From Measure Theory to Code (Neural Networks)

AI SG #3

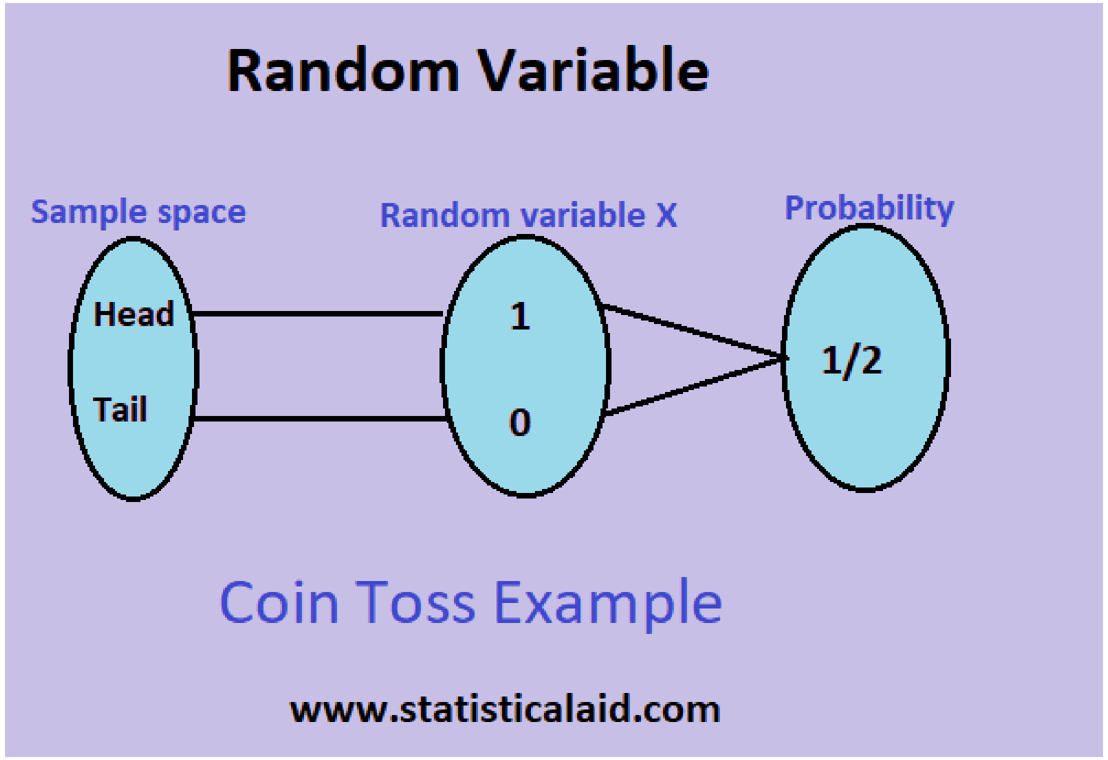
Probability Space

- Sample Space: The set of all elementary outcomes
- Sigma-Algebra/Event Space: A collection of subsets of SS (events) that we want to measure…
- Probability Measure: A function mapping events to [0, 1]


+ Γράμματα = ???

Random Variables

- A function $X(\theta): \Theta \rightarrow R$ mapping outcomes to real numbers



Expectation

$$G(\theta) = E[H(x, \theta)] = \int_{-\infty}^{\infty} H(x, \theta) f_x(x) dx$$

- The Expectation of a function H(x) is the integral of that function weighted by how likely each value x is!

Τότε γιατί χρειαζόμαστε νευρωνικά;

- Ποια είναι η PDF του 0 ή του 8; Ποια είναι η PDF της γάτας;
- Έχουμε στην διάθεσή μας μόνο δείγματα από τα δεδομένα!!!

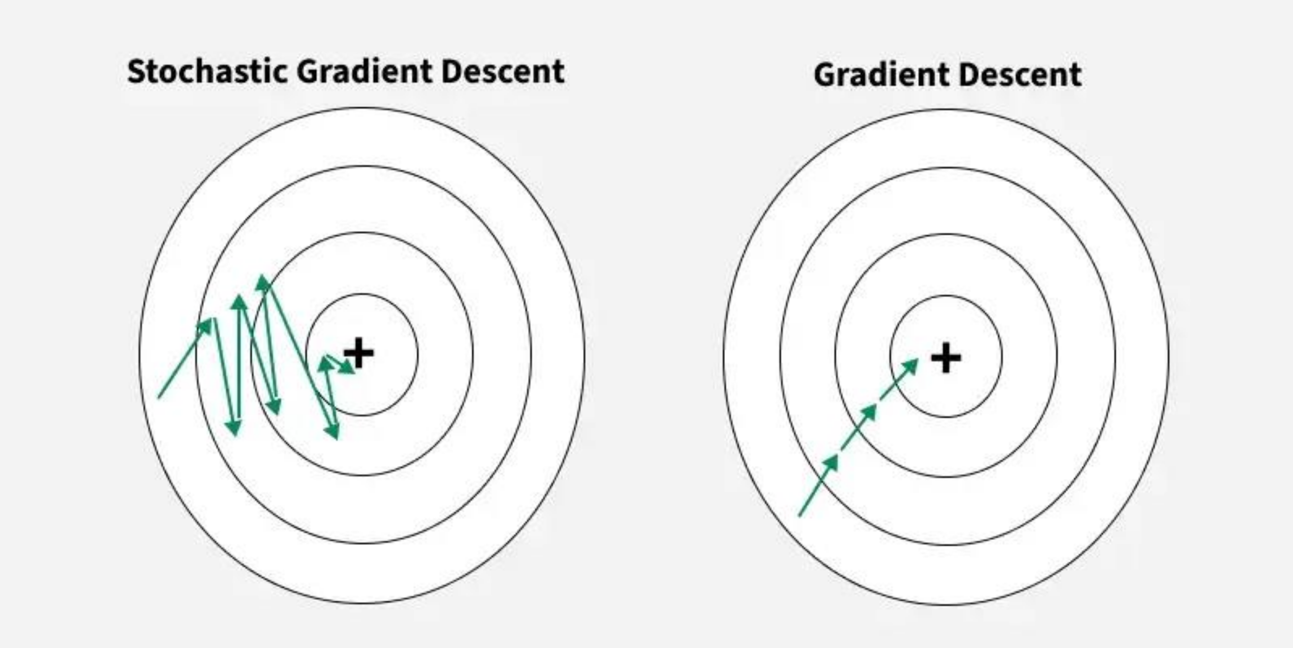
Law of Large Numbers

$$\nabla_\theta G(\theta) \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta H(x_i, \theta)$$

- Batch Gradient? (slow…)

- Stochastic Gradient Descent [SGD] (just 1 single sample)!!!

$$\theta_t = \theta_{t-1} - \mu \nabla_\theta H(x_t, \theta_{t-1})$$



Cybenko's Theorem

- From 1989, reveals how a finite sum of sigmoidal functions can densely approximate any continuous function!

Theorem (Cybenko)

Let σ be any continuous discriminatory function. Then finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j), \text{ where } w_j \in \mathbb{R}^n,\ \alpha_j,\ b_j \in \mathbb{R}$$

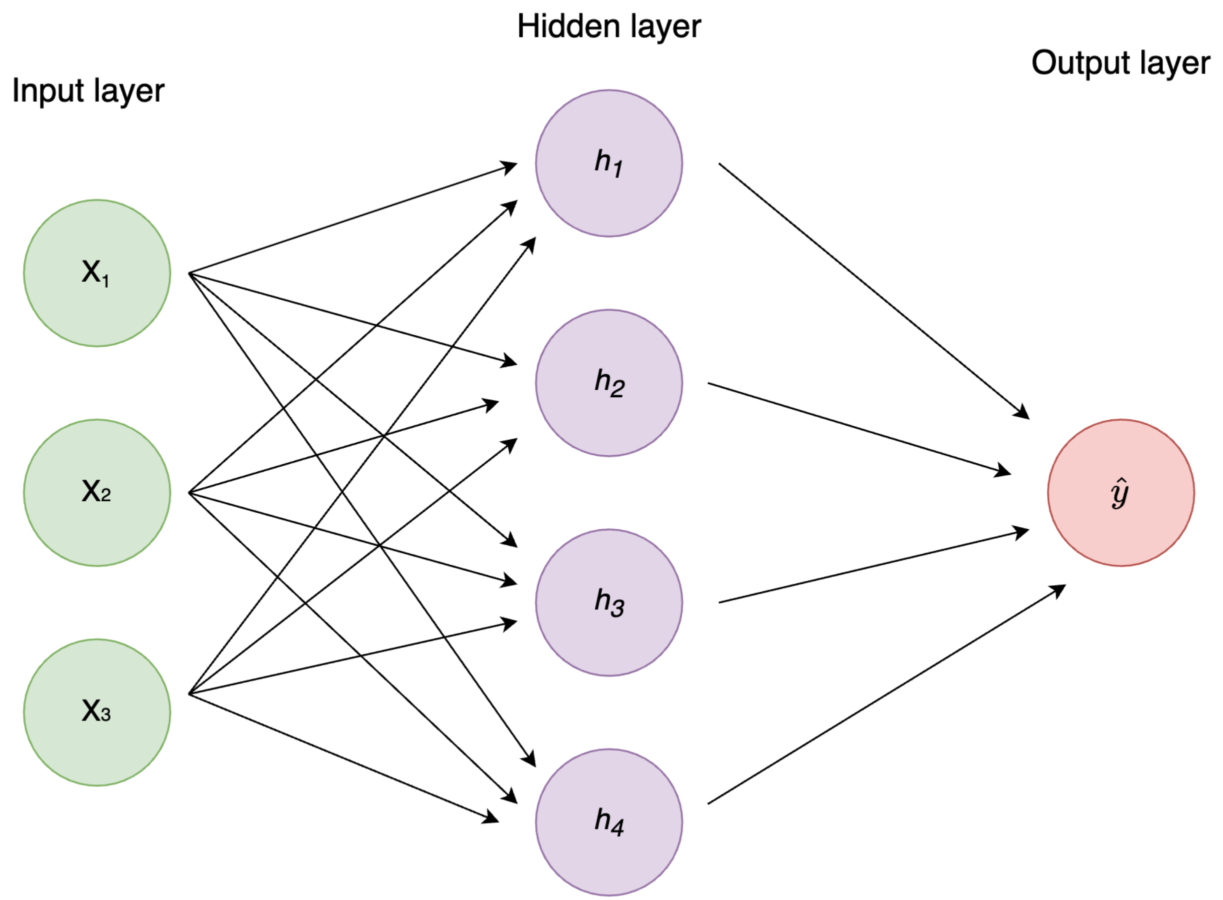
are dense in $C(I_n)$.

In other words, given any $\varepsilon > 0$ and $f \in C(I_n)$, there is a sum $G(x)$ of the above form such that

$$|G(x) - f(x)| < \varepsilon, \quad \forall x \in I_n$$

https://medium.com/@ML-STATS/understanding-the-universal-approximation-theorem-8bd55c619e30

NN with 1 hidden layer (approximate any foo!)



Likelihood Ratio Test (LRT)

$$r(x) = \frac{f_1(x)}{f_0(x)} \gtrless \eta$$

- Optimal Decision Rule

ROC Curve (Receiver Operating Characteristic)

- Η διακεκομμένη γραμμή είναι ουσιαστικά η ρίψη νομίσματος μεταξύ 0 και 8! Αν η πορτοκαλί ταυτιζόταν με αυτή , τότε το μοντέλο τυχαία (50 - 50) δίνει 0 ή 8!
- Το ότι η πορτοκαλί καμπύλη ταυτίζεται με τον άξονα Y = 1, σημαίνει ότι βρήκε σχεδόν όλα τα 8 χωρίς να τα κάνει mislabel ως 0...