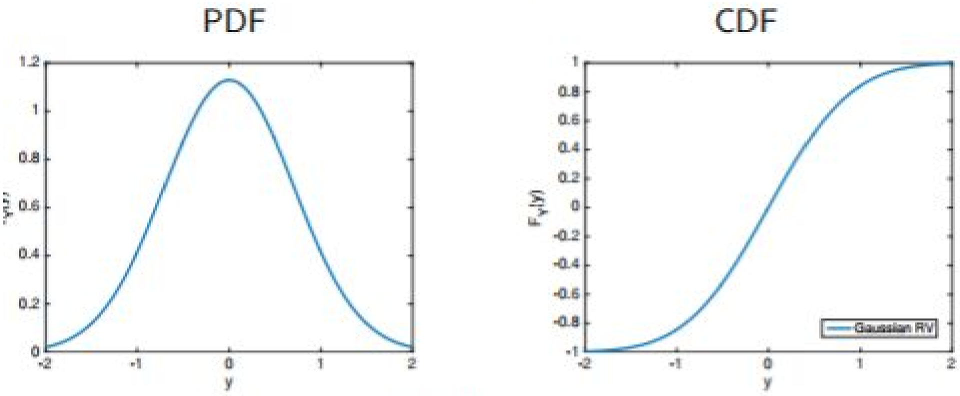
Probabilistic Interpretation

- **Probability Refresh**
  - **Sample space S:** set of possible outcomes
  - **Random Event E:** subset of S
  - **Random Variable:** maps event E to a real value (**denoted by Capitals**)
  - **Conditional Probability**
    - **Events:** $P(E|F) = \frac{P(E \cap F)}{P(F)}$ $when\ P(F) > 0$
    - **RVs:** $P(X = x|Y = y) = \frac{P(X=x\ and\ Y=y)}{P(Y=y)}$
  - **Chain Rule:** $P(X = x\ and\ Y = y) = P(X = x|Y = y)P(Y = y)$
  - **Consequences of the Chain Rule**
    - **Marginalisation:** Suppose RV Y takes values in $\{y_1, y_2, ..., y_n\}$ Then
      - $P(X = x) = P(X = x\ and\ Y = y_1) + ... + P(X = x\ and\ Y = y_n)$
      - $= \sum_{i=1}^{n} P(X = x|Y = y_i)P(Y = y_i)$
    - **Bayes Rule:** $P(X = x|Y = y) = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$
    - **Independence:** Random variables X and Y are independent if…
      - $P(X = x\ and\ Y = y) = P(X = x)P(Y = y)$
      - For all x and y, in which case: $P(X = x|Y = y) = P(X = x)$
  - **Continuous-valued random variables**
    - $P(X = x) = 0$ for continuous-valued random variables, so we consider intervals instead: $P(a \le X \le b)$
    - $F_Y(y) := P(Y \le y)$ is the **cumulative distribution function (CDF) and** $P(a < Y \le b) = F_Y(b) - F_Y(a)$
    - For a continuous-valued random variable, Y, there exists a **probability density function** $f_Y(y) \ge 0$ such that:
      - $F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt$
      - And so…
      - $P(a < Y \le b) = \int_{-\infty}^{b} f_Y(t)dt - \int_{-\infty}^{a} f_Y(t)dt = \int_{a}^{b} f_Y(t)dt$
      - **The probability density function f(y) for random variable Y is <u>not</u> a probability (it can take values greater than 1) - the <u>area under the PDF is the probability</u>** $P(a < Y \le b)$
    - $\int_{-\infty}^{\infty} f(y)dy = 1\ (since\ \int_{-\infty}^{\infty} f(y)dy = F_Y(\infty) = P(Y \le \infty) = 1)$
  - **CDF for X and Y:** $F_{XY}(x, y) = P(X \le x\ and\ Y \le y)$
    - Well defined for both continuous and discrete valued RVs

- When X and Y are continuous-valued RVs there exists a PDF $f_{XY}(x,y) \geq 0$ such that: $F_{XY}(x,y) = \int_{\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)du\,dv$
- Define conditional PDF: $f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$
- Then the chain rule holds for PDFs:
  - $f_{XY}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$
  - **So marginalisation, Bayes rule and independence carry over to PDFs similarly to discrete-valued RVs**

- **Y is a Normal or Gaussian RV Y~N(μ,σ²) when it has the PDF:**

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



$\mu = 0,\ \sigma = 1$

- $E[Y] = \mu,\ Var(Y) = \sigma^2$
- Symmetric about μ and defined for all real-valued x

-

- **Probabilistic Interpretation of Linear Regression**
  - Assume output Y is generated by:
    - $Y = \theta^T x + M = h_\theta(x) + M$
    - Where $h_\theta(x) = \theta^T x$ and M is Gaussian noise with mean 0 and variance 1
  - So training data d is:
    - $\{(x^{(1)}, h_\theta(x^{(1)}) + M^{(1)}), (x^{(2)}, h_\theta(x^{(2)}) + M^{(2)}), ..., (x^{(m)}, h_\theta(x^{(m)}) + M^{(m)})\}$
    - Where $M^{(1)}, M^{(2)}, ..., M^{(m)}$ are independent RVs each of which is Gaussian with mean 0 and variance 1
  - A Gaussian RV Z with mean μ and variance σ^2 has PDF:
    - $f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(Z-\mu)^2}{2\sigma^2}}$
  - So we are assuming:
    - $f_M(m) = \frac{1}{\sqrt{2\pi}}e^{-\frac{m^2}{2}}, f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(y-h_\theta(x))^2}{2}}$

- The **likelihood** $f_{D|\Theta}(d|\theta)$ of the training data d is therefore:
  - $f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}}$
- Taking logs: $log\, f_{D|\Theta}(d|\theta) = log\, \frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{m} \frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}$
- And the maximum likelihood estimate of θ maximises:
  - $max_\theta - \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$
  - **I.e. it minimises:** $min_\theta \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$

- **Why do we care about probabilistic interpretation?**
  - Probability is the reasoning about uncertainty, it would be strange if machine learning algorithms didn't make sense from a probability perspective
  - Casting an ML approach within a statistical framework clarifies the assumptions we (implicitly) make e.g. in Linear Regression:
    - **Noise is additive:** $Y = \theta^T x + M$
    - **Noise** on each observation **is independent and identically distributed**
    - **Noise is Gaussian** - this is what drives our usage of square loss. Changing the noise model would lead to a different loss function
  - Allows us to utilise the results and approaches of probability/statistics, and perhaps gain new insights. E.g. in linear regression:
    - Without regularisation, our estimate of θ is the maximum likelihood estimate. Would a MAP (Maximum A Posteriori) estimate be more/less useful?

- **Probabilistic Interpretation of Logistic Regression**
  - Assume:
    - $P(Y = y|\theta, x) = \frac{1}{1+e^{-y\theta^T x}}$
    - And recall y = 1 or y = -1 only
  - The **likelihood** of training data d is: $f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{1+e^{-y\theta^T x}}$
  - **Taking logs:** $log\, f_{D|\Theta}(d|\theta) = \sum_{i=1}^{m} log\, \frac{1}{1+e^{-y\theta^T x}}$
  - **And the maximum likelihood estimate of θ minimises:**
    - $- \sum_{i=1}^{m} log\, \frac{1}{1+e^{-y\theta^T x}} = \sum_{i=1}^{m} log(1 + e^{-y\theta^T x})$

- Since -log(z)=log(1/z)
- **The probabilistic formulation of logistic regression provides us with new insight:**
    - $P(Y=y|\theta,x) = \frac{1}{1+e^{-y\theta^T x}}$
- So in addition to our prediction $h_\theta(x) = sign(\theta^T x)$ we also have a confidence in the prediction: $\frac{1}{1+e^{-y\theta^T x}}$
    - **When $\frac{1}{1+e^{-y\theta^T x}}$ is close to 1 we are confident, close to zero we are less confident**

- **Probabilistic Interpretation of Regularisation**

$$\underbrace{P(\Theta=\vec{\theta}|D=d)}_{\text{posterior}} = \frac{\underbrace{P(D=d|\Theta=\vec{\theta})}_{\text{likelihood}}\underbrace{P(\Theta=\vec{\theta})}_{\text{prior}}}{P(D=d)}$$

- **Bayes Rule:**
    - **Likelihood:** probability of seeing the data d, given the model with parameter $\Theta=\theta$ where θ is a vector
    - **Prior:** Before seeing any data what is our belief of the model… i.e. **what is probability of parameter values Θ**
    - **Posterior:** after seeing the data, what is our belief about probability of parameter values Θ now that we have seen the data
- **Maximum A Posteriori (MAP):** estimate of vector θ is value that maximises $P(\Theta=\theta|D=d)$
- **Maximum Likelihood estimation:** Select value that maximises $P(D=d|\Theta=\theta)$
- **Taking logs in Bayes rule:**
    - $log\,P(\Theta=\theta|D=d) = log\,P(D=d|\Theta=\theta) + log\,P(\Theta=\theta) - log\,P(D=d)$
    - Can drop the $log\,P(D=d)$ as d is fixed, so we select θ to maximise:
        - $log\,P(D=d|\Theta=\theta) + log\,P(\Theta=\theta)$
    - Or for continuous-valued RVs:
        - $log\,f_{D|\Theta}(D=d|\Theta=\theta) + log\,f_{\Theta}(\Theta=\theta)$
- **Ridge regression variant of linear regression:**

- $Y = \Theta x + M$, $M \sim N(0,1)$ as before.
- $\Theta_j, \sim N(0, \sigma^2)$ (this is our prior on θ_j), $j = 1, \ldots, n$
- log-likelihood: $-\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2$
- log-prior: $-\theta_j^2/\sigma^2$
- So MAP estimate selects θ to maximise:

$$-\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 - \sum_{j=1}^{n} \theta_j^2/\sigma^2$$

i.e. to minimise:

$$\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 + \sum_{j=1}^{n} \theta_j^2/\sigma^2$$

-