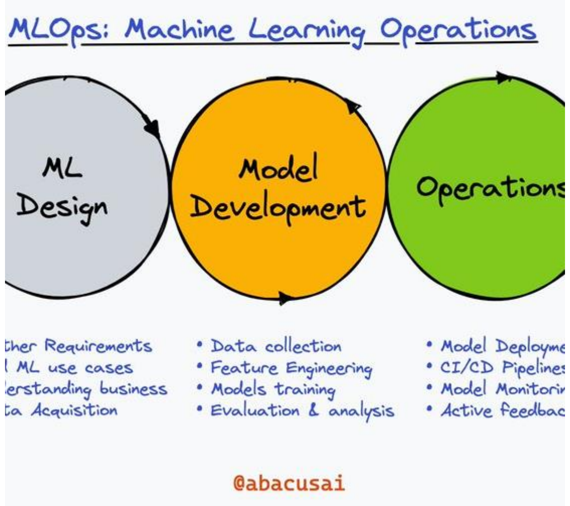
$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

Toy Example

https://github.com/peeyushsinghal/ContinualLearning/blob/main/EWC_experiment.ipynb

Relevance in Mapmaking

Where all we can use continual learning



- Cerebro – One model for MoMa and Mapillary images
- Same model for APTs and POIs

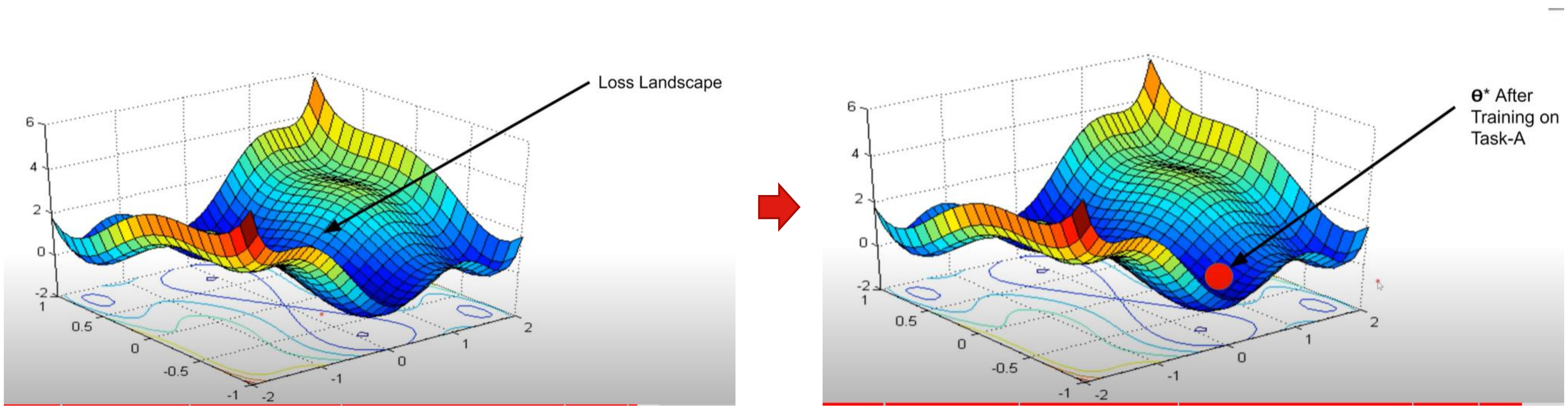
- Extending Models to similar datasets, BFP creation models

- Reduction in number of models, less burden for MLOPs

- … many more

Fisher Information Matrix – view 1

Intuition and match

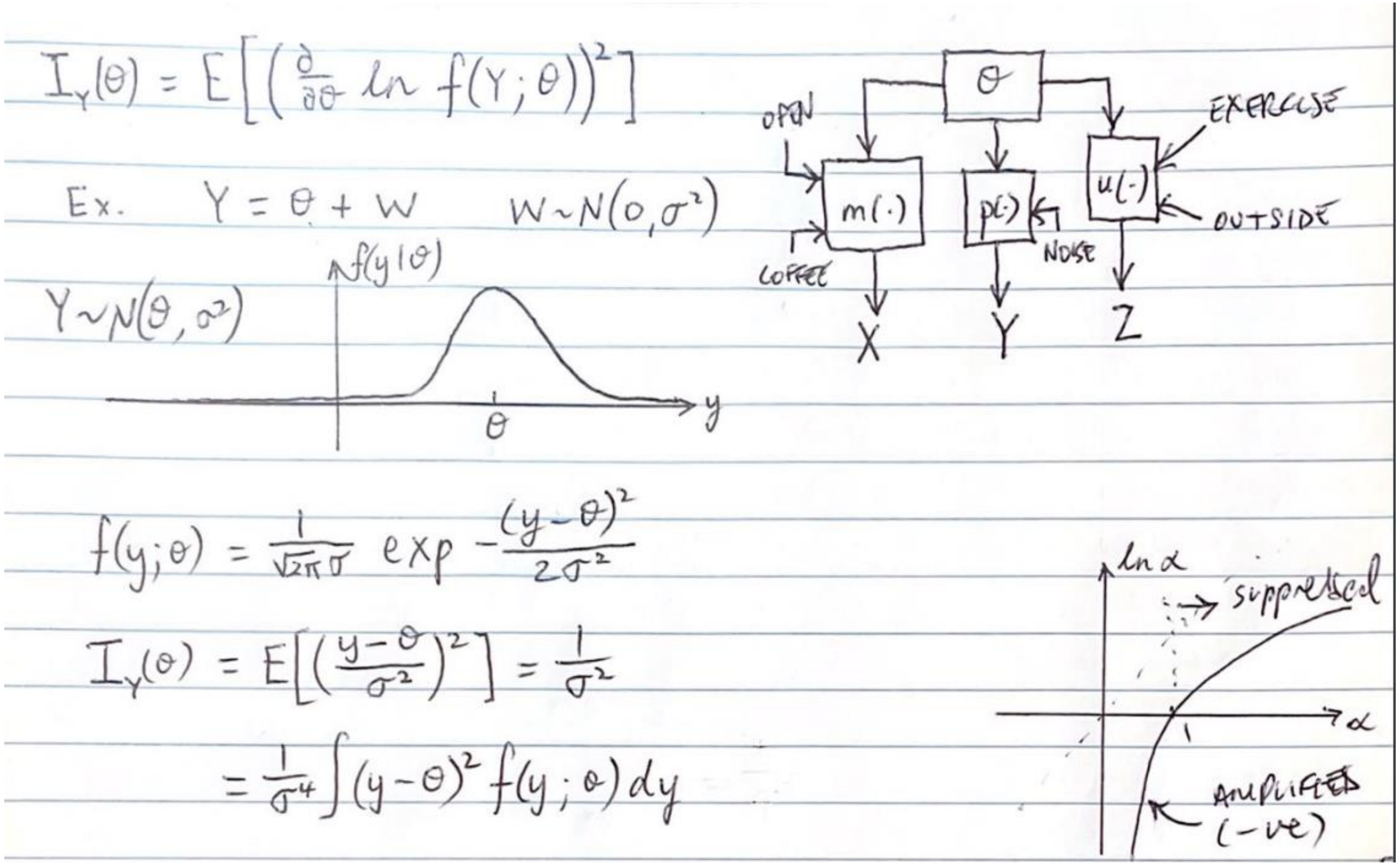


Perturbing the weight in different direction helps us understand where the impact of movement is high

- Ideally, we would like to understand the curvature of $\mathcal{L}(\theta)$, using the Hessian (second derivative), but that is intractable due to a large number of parameters. Please note that already first derivative is 0 at θ^*
- Instead, we approximate Hessian with the diagonal of the empirical Fisher Matrix. It provides a view of the loss landscape using double derivative
- Loss takes form of multivariate Gaussian with diagonal covariance

Fisher Information Matrix – view 2

Intuition and math



Source : https://www.iaincollings.com/probability-and-random-variables

Fisher Information Matrix – view 3

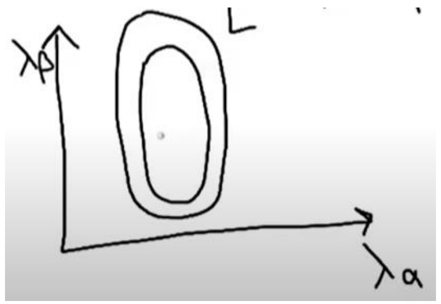
Intuition and math

$$F_{\alpha\beta}^{-1} = C_{\alpha\beta}$$

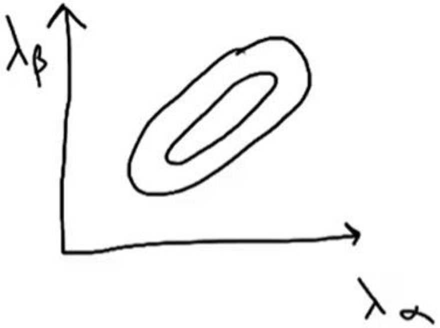
α, β *are two weights of* θ

C is covariance matrix

$$C_{\alpha\beta} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\beta\alpha} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$



$\sigma_{\alpha\beta} = 0$ and $\sigma_{\beta\alpha} = 0$



$\sigma_{\alpha\beta} \neq 0$ and $\sigma_{\beta\alpha} \neq 0$

Fisher Information Matrix looks to be curvature matrix : Bigger the Fisher Information Matrix, smaller the covariance matrix (therefore the variances), the smaller the contours, the peakier / curved our loss landscape is.

For simplified perspective, we take $\sigma_{\alpha\beta} = 0$ and $\sigma_{\beta\alpha} = 0$, therefore we look at the diagonal of Fisher Information Matrix

Thank You