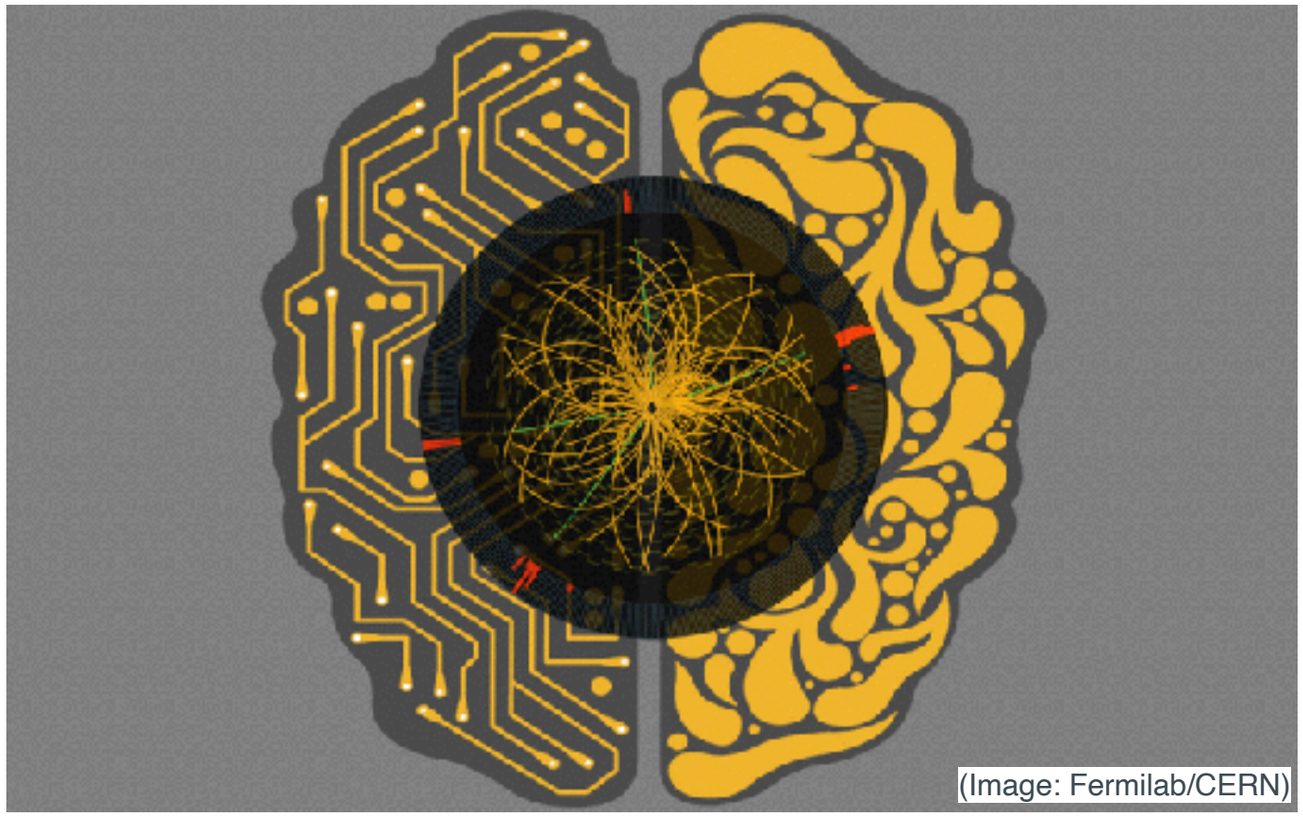
PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



Lecture 6: Information Content

Recap of Lecture 5

- Binary classification Logistic sigmoid

- Binary cross-entropy as loss function

- Multi-class classification

- 3 Examples: Phase classification 2D Ising, SUSY datasets, handwritten digits MNIST.

Multi-class Classification

Softmax Regression

- Going from 2 labels to M labels.
- Treat the label as a vector $\mathbf{y}_i \in \mathbb{Z}_2^M$, i.e., **one hot-encoding** of labels:

$$0 \to \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad 1 \to \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \quad M-1 \to \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

- The probability of being in class m' is the **softmax function**:

$$P(y_{im'} = 1 | \boldsymbol{x}_i, \{\mathbf{w}_k\}_{k=0}^{M-1}) = \frac{\mathrm{e}^{-\mathbf{x}_i^T \mathbf{w}_{m'}}}{\sum_{m=0}^{M-1} \mathrm{e}^{-\mathbf{x}_i^T \mathbf{w}_m}}$$

where $y_{im'} \equiv \{\mathbf{y}_i\}_{m'}$ is the m'-th component of vector $\mathbf{y}$.

Softmax Regression

- Likelihood of this M-class classifier:

$$P(\mathcal{D}|\{\mathbf{w}_k\}_{k=0}^{M-1}) = \prod_{i=1}^{n}\prod_{m=0}^{M-1}[P(y_{im}=1|\boldsymbol{x}_i,\mathbf{w}_m)]^{y_{im}} \times [1-P(y_{im}=1|\boldsymbol{x}_i,\mathbf{w}_m)]^{1-y_{im}}$$

- Cost function:

$$\mathcal{C}(\mathbf{w}) = -\sum_{i=1}^{n}\sum_{m=0}^{M-1} y_{im}\log P(y_{im}=1|\boldsymbol{x}_i,\mathbf{w}_m) + (1-y_{im})\log\left(1-P(y_{im}=1|\boldsymbol{x}_i,\mathbf{w}_m)\right).$$

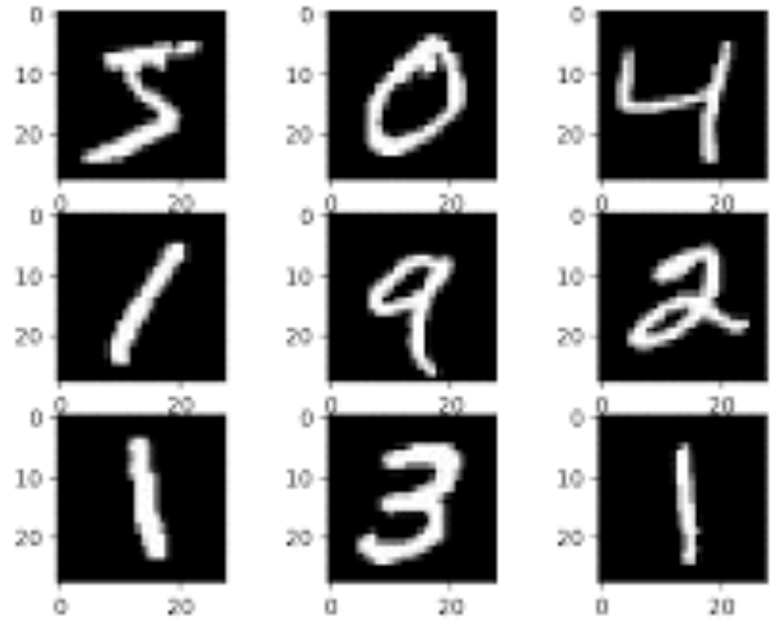
- Activations (one-hot encoding), can be thought of as activating particular cells (e.g. in your brain).
- Becomes a lot harder for a larger number of classes.

Classifying Digits — MNIST

MNIST

- Classifying digits ⇒ M=10 categories

- MNIST = Dataset of handwritten digits, 28x28=784 pixel grid, each assumes 256 grayscale values, interpolating between white and black.



Yann LeCun, Corinna Cortes, Christopher Burges

http://yann.lecun.com/exdb/mnist/

- 60,000 images: 50,000 for training, 10,000 for testing

Experiment with Notebook 7 using softmax:

https://physics.bu.edu/%7Epankajm/MLnotebooks.html

Outline for today

- Information Content
- Shannon Entropy
- Kullback-Leibler Divergence
- Capacity of Perceptron

Reference: MacKay Chapter 2, 39, 40; 1805.11965

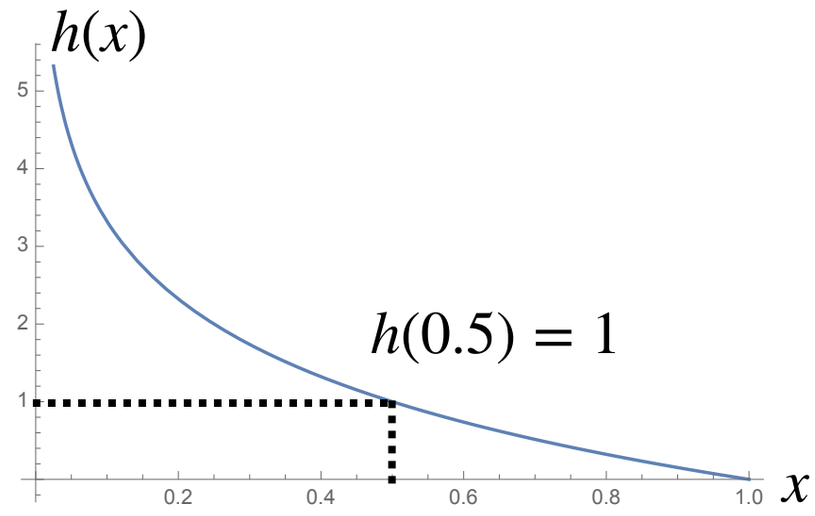
Quantifying Information

- Consider the following two sentences:
  - I am taking three physics courses.
  - I am taking Physics 835, 910, 922
- Which sentence contains more information?
- The amount of information of event A = $-\log P(\text{event } A)$
- When the probability is low, the amount of information is large.

Shannon Entropy

- **Shannon information content** of an outcome x (in units of **"bits"**):

$$h(x) = \log_2 \frac{1}{P(x)}.$$



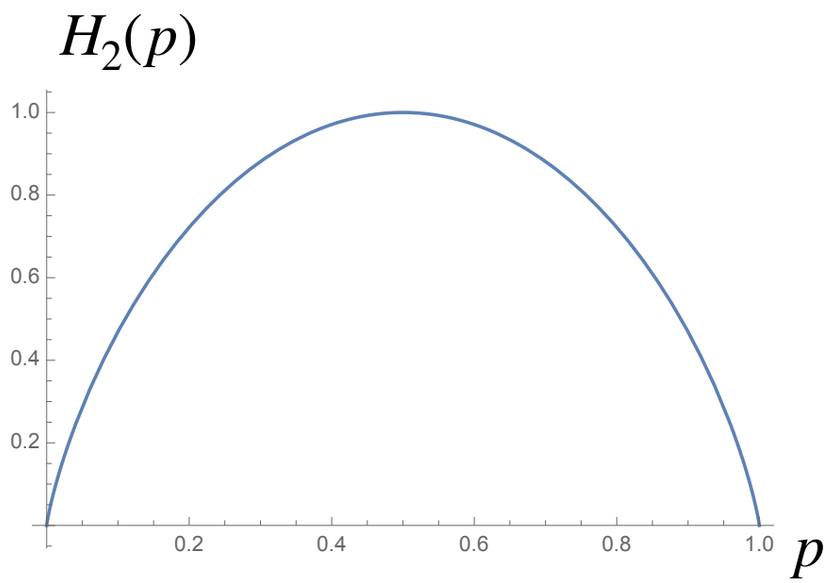
- **Shannon entropy** (X's average info content)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}, \qquad X = \text{ensemble}$$

- $P(x) = 0$ gives no contribution to $H(X)$ due to L'Hospital's Rule.

- Consider an ensemble with 2 outcomes:

$$H_2(p) = H(p, 1-p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$



No info conveyed at p=0, 1

Information Content

- We can extrapolate from this 2-outcome example these properties:
  - $H(X) \geq 0$ with equality iff $p_i = 1$ for one i.
  - Entropy is maximized if $\mathbf{p}$ is uniform:

$$H(X) \leq \log(|\mathcal{A}_X|) \quad \text{with equality iff } p_i = 1/|\mathcal{A}_X| \text{ for all } i.$$

- The following slogan may be useful:
  - Large information entropy ⇔ difficult to predict
  - Small information entropy ⇔ easy to predict

Joint Entropy

- The joint entropy of X, Y is:

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}.$$

- Entropy is additive for independent random variables:

$$H(X,Y) = H(X) + H(Y) \text{ iff } P(x,y) = P(x)P(y).$$

- Our definitions for info content so far apply only to discrete probability distribution over finite sets $\mathcal{A}_X$.

- Definitions can be extended to **infinite sets** (entropy may be infinite), or probability density over a **continuous set** (MacKay Chapter 11).

Relative Entropy or KL Divergence

- Let us start by motivating the notion of relative entropy for finite sets $\mathscr{A}_X$: possible events are $A_1, A_2, \ldots, A_W$ with probability $p_1, p_2, \ldots p_W$.

- In practice, p_i is not known. Instead we only know how many times A_i has occurred.

- $\begin{cases} A_1 : \#_1 \text{ times}, A_2 : \#_2 \text{ times}, \ldots, A_W : \#_W \text{ times}, \\ \#(= \sum_{i=1}^{W} \#_i) \text{ times in total.} \end{cases}$

- **Task:** find an expected probability q_i that is as close as possible to p_i in producing the above observed outcomes ⇒ **maximum likelihood estimation**

- Suppose each A_i occurs with probability q_i:

$$p(\text{probability of } A_i \text{ occurring } \#_i \text{ times}) = q_i^{\#i}$$

Relative Entropy or KL Divergence

- The A_i's can occur in any order, e.g., $[A_1, A_1, A_2]$ or $[A_2, A_1, A_1]$. To account for the combinatorics, we multiply the probability by:

$$\binom{\#}{\#_1, \#_2, \ldots, \#_W} = \frac{\#!}{\#_1!\#_2! \cdots \#_W!}$$

- The probability of observing $(\cdot)$ is

$$q_1^{\#_1} q_2^{\#_2} \ldots q_W^{\#_W} \frac{\#!}{\#_1!\#_2! \ldots \#_W!}$$

- **Task:** find q_i's that maximize this probability.

- This problem can be solved by the Lagrange multiplier method:

$$L(q_i, \lambda) = \log \left[q_1^{\#_1} q_2^{\#_2} \ldots q_W^{\#_W} \frac{\#!}{\#_1!\#_2! \ldots \#_W!} \right] + \lambda \left(1 - \sum_{i=1}^{W} q_i \right)$$

Relative Entropy or KL Divergence

- Extremizing the Lagrangian:

$$q_i = \frac{\#_i}{\#}, \quad \lambda = \#.$$

- In the limit of large dataset ($\# \to \infty$), the probability is maximized when $q_i = p_i$.

$$\frac{\#_i}{\#} \approx p_i \quad \Leftrightarrow \quad \#_i \approx \# \cdot p_i \, .$$

- Even though we know what q_i maximize the probability, let's keep going to find our loss function.
- Since the individual $\#_i$'s must also be large in this limit unless $p_i = 0$, using the Stirling's formula:

$$\#_i ! \approx \#_i^{\#_i}$$

Relative Entropy or KL Divergence

- Maximizing the probability amounts to maximizing:

$$
\begin{aligned}
q_1^{\#_1} q_2^{\#_2} \dots q_W^{\#_W} \frac{\#!}{\#_1!\#_2!\dots\#_W!} &\approx q_1^{\#\cdot p_1} q_2^{\#\cdot p_2} \dots q_W^{\#\cdot p_W} \frac{\#!}{(\#\cdot p_1)!(\#\cdot p_2)!\dots(\#\cdot p_W)!} \\
&\approx q_1^{\#\cdot p_1} q_2^{\#\cdot p_2} \dots q_W^{\#\cdot p_W} \frac{\#^{\#}}{(\#\cdot p_1)^{\#\cdot p_1}(\#\cdot p_2)^{\#\cdot p_2}\dots(\#\cdot p_W)^{\#\cdot p_W}} \\
&= q_1^{\#\cdot p_1} q_2^{\#\cdot p_2} \dots q_W^{\#\cdot p_W} \frac{1}{p_1^{\#\cdot p_1} p_2^{\#\cdot p_2} \dots p_W^{\#\cdot p_W}} \\
&= \exp\left[-\# \sum_{i=1}^{W} p_i \log \frac{p_i}{q_i} \right].
\end{aligned}
$$

- Goal: to make this probability as close to 1 as possible, i.e., make

$$
D_{KL}(P\,||\,Q) = \sum_{i=1}^{W} p_i \log \frac{p_i}{q_i} \qquad \text{where } P = \{p_1, p_2, \dots, p_W\}, Q = \{q_1, q_2, \dots q_W\}
$$

as close to 0 as possible. This quantity is known as the **relative entropy** (or **Kullback-Leibler divergence** in information theory). This is our loss function.

Gibbs' Inequality

- $D_{KL}(P||Q)$ is also known as the **KL distance** even though it is not strictly a distance because in general:

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

- Relative entropy (KL divergence) can be similarly defined for continuous probability distributions $P(x), Q(x)$:

$$D_{KL}(P||Q) = \int dx \; P(x) \; \log \frac{P(x)}{Q(x)}$$

- Moreover, the relative entropy satisfies the Gibbs' inequality:

$$D_{KL}(P||Q) \geq 0$$

- Relative entropy is important in pattern recognition, design of neutral network loss functions, and information theory.

Proving Gibbs' Inequality

- First prove the Gibbs' inequality for finite sets $\mathscr{A}_X$ & using natural log; the later assumption can be relaxed by scaling relationships below.

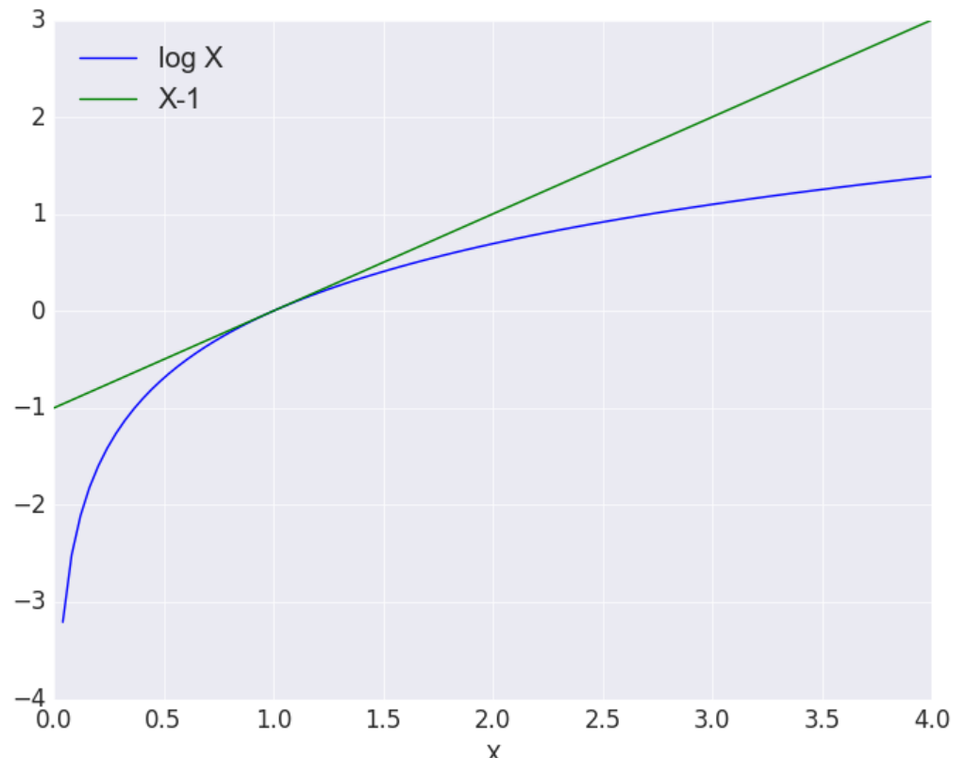
- Let I denotes the set of all i for which $p_i \neq 0$:

$$-\sum_{i\in I} p_i \ln\frac{q_i}{p_i} \geq -\sum_{i\in I} p_i\left(\frac{q_i}{p_i}-1\right) = -\sum_{i\in I} q_i + \sum_{i\in I} p_i = -\sum_{i\in I} q_i + 1 \geq 0$$

- The 1st inequality follows from $\ln x \leq x-1,\ \forall x > 0$ which is saturated only for $x = 1$.

- The 2nd inequality is a consequence of P and Q being probability distributions.

$$\sum_{i\in I} p_i = 1; \qquad \sum_{i\in I} q_i \leq \sum_{i}^{W} q_i = 1$$



Proving Gibbs' Inequality

- We have shown so far that: $-\sum_{i\in I} p_i \ln q_i \geq -\sum_{i\in I} p_i \ln p_i$
- Both sums can be extended to all $i = 1,2,\ldots, W$ by noting that:

$$p_i \ln p_i \to 0 \text{ as } p_i \to 0 \qquad \ln q_i \to \infty \text{ as } q_i \to 0$$

- We arrive at the Gibbs' inequality: $-\sum_{i=1}^{W} p_i \ln q_i \geq -\sum_{i=1}^{W} p_i \ln p_i$
- For the equality to hold, both conditions must be satisfied:
  - $\frac{q_i}{p_i} = 1 \ \forall i \in I$ in order for $\ln \frac{q_i}{p_i} = \frac{q_i}{p_i} - 1$ to hold
  - $\sum_{i\in I} q_i = 1$ which means $q_i = 0 \text{ if } p_i = 0$

} $p_i = q_i \ \forall i = 1,2,\ldots W$

Example: Gaussian Distribution

- Consider two Gaussian distributions:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{1}{2\sigma_p^2}(x-\mu_p)^2},$$

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma_q} e^{-\frac{1}{2\sigma_q^2}(x-\mu_q)^2}.$$

- The relative entropy:

$$\begin{aligned} D_{KL}(p||q) &= \int_{-\infty}^{\infty} dx\ p(x)\left(\log\frac{\sigma_q}{\sigma_p} - \frac{1}{2\sigma_p^2}(x-\mu_p)^2 + \frac{1}{2\sigma_q^2}(x-\mu_q)^2\right) \\ &= \log\frac{\sigma_q}{\sigma_p} - \frac{1}{2\sigma_p^2}(\sigma_p)^2 + \frac{1}{2\sigma_q^2}\int_{-\infty}^{\infty} dx\ p(x)\underbrace{(x-\mu_q)^2}_{(\mu_p-\mu_q)^2+2(\mu_p-\mu_q)(x-\mu_p)+(x-\mu_p)^2} \\ &= \log\frac{\sigma_q}{\sigma_p} - \frac{1}{2\sigma_p^2}(\sigma_p)^2 + \frac{1}{2\sigma_q^2}\left((\mu_p-\mu_q)^2 + 0 + \sigma_p^2\right) \\ &= \frac{1}{2}\left(-\log\frac{\sigma_p^2}{\sigma_q^2} + \left(\frac{\sigma_p^2}{\sigma_q^2} - 1\right) + \frac{1}{\sigma_q^2}(\mu_p-\mu_q)^2\right) \end{aligned}$$

Example: Gaussian Distribution

- Relative entropy as the “distance” between probability distributions.
- Consider two “nearby points” in the (μ, σ) space:

$$\sigma_p = \sigma\,, \quad \mu_p = \mu\,, \qquad\qquad \sigma_q = \sigma + d\sigma\,, \quad \mu_q = \mu + d\mu\,.$$

- Up to second order in $d\sigma, d\mu$:

$$\begin{aligned} D_{KL}(p||q) &= \frac{1}{2}\left(-\log\frac{\sigma^2}{(\sigma+d\sigma)^2} + \left(\frac{\sigma^2}{(\sigma+d\sigma)^2} - 1\right) + \frac{1}{(\sigma+d\sigma)^2}(\mu-\mu-d\mu)^2\right) \\ &= \frac{1}{2}\left(\log(1+\frac{d\sigma}{\sigma})^2 + \left(\frac{1}{(1+\frac{d\sigma}{\sigma})^2} - 1\right) + \frac{1}{\sigma^2(1+\frac{d\sigma}{\sigma})^2}d\mu^2\right) \\ &\approx \frac{1}{2}\left(2\frac{d\sigma}{\sigma} - \frac{d\sigma^2}{\sigma^2} + \left(1 - 2\frac{d\sigma}{\sigma} + 3\frac{d\sigma^2}{\sigma^2} - 1\right) + \frac{d\mu^2}{\sigma^2}\right) \\ &= \frac{1}{2}\left(2\frac{d\sigma^2}{\sigma^2} + \frac{d\mu^2}{\sigma^2}\right) = \frac{d\sigma^2 + d\tilde{\mu}^2}{\sigma^2}. \qquad \text{where } \tilde{\mu} = \mu/\sqrt{2} \end{aligned}$$

- Metric of a hyperboloid, a part of AdS spacetime. AdS metric diverges at its infinite boundary ($\sigma = 0$). A curiosity or maybe more? https://arxiv.org/abs/2001.02683

Summary

- Multi-class classification, e.g., handwritten digits MNIST.
- Information Content
- Shannon entropy (relative entropy, KL divergence, KL distance)
- Gibbs' inequality