Introduction to Deep Learning

5. Maximum Likelihood and Logistic Regression

MGMT-735

SLIDES From Mu Li and Alex Smola

courses.d2l.ai/berkeley-stat-157

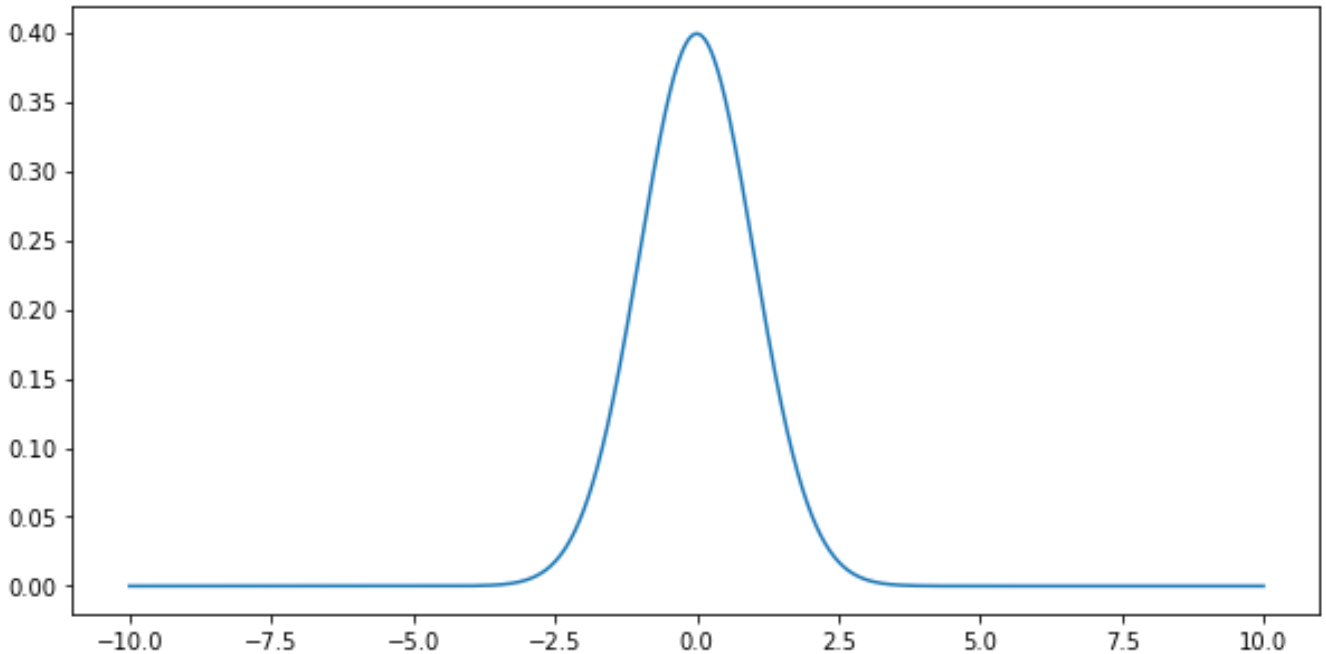
Outline

- **Maximum Likelihood**
  - Gauss and means
  - More loss functions (l_1 loss, trimmed mean)
  - Regression revisited
- **Classification**
  - Computing discrete probabilities
  - Likelihood and loss functions
- **Information Theory**

Maximum Likelihood & MAP

Flashback - Normal Distribution

Density $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Estimating the parameters in a Gaussian

- **Mean**

$$\mu = \mathbf{E}[x] \text{ hence } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Variance**

$$\sigma^2 = \mathbf{E}\left[(x-\mu)^2\right] \text{ hence } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Why?

Likelihood

- Observe some data $X = \{x_1, \ldots x_n\}$
- Assume that the data is drawn from a Gaussian

$$p(X; \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- **Fitting parameters is maximizing** $p(X; \mu, \sigma^2)$ **wrt.** μ, σ^2 (maximize likelihood that data was generated by model)
- **Practical simplification**

$$\underset{\mu,\sigma^2}{\text{maximize}}\, p(X; \mu, \sigma^2) \Longleftrightarrow \underset{\mu,\sigma^2}{\text{minimize}} - \log p(X; \mu, \sigma^2)$$

Maximum Likelihood

- Estimate parameters by finding ones that explain the data

$$\underset{\mu,\sigma^2}{\text{minimize}} - \log p(X; \mu, \sigma^2)$$

- **Decompose likelihood**

$$-\log p(X; \mu, \sigma^2) = \sum_{i=1}^{n} \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x_i - \mu)^2 = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Minimized for $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$

- Estimating the variance

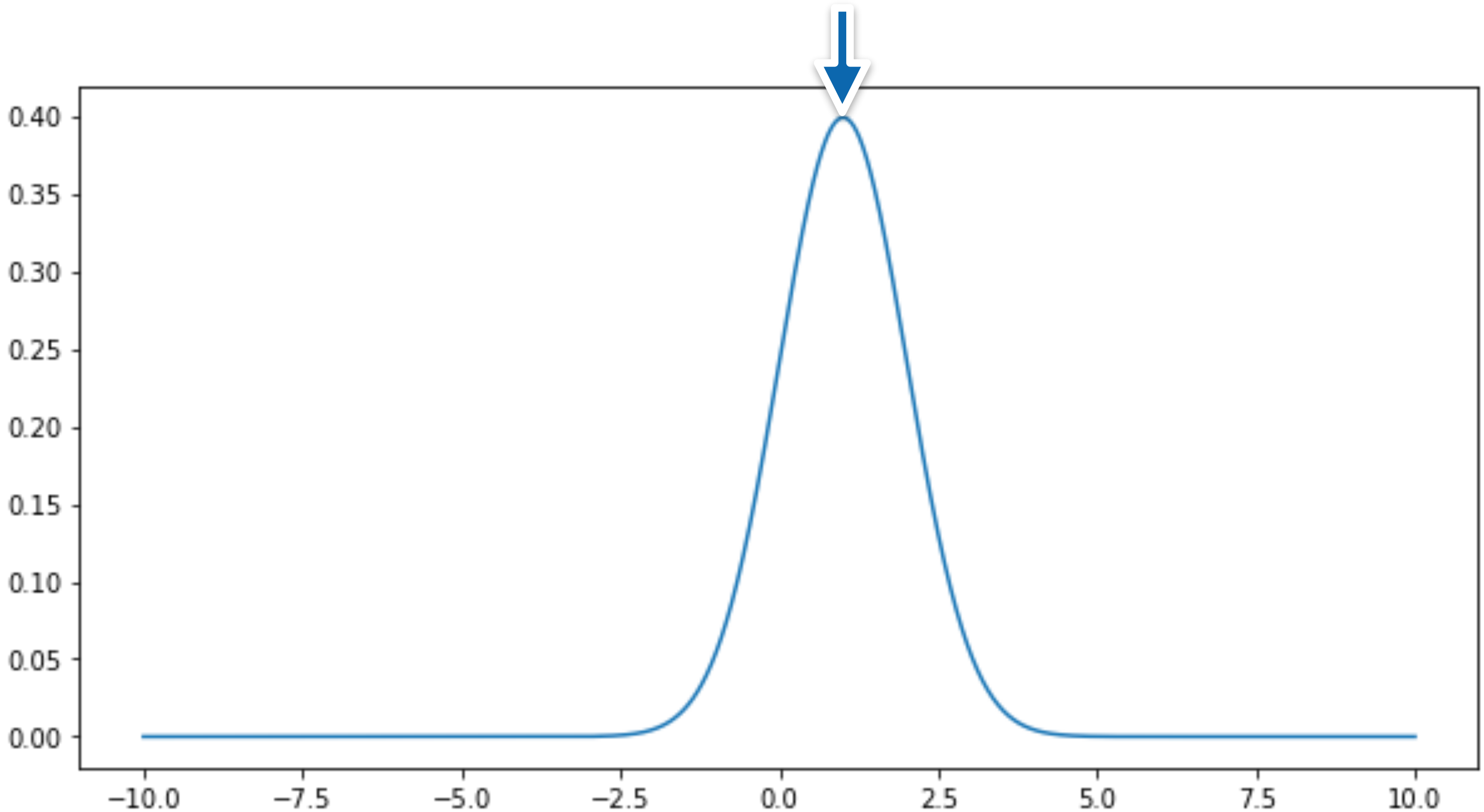
$$\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Take derivatives with respect to it

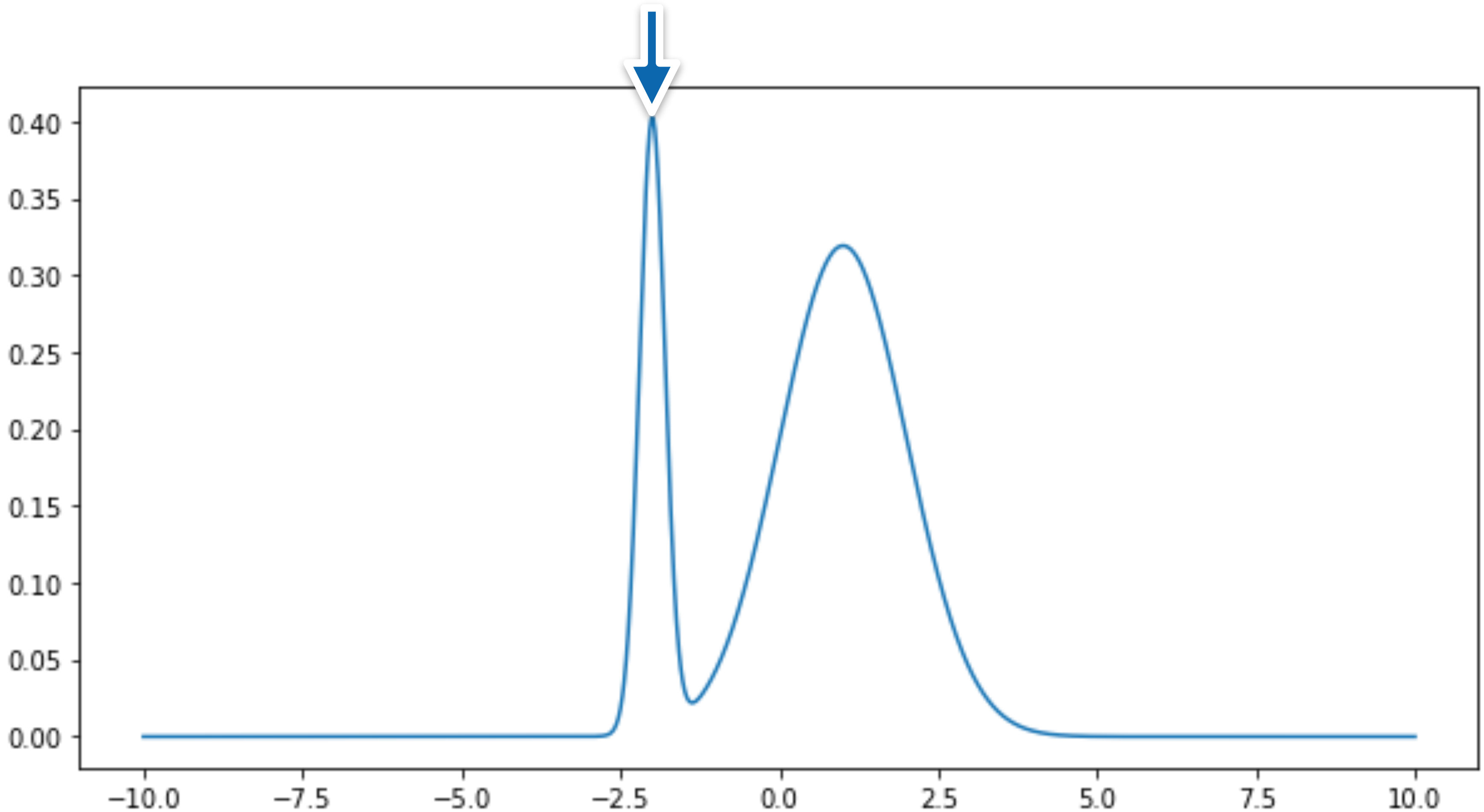
$$\partial_{\sigma^2}[\,\cdot\,] = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\implies \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

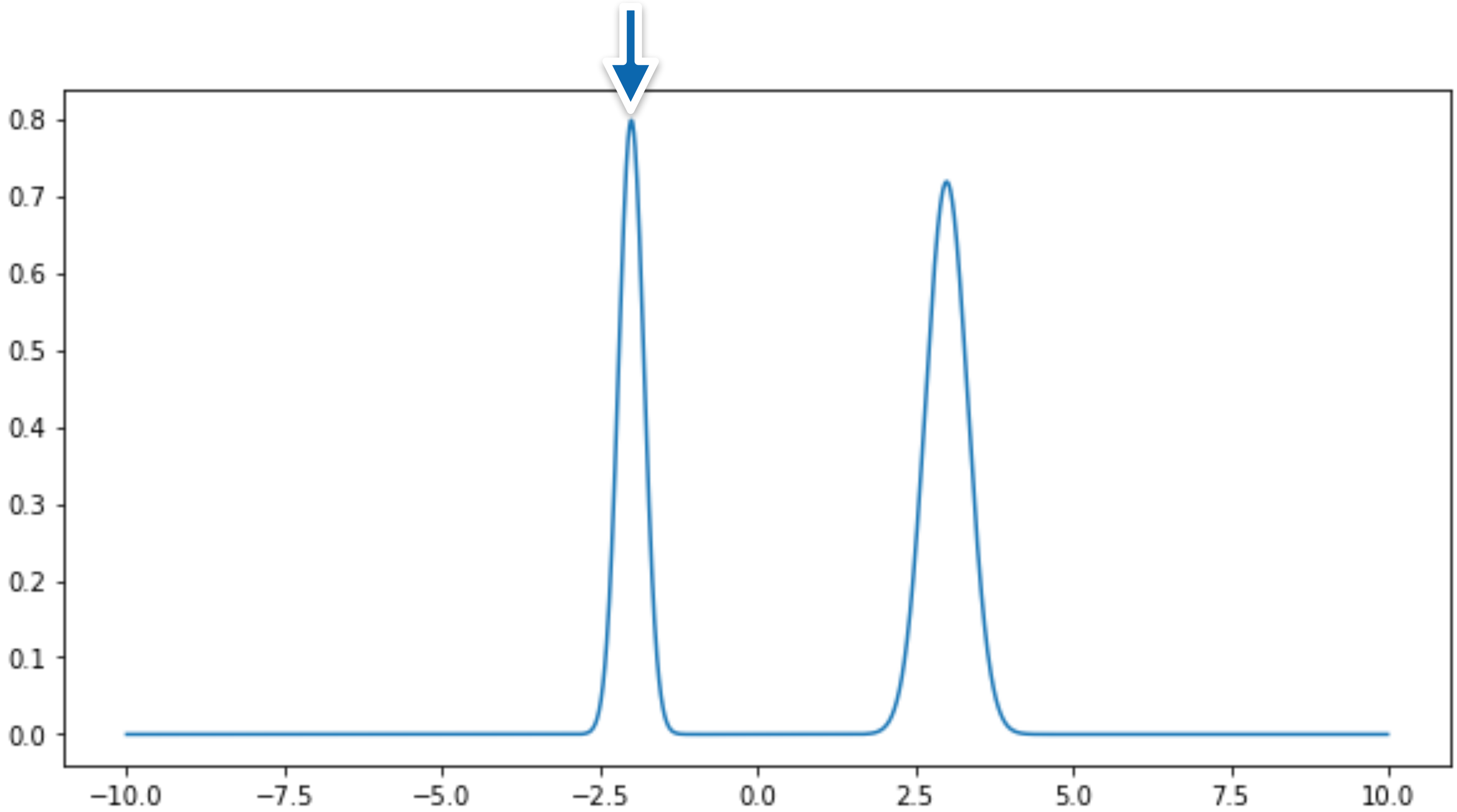
Maximum likelihood estimation



Maximum likelihood estimation



Maximum likelihood estimation



Maximum likelihood estimation

- Data - ‘student didn’t do homework’
- Possible parameters
  - **‘dog ate homework’**
  - **‘abducted by aliens’**
  - **‘too lazy’**
  - **‘sick grandmother’**

Should we believe all?

- **All parameters explain the data.**

Maximum a posteriori estimation

penalty

- Posterior Probability

$$p(w|X) \propto p(X|w)p(w)$$

hence $-\log p(w|X) = -\log p(X|w) - \log p(w) + c$

- Maximum a Posteriori Estimation

$$\underset{w}{\text{minimize}} - \log p(X;w) - \log p(w)$$

- No homework example
  p('no homework'|explanation) = 1 (all explanations work)

lazy student	grandma sick	dog ate it	alien abduction
0.8	0.19	0.0099	0.0001

What does this have to do with regression?

Regression

- Recall optimization problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i, w))^2 + \text{penalty}(w)$$

$-\log p(w)$

Does the model work?

Additive Gaussian Noise

- Data generation model

$$y_i = f(x_i, w) + \epsilon_i \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Gaussian Prior $p(w)$ hence $-\log p(w) = \frac{1}{2\bar{\sigma}^2} \|w\|^2 + \text{const}\,.$

Regression

- Maximum a posteriori

$$\underset{w}{\text{minimize}} \; -\log p(w \mid X, Y)$$

$$\Longleftrightarrow \underset{w}{\text{minimize}} \; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i, w))^2 + \frac{1}{2\bar{\sigma}^2}\|w\|^2 + \text{const}\,.$$

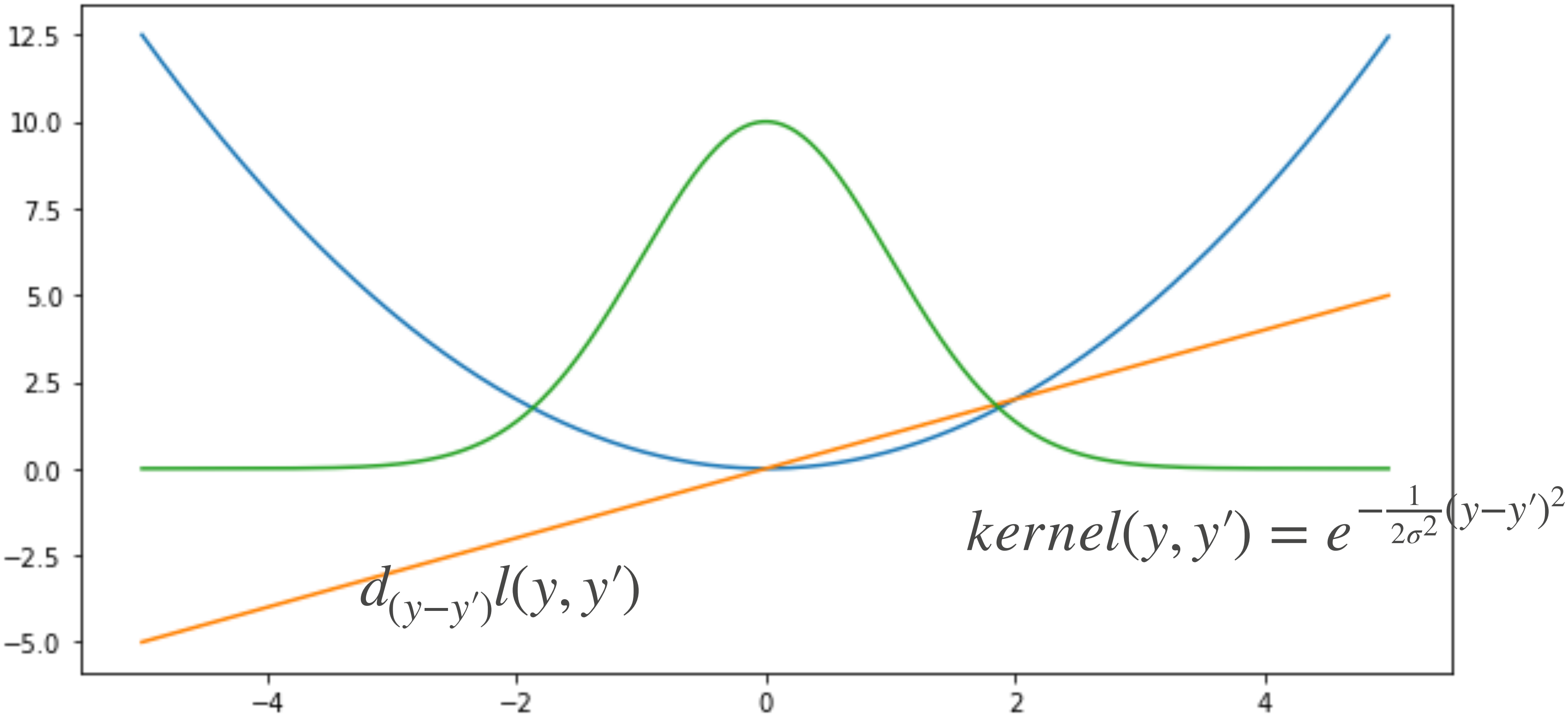
$$\Longleftrightarrow \underset{w}{\text{minimize}} \; \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i, w))^2 + \frac{\lambda}{2}\|w\|^2$$

Implement this

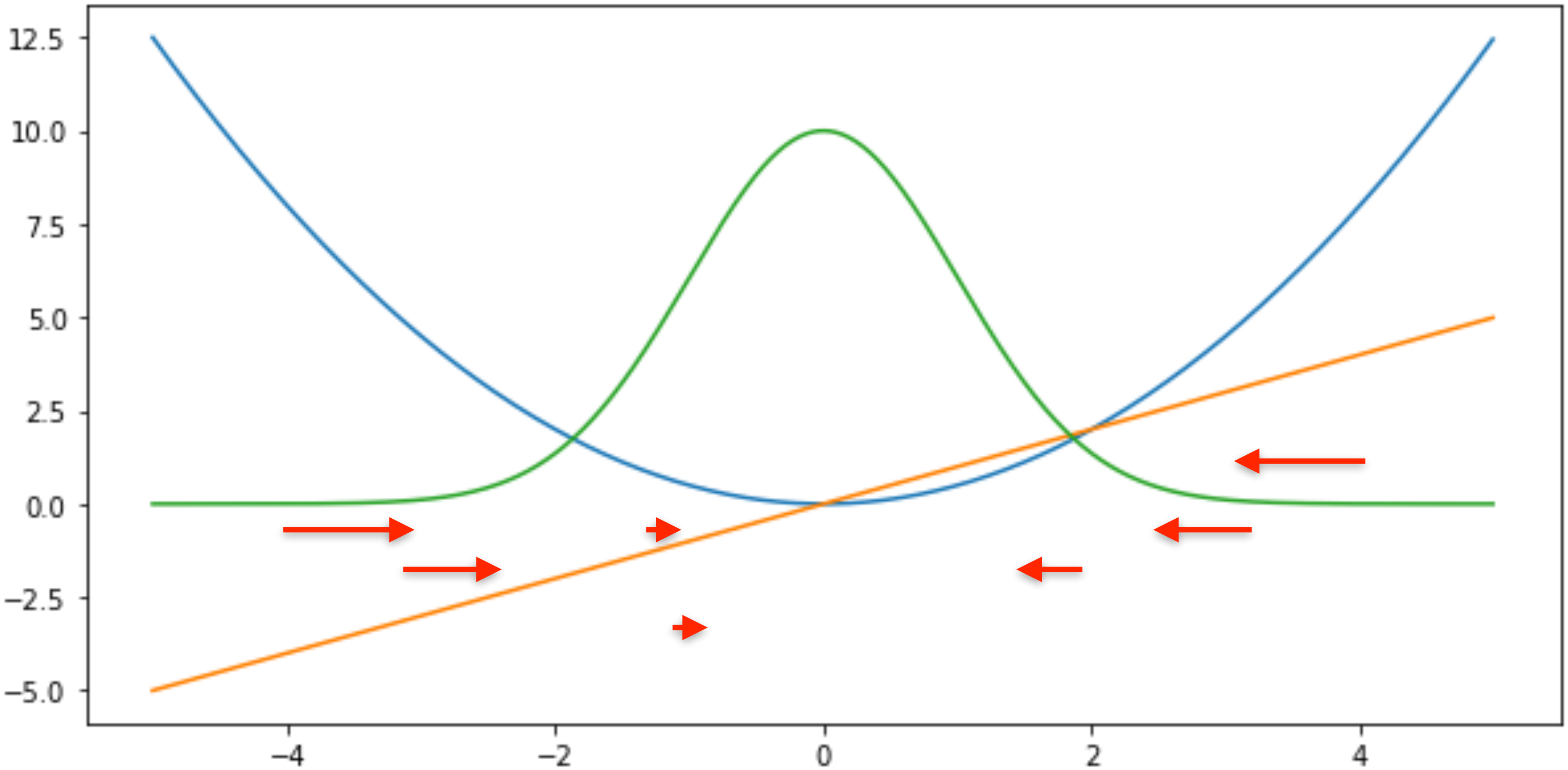
Loss Functions

L2 Loss

$$l(y, y') = \frac{1}{2}(y - y')^2$$

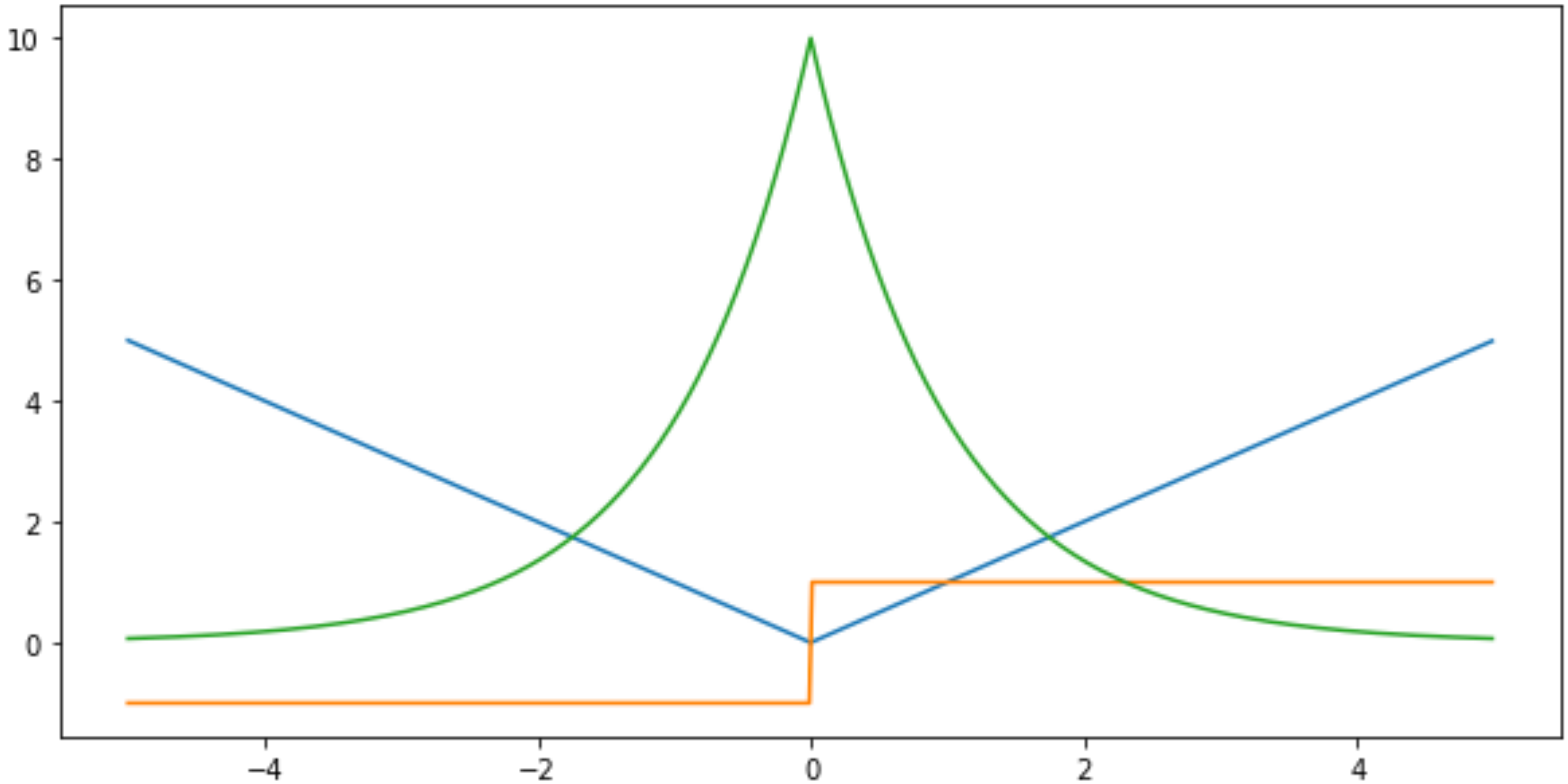


L2 Loss - mean $$l(y, y') = \frac{1}{2}(y - y')^2$$



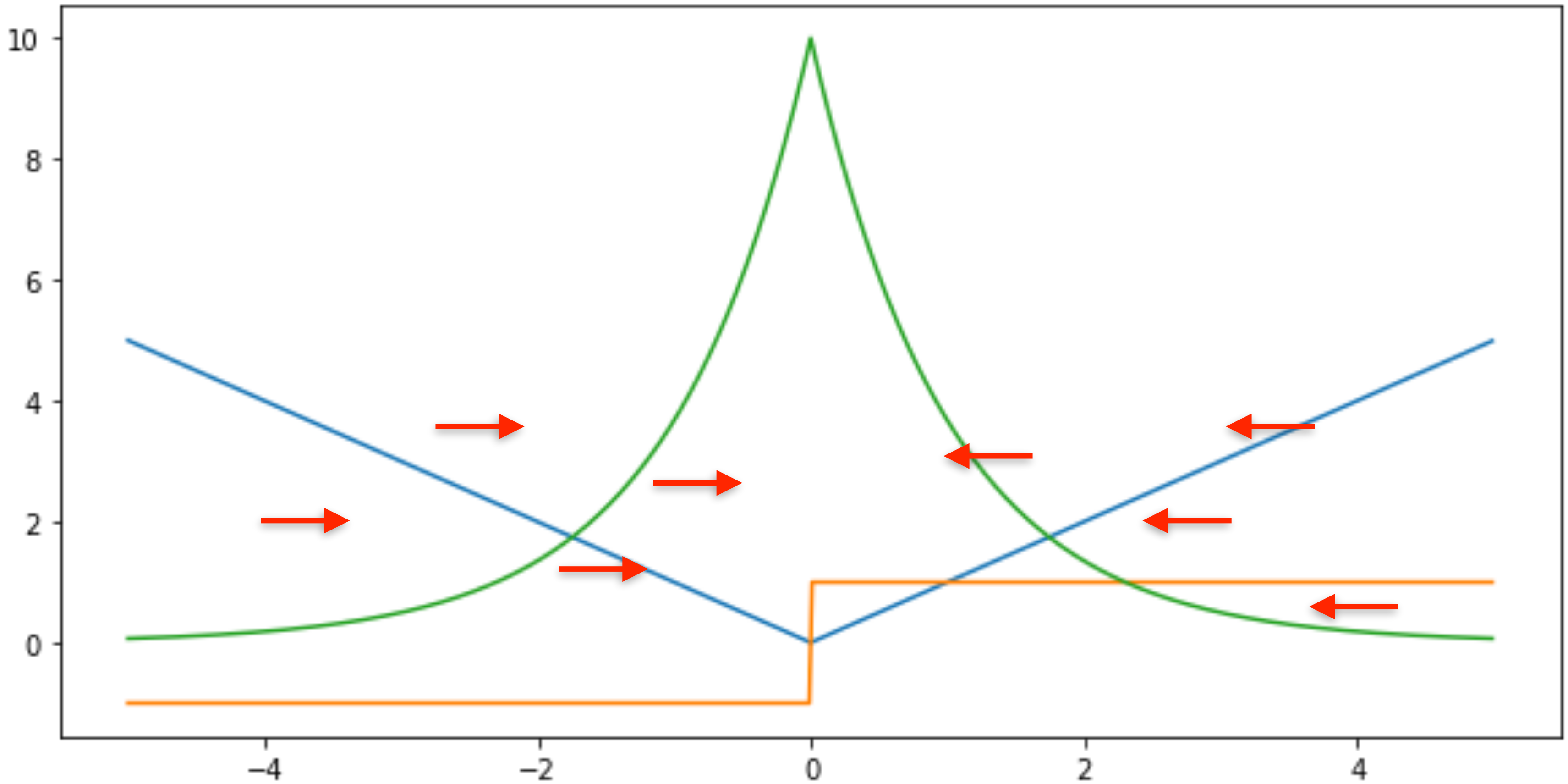
L1 Loss

$$l(y, y') = |y - y'|$$



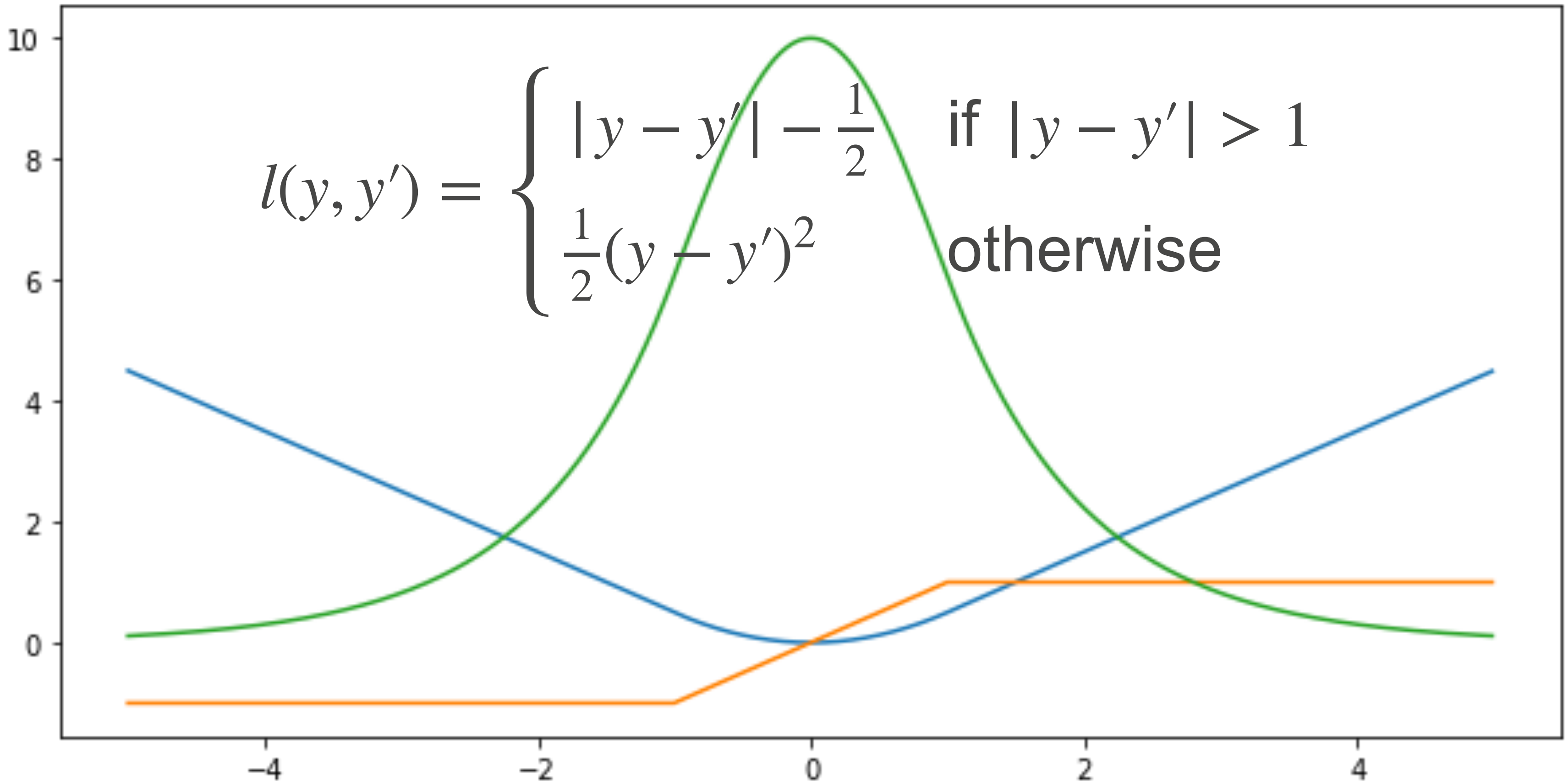
L1-Loss

$$l(y, y') = |y - y'|$$



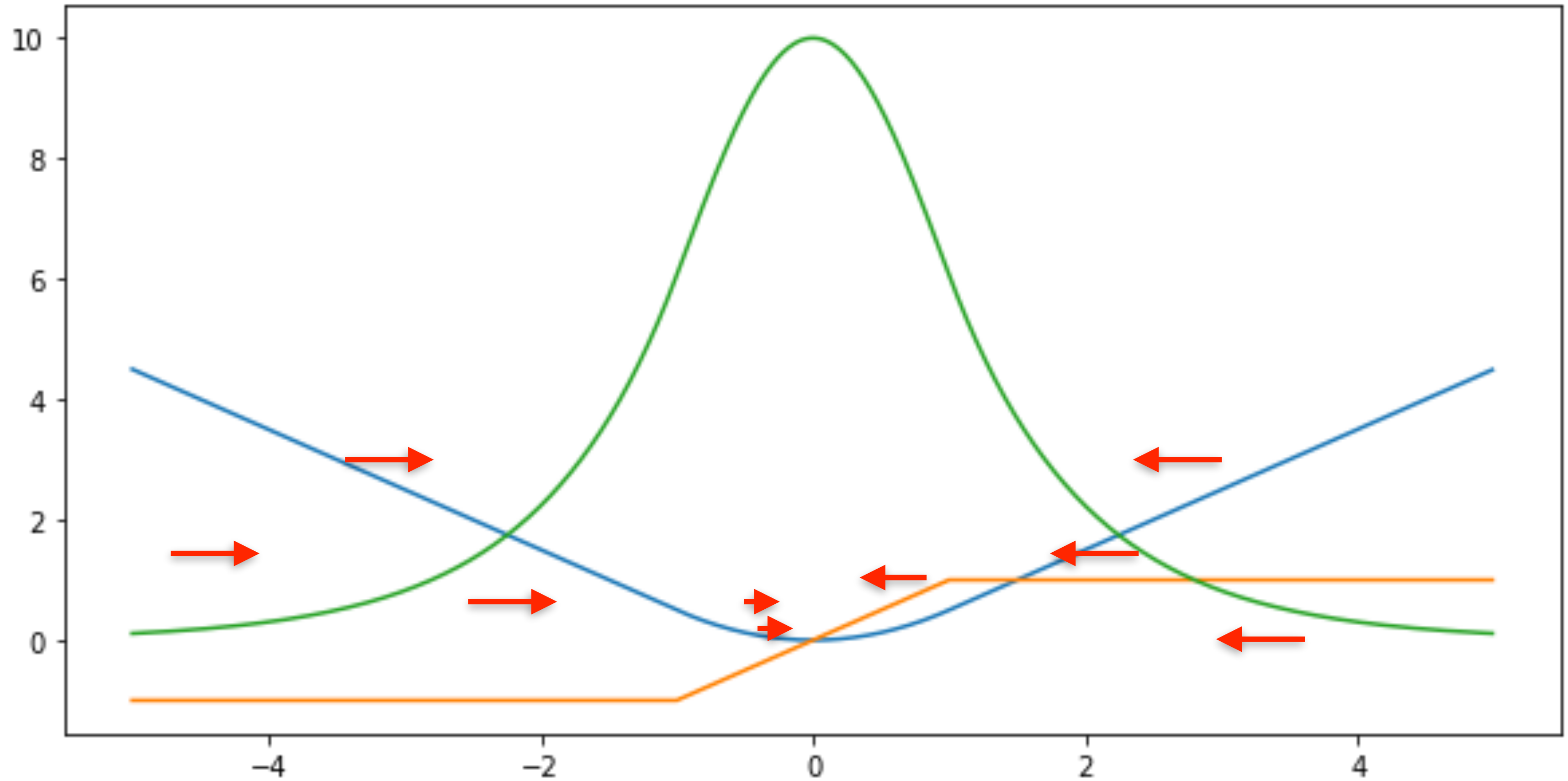
The same number of data on the left and right. The medium

Huber's Robust Loss



$$l(y, y') = \begin{cases} |y - y'| - \frac{1}{2} & \text{if } |y - y'| > 1 \\ \frac{1}{2}(y - y')^2 & \text{otherwise} \end{cases}$$

Huber's Robust Loss - trimmed mean

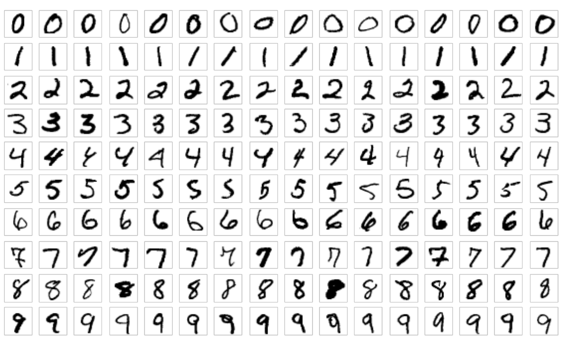



MAGIC Etch A Sketch® SCREEN
Logistic
Regression
OHIO ART
The World of Toys®
Horizontal Dial
Vertical Dial
MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
USE WITH CARE

Regression vs. Classification

- Regression estimates a continuous value
- Classification predicts a discrete category

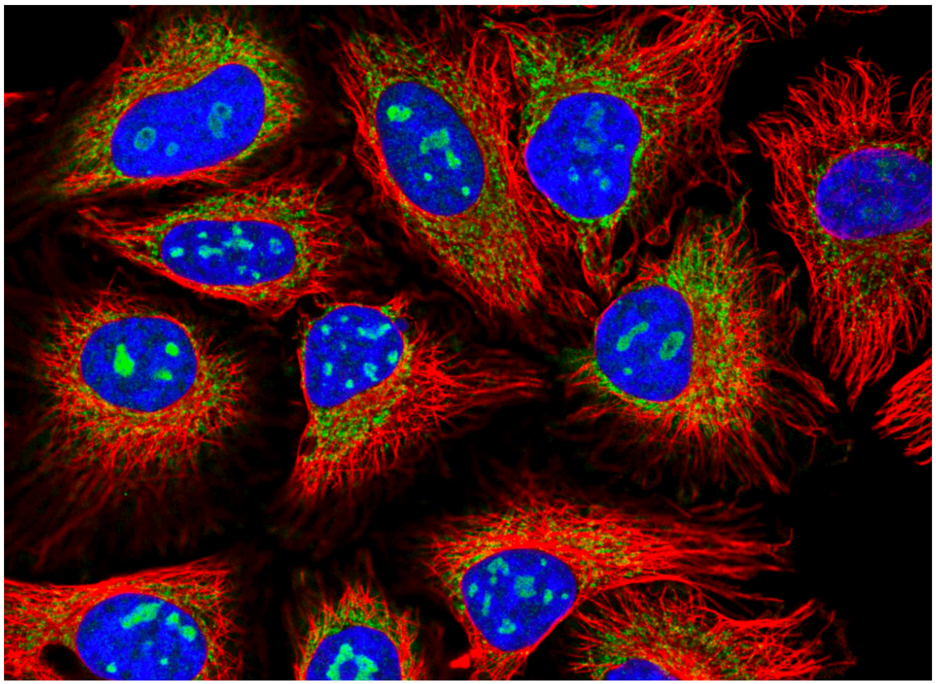
MNIST: classify hand-written digits (10 classes)



ImageNet: classify nature objects (1000 classes)

Classification Tasks at Kaggle

Classify human protein microscope images into 28 categories



```
0.  Nucleoplasm
1.  Nuclear membrane
2.  Nucleoli
3.  Nucleoli fibrillar
4.  Nuclear speckles
5.  Nuclear bodies
6.  Endoplasmic reticu
7.  Golgi apparatus
8.  Peroxisomes
9.  Endosomes
10.  Lysosomes
11.  Intermediate fila
12.  Actin filaments
13.  Focal adhesion si
14.  Microtubules
15.  Microtubule ends
16.  Cytokinetic bridg
```

https://www.kaggle.com/c/human-protein-atlas-image-classification

Classification Tasks at Kaggle

Classify malware into 9 categories

https://www.kaggle.com/c/malware-classification

Classification Tasks at Kaggle

Classify toxic Wikipedia comments into 7 categories

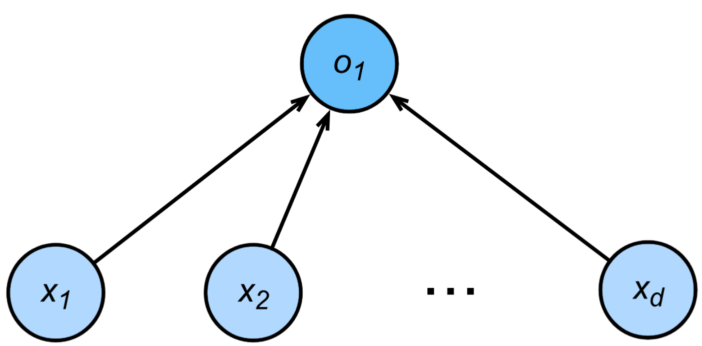
comment_text	toxic	severe_toxic	obs
Explanation\nWhy the edits made under my usern...	0	0	0
D'aww! He matches this background colour I'm s...	0	0	0
Hey man, I'm really not trying to edit war. It...	0	0	0
"\nMore\nI can't make any real suggestions on ...	0	0	0
You, sir, are my hero. Any chance you remember...	0	0	0

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

From Regression to Multi-class Classification

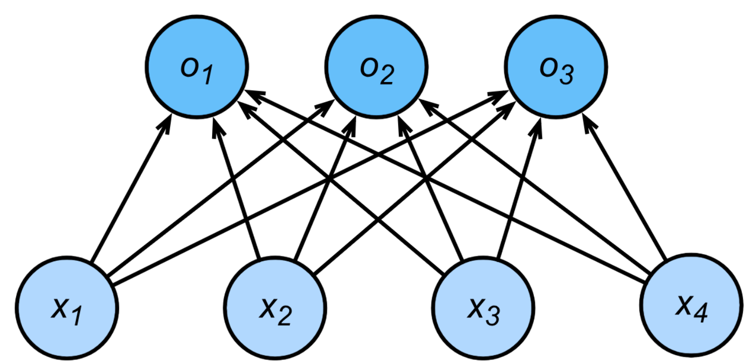
Regression

- Single continuous output
- Natural scale in $\mathbb{R}$
- Loss given e.g. in terms of difference $y - f(x)$



Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence …



From Regression to Multi-class Classification

Square Loss

- One hot encoding per class

$$\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$$

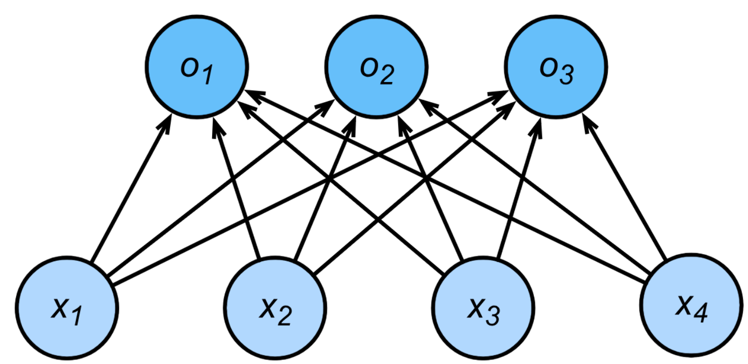
$$y_i = \begin{cases} 1 \text{ if } i = y \\ 0 \text{ otherwise} \end{cases}$$

- Train with squared loss
- Largest output wins

$$\hat{y} = \underset{i}{\operatorname{argmax}}\, o_i$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence …



From Regression to Multi-class Classification

Uncalibrated Scale

- One output per class
- Largest output wins

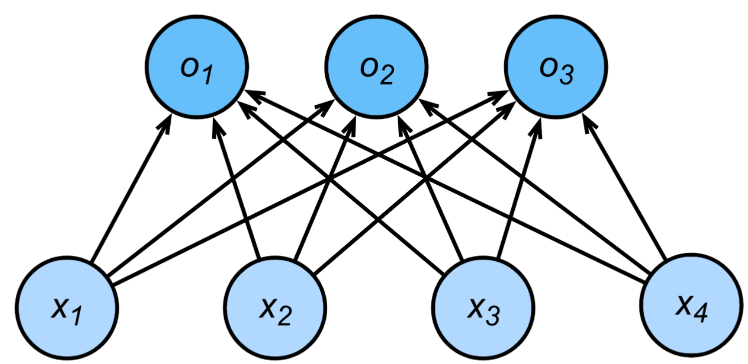
$$\hat{y} = \operatorname*{argmax}_i o_i$$

- Want that correct class is recognized confidently (**large margin**)

$$o_y - o_i \geq \Delta(y, i)$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence …



From Regression to Multi-class Classification

Calibrated Scale

- Output matches probabilities (nonnegative, sums to 1)

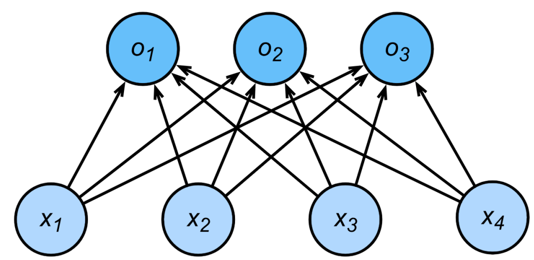
$$p(y \mid o) = \mathrm{softmax}(o) = \frac{\exp(o_y)}{\sum_i \exp(o_i)}$$

- Negative log-likelihood

$$-\log p(y \mid \mathbf{o}) = \log \sum_i \exp(o_i) - o_y$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence …



Softmax and Cross-Entropy Loss

- Negative log-likelihood (for given label y)

$$-\log p(y \mid o) = \log \sum_i \exp(o_i) - o_y$$

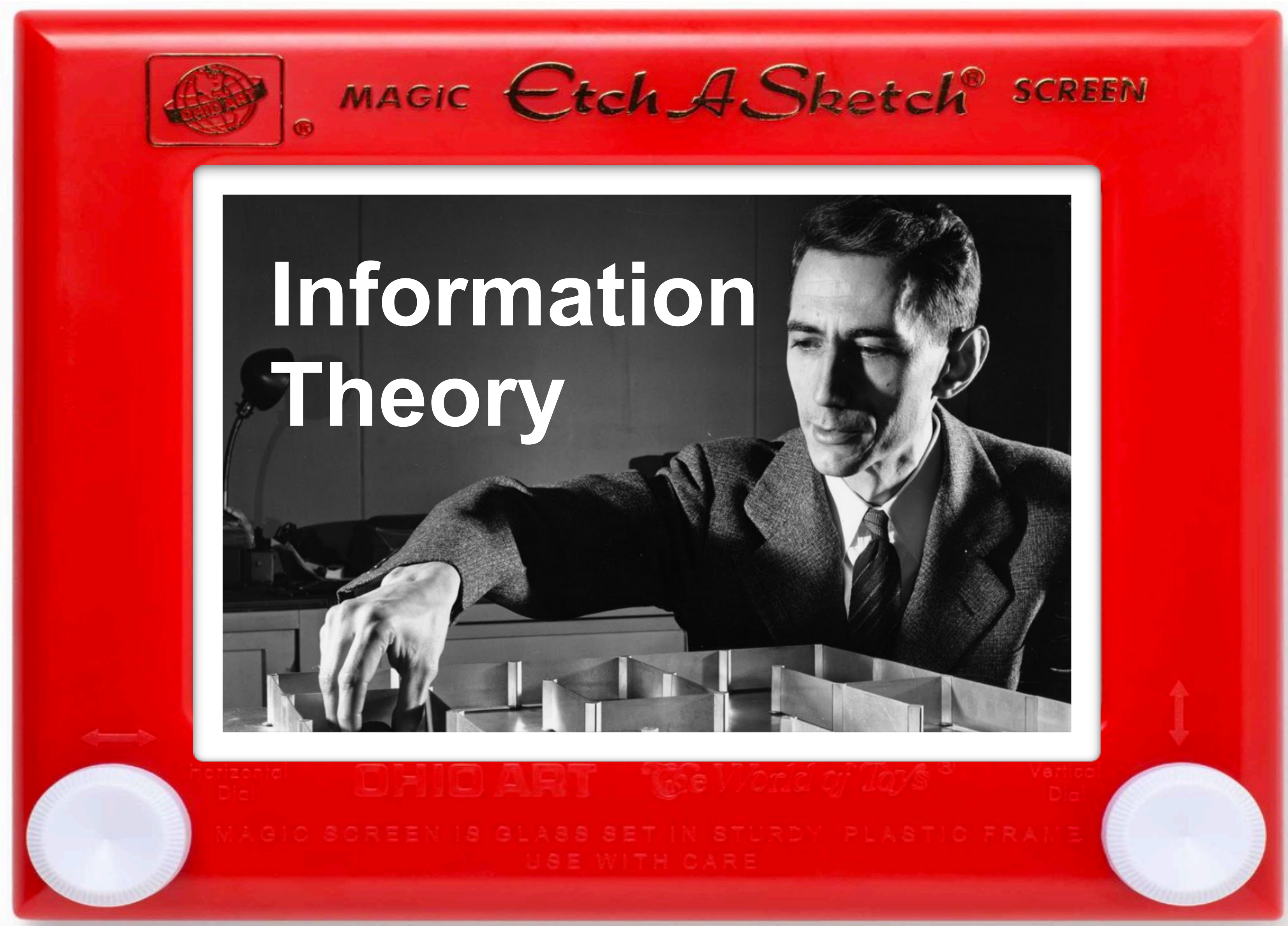
- Cross-Entropy Loss (for probability distribution y)

$$l(y, o) = \log \sum_i \exp(o_i) - y^\top o$$

- **Gradient**

Difference between true and estimated probability

$$\partial_o l(y, o) = \frac{\exp(o)}{\sum_i \exp(o_i)} - y^\top$$


MAGIC Etch A Sketch SCREEN
Information
Theory
OHIO ART
The World of Toys
MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME
USE WITH CARE

Entropy

- Data source producing observations $x_1 \ldots x_n$
- **How much 'information' is in this source?**
    - Tossing a fair coin - at each step the surprise is whether it's heads or tails
    - Rolling a fair dice - we have 1 out of 6 outcomes. This should be *more* surprising than the dice
    - Picture of a white wall vs. picture of a football stadium (the football stadium should have more information)
- **Measure is minimum number of bits needed**

Entropy

- Data source producing data $x_1 \ldots x_n$ with probability $p(x)$
- **Definition**

$$H[p] = -\sum_j p_j \log p_j$$

- **Coding theorem**
  Entropy is lower bound on bits (or rather nats - base e)

$$2^a = e^b \text{ hence } a \log 2 = b \text{ hence } \mathrm{bits} = \frac{H[p]}{\log 2}$$

- Entropy is concave since $p \log p$ is convex

$$H[\lambda p + (1-\lambda)q] \geq \lambda H[p] + (1-\lambda)H[q]$$

Entropy (binary form)

- Fair coin (p = 0.5)

$$H[p] = -0.5 \cdot \log_2 0.5 - 0.5 \cdot \log_2 0.5 = 1 \text{ bit}$$

- Biased coin (p = 0.9)

$$H[p] = -0.9 \cdot \log_2 0.9 - 0.1 \cdot \log_2 0.1 = 0.47 \text{ bit}$$

- Dungeons and Dragons (20-sided dice)

$$H[p] = -\log_2 \frac{1}{20} = 4.32 \text{ bit}$$

Kullback-Leibler Divergence

- Distance between distributions (e.g. truth & estimate)
  Number of extra bits when using the wrong code

$$D[p\|q] = \int dp(x) \log \frac{p(x)}{q(x)} = \int dp(x) \underbrace{\left[-\log q(x)\right]}_{\text{Inefficient bits}} - \underbrace{\left[-\log p(x)\right]}_{\text{Optimal bits}}$$

- Nonnegativity of KL Divergence

$$D[p\|p] = \int dp(x) \log \frac{p(x)}{p(x)} = 0$$

Jensen Inequality

$$D[p\|q] = -\int dp(x) \log \frac{q(x)}{p(x)} \geq -\log \int dp(x) \frac{q(x)}{p(x)} = 0$$

$$\psi(E[X]) \leq E[\psi(X)] (Jensen\ inequality)$$

Back to the Cross Entropy Loss

- Cross entropy loss

$$l(y, o) = \log \sum_i \exp(o_i) - y^\top o$$

- Kullback Leiber divergence between target q and prediction softmax (o)

$$D(q \| \mathrm{softmax}(o)) = \sum_i q_i \log q_i - q_i \log \mathrm{softmax}(o)_i$$

$$= -H[q] + \log \sum_i \exp(o_i) - \sum_i q_i o_i$$

Independent of o

Extended Reading

- Cover and Thomas (Elements of Information Theory)
  dl.acm.org/citation.cfm?id=1146355
- Information theory course (Entropy primer)
  slpl.cse.nsysu.edu.tw/cpchen/courses/ita/l1_entropy.pdf
- David MacKay (Information Theory and Learning)
  www.inference.org.uk/itprnn/book.html

- Conditional Entropy, Mutual Information,
  Exponential Families, Maximum Entropy Estimation

Summary

- **Maximum Likelihood**
  - Gauss and means
  - More loss functions (l_1 loss, trimmed mean)
  - Regression revisited
- **Classification**
  - Computing discrete probabilities
  - Likelihood and loss functions
- **Information Theory**