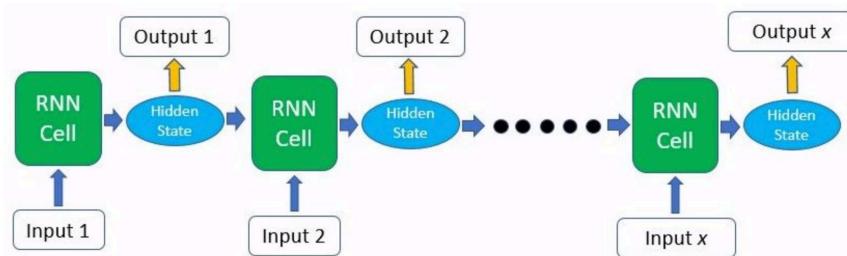


# Recurrent Networks and LSTMs

Olivier Dubrule and Navjot Kukreja

1

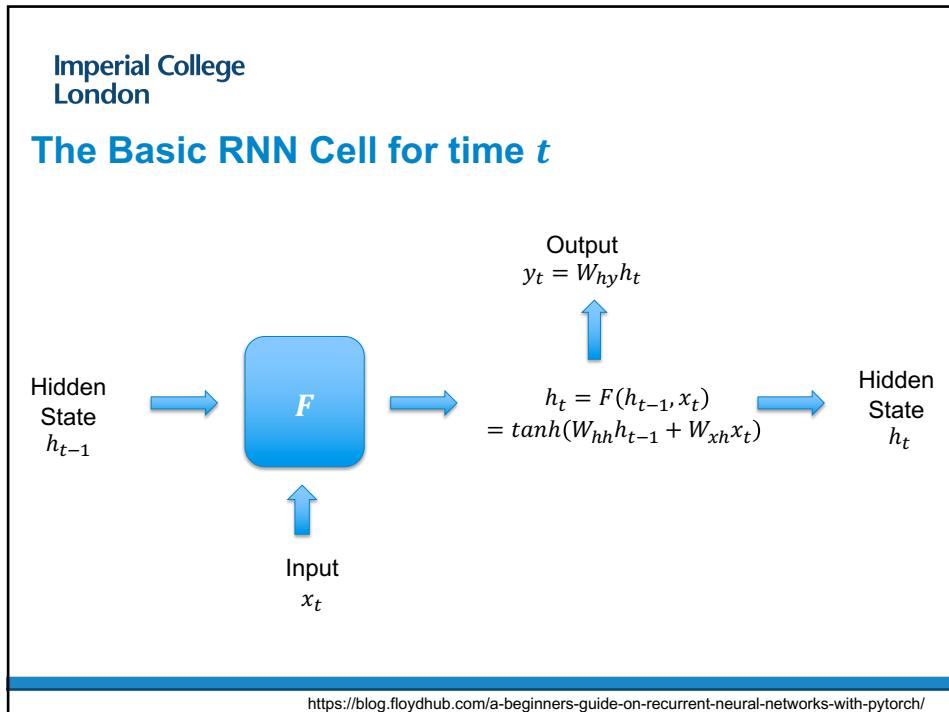
## Basic Recurrent Network Structure

**Example:**

Based on a (usually long) text used for Training, we wish to generate “similar” text on a character by character basis.

<https://blog.floydhub.com/a-beginners-guide-on-recurrent-neural-networks-with-pytorch/>

2



3

Imperial College London

### A Simplistic Character Sequence Example (3)

**Training Workflow:**

For a sequence of  $p$  (say  $p = 20$ ) characters:

- Take first character  $x_1$ , combine with current hidden state  $h_0$  (the size of  $h_0$  is a hyper-parameter) to derive  $h_1$ , calculate associated output vector  $y_1$  of size equal to vocabulary, transform  $y_1$  into Softmax vector, calculate cross-entropy with second character  $x_2$  of sequence.
- Take this second character as input to re-do what we just did with the first character and again calculate loss function using the value of the third character of the sequence, and so on until we reach the end of the sequence.
- Add up all the above loss functions.
- Do back-propagation to calculate gradients according to each trainable parameter.
- Modify parameters by gradient descent

Move to next sequence of  $p$  characters, until end of Training Set is reached.

4

Imperial College  
London

## A Simplistic Character Sequence Example (4)

### Text Generation Workflow:

Generate a new sequence of  $n$  characters:

Randomly sample first character  $x_1$ , combine with initial null state vector value  $h_0$  to derive new vector  $h_1$ , calculate associated output vector  $y_1$  of size equal to vocabulary size, transform  $y_1$  into Softmax vector, sample new character  $x_2$  based on Softmax probability, and so on, until the required number of characters have been generated.

5

Imperial College  
London

## A Simple Language Processing Python Code

<https://gist.github.com/karpathy/d4dee566867f8291f086>

Learning to write Shakespeare plays

After many iterations

at first:

```
tyntd-lafhatawiaoahrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tkirgd t o idee ns,smtt h ne etie h,hregtrs nigtike,aoennts lbg
```

↓ train more

```
"Tmont thithey" fomesscerlind
Keushey. Thom here
sheulke, amerenith ol sivh I lalterthend Bleipile shuw fil on asterlome
coanigennc Phe lism thond hon at. Meidmorotion in ther thize."
```

↓ train more

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwege fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

↓ train more

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had offended him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

↓ train more

```
VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.
```

↓ train more

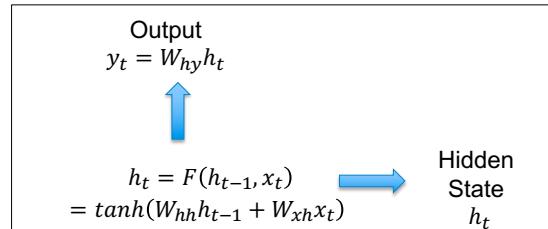
```
KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
```

Andrej Karpathy's Blog: *The Unreasonable Effectiveness of Recurrent Neural Networks*

6

Imperial College  
London

## A Summary



The weight matrices  $W_{hh}$ ,  $W_{xh}$  and  $W_{hy}$  contain the trainable parameters and are the same for each time  $t$ .

The hidden vector  $h_{t-1}$ , the size of which is a hyperparameter, changes with  $t$  and contains recent historical information about the  $x_t$  sequence before  $t$ .

$x_t$  is the new data at time  $t$  which is combined with  $h_{t-1}$  to predict  $x_{t+1}$ .

<https://www.youtube.com/watch?v=6niqTuYFZLQ>

7

Imperial College  
London

## Visualizing one coordinate of $h_t$ as a function of $t$

Beginning of Quote  

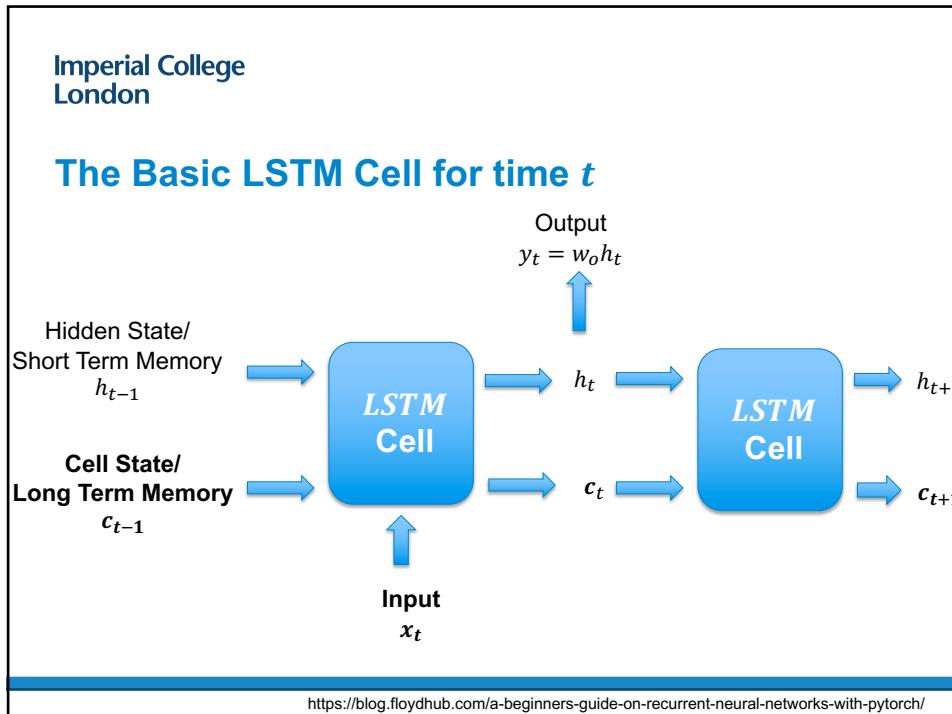
 "You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.  
 Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."  
 End of Quote

Beginning of Quote  

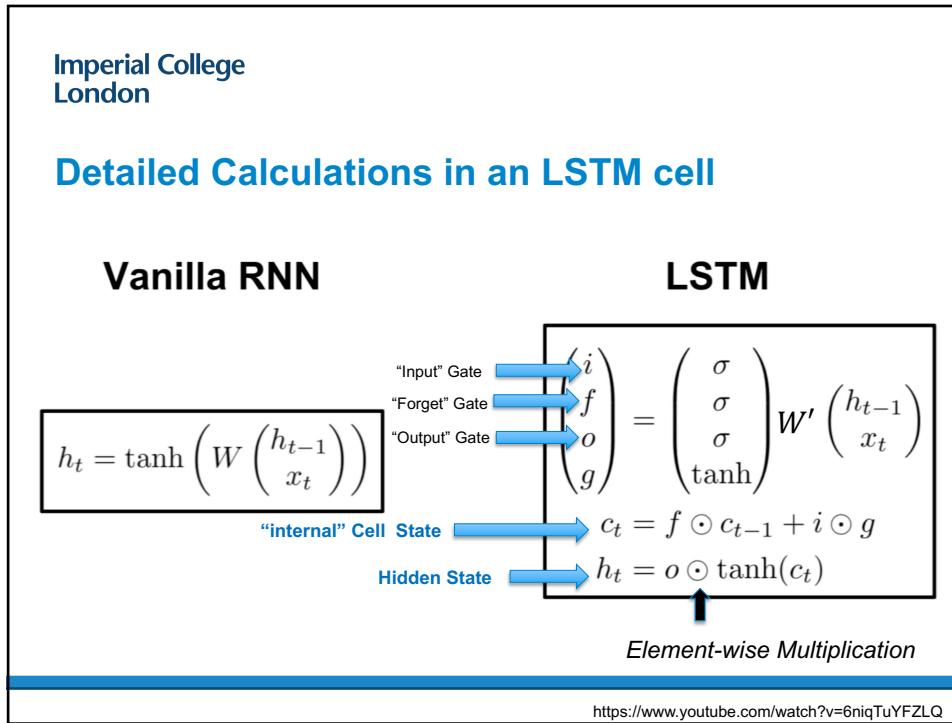
 "You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.  
 Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."  
 End of Quote

From Karpathy, Johnson and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016

8



9



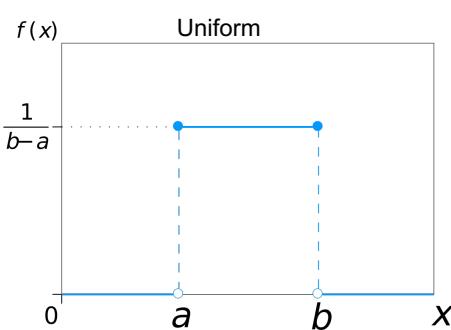
10

# Probabilities for Deep Learning

Olivier Dubrule and Navjot Kukreja

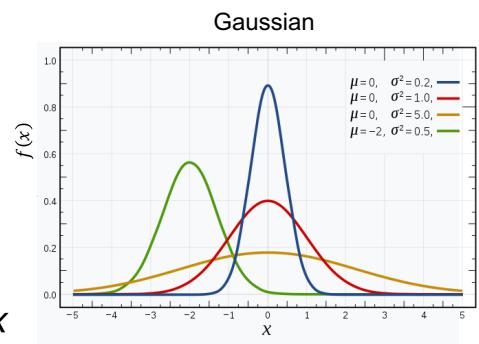
11

## The Uniform and the Normal (or Gaussian) pdfs



$$f(x) = \frac{1}{b-a} \text{ if } x \in [a, b] \text{ and } f(x) = 0 \text{ if } x \notin [a, b]$$

$f(x)$  also written  $U(x; a, b)$



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$f(x)$  also written  $N(x; \mu, \sigma^2)$

[https://en.wikipedia.org/wiki/Normal\\_distribution](https://en.wikipedia.org/wiki/Normal_distribution)

12

Imperial College  
London

### General Properties of a PDF $f(x)$

$$P(a < X < b) = \int_a^b f(x)dx \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

**Mean**  $E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$

**Variance**  $Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$   
 $= E[(X - \mu)^2] = E(X^2) - \mu^2$

13

Imperial College  
London

### Multivariate Normal pdf of a Random Vector $(X_1, \dots, X_n)$

$$f(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right]$$

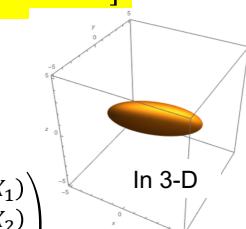
$x$  is the  $n \times 1$  vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$\mu$  is the  $n \times 1$  expectation vector

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

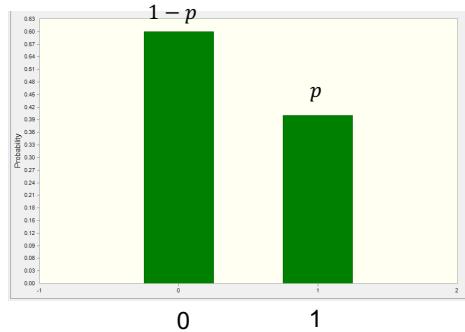
$\Sigma$  is the  $n \times n$  variance-covariance matrix  $\Sigma = \left( \left( Cov(X_i, X_j) \right)_{i,j=1,\dots,n} \right)$



14

Imperial College  
London

### Probability Distribution of a Bernoulli Random Variable



A Bernoulli variable takes the value 1 with probability  $p$  and 0 with probability  $(1 - p)$ , and we can write that the probability that it is equal to  $x$  is:

$$b(x) = p^x(1 - p)^{1-x}$$

15

Imperial College  
London

### Independent and Identically Distributed (IID)

In probability theory, a sequence or collection of random variables is independent and identically distributed (**i.i.d.** or **iid** or **IID**) if each random variable has the same probability distribution as the others and they all are mutually independent.

When treating  $m$  samples from a training or test dataset (for instance a set of images), it is assumed they are IID.



16

Imperial College  
London

### Problem Addressed by the Maximum Likelihood Method

Assume we have  $m$  independent (IID) data points  $x^{(1)}, x^{(2)}, \dots, x^{(m)}$

Each data point is  $x^{(i)}$  is composed of  $n$  features.

We want to fit a multivariate (for instance multivariate Gaussian) pdf  $p_\theta(x)$  of dimension  $n$  to these  $m$  samples.

Maximum likelihood consists of calculating the parameters  $\theta$  such that the  $m$  samples maximize their likelihood.

But what is the likelihood? The likelihood is the probability to obtain the  $m$  sample values  $x^{(1)}, x^{(2)}, \dots, x^{(m)}$  assuming that  $p_\theta(x)$  is the probability of  $x$ .

17

Imperial College  
London

### Likelihood and Maximum Likelihood for a Dataset of Images

The pdf  $p_\theta(x_1, x_2, \dots, x_n)$  is parametrized by  $\theta$ . If there is just one image  $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$  in the dataset, its likelihood is defined as:

$$\text{Likelihood of image } x^{(1)} = p_\theta(x^{(1)}) = p_\theta(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$$

The Maximum Likelihood estimate  $\theta_{ML}$  of  $\theta$  is calculated as

$$\theta_{ML} = \operatorname{argmax}(p_\theta(x^{(1)})) \quad \text{or} \quad \theta_{ML} = \operatorname{argmax}(\log p_\theta(x^{(1)}))$$

If there are  $m$  IID images  $x^{(1)}, x^{(2)}, \dots, x^{(m)}$  in the dataset

$$\begin{aligned} \theta_{ML} &= \operatorname{argmax}(p_\theta(x^{(1)})p_\theta(x^{(2)}) \dots p_\theta(x^{(m)})) \\ &= \operatorname{argmax}(\log(p_\theta(x^{(1)})p_\theta(x^{(2)}) \dots p_\theta(x^{(m)}))) = \theta_{ML} = \operatorname{argmax}\left(\sum_{i=1}^m \log p_\theta(x^{(i)})\right) \end{aligned}$$

18

Imperial College  
London

### Maximum Likelihood for a Bernouilli Distribution

A Bernouilli distribution has just one parameter  $p$

Suppose we have just one sample  $x_1$  (which has the value 1 or 0)

Its likelihood is:  $b(x_1) = p^{x_1}(1 - p)^{1-x_1}$

Its log-likelihood is:  $\log b(x_1) = x_1 \log p + (1 - x_1) \log(1 - p)$

If we have  $m$  samples  $x_i$ , their log-likelihood is:

$\sum_{i=1 \dots m} (x_i \log p + (1 - x_i) \log(1 - p))$  (this is equal to minus the cross-entropy!)

The maximization leads, unsurprisingly, to :  $p_{ML} = \frac{1}{m} \sum_{i=1 \dots m} x_i$ .  
( $p_{ML}$  is the proportion of samples equal to 1)

19

Imperial College  
London

### Compare two pdfs: the Kullback-Leibler (KL) Divergence

For two pdfs  $p(x)$  and  $q(x)$  :

$$D_{KL}(p\|q) = \int_{-\infty}^{+\infty} p(x) \log p(x) dx - \int_{-\infty}^{+\infty} p(x) \log q(x) dx$$

**Two Fundamental Properties**

$D_{KL}(p\|q)$  is positive, and 0 if  $p(x)$  and  $q(x)$  identical  
 Asymmetry:  $D_{KL}(p\|q) \neq D_{KL}(q\|p)$

20

Imperial College  
London

## KL Divergence versus Maximum Likelihood

If the Training Set consists of  $m$  IID data points  $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ , the KL Divergence  $D_{KL}(p_d \| p_\theta)$  between the experimental distribution  $p_d$  of the Training Set and any theoretical distribution  $p_\theta$  is:

$$D_{KL}(p_d \| p_\theta) = \int_{-\infty}^{+\infty} p_d(x) \log p_d(x) dx - \int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

Minimizing  $D_{KL}(p_d \| p_\theta)$  in the parameters  $\theta$  is equivalent to maximizing the second term:

$$\int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

But this is the expression of the expectation of  $\log p_\theta(x)$  calculated over the Training Set! Hence

$$\theta = \operatorname{argmax} \left( \frac{1}{m} \sum_{i=1}^m \log p_\theta(x^{(i)}) \right)$$

***Minimizing the KL Divergence is equivalent to maximizing the Likelihood!***