Machine Learning 2

Conditional Independence

Patrick Forré

Recap: Marginals & Conditionals

- Let X, Y, Z be random variables with joint distribution $p(x, y, z)$.
- Marginal distribution of Z:

$$p(z) = \int p(x, y, z)\, dx\, dy$$

- (Marginal) joint distribution of X and Y given $Z = z$:

$$p(x, y \mid z) = \begin{cases} \dfrac{p(x, y, z)}{p(z)}, & p(z) > 0 \\ 0, & p(z) = 0 \end{cases}$$

Recap: Independence

- Let X, Y be random variables with joint distribution $p(x, y)$.
  We say that X is independent of Y iff:

$$\forall x, y, p(x, y) = p(x) \cdot p(y)$$

- In symbols: $X \perp\!\!\!\perp Y$

Conditional Independence - Definition

- Let X, Y, Z be random variables with joint distribution $p(x, y, z)$.
  We say that X is independent of Y conditioned on Z iff:

$$\forall x, y, z: \; p(x, y \mid z) = p(x \mid z) \cdot p(y \mid z)$$

  (actually, we only need to check that for z with $p(z) > 0$).

- In symbols: $X \perp\!\!\!\perp Y \mid Z$

Conditional Independence - Remarks

- We consider (unconditional) independence as a special case of conditional independence (as we need to say what we condition on anyways):

$$X \perp\!\!\!\perp Y \iff X \perp\!\!\!\perp Y \mid \varnothing \iff X \perp\!\!\!\perp Y \mid \text{(const. RV)}$$

- Independence heuristically means that knowing Y does not give us any information about X.
- Conditional Independence of X and Y given Z means that there is no additional information in the state of Y beyond the information already in Z to determine the state of X.

Conditional Independence - Example 1

- Alice and Bob each independently throw a coin.
- They try to guess what the other person has thrown.
- Is that possible by just looking at their own coin?

$A \perp B$

- Eve just happens to see both coin outcomes and tells Bob: “I don’t tell you what Alice has, the outcomes are different anyways.”
- What about now?

$A \not\perp B \mid E$

Conditional Independence - Example 2

- Alice, Bob and Casper each throw a coin.
- Alice tries to guess the number of ‘tails’ T.
- Alice has a ‘head’, what can she tell?

A ~~H~~ T

- Eve just saw that Alice and Bob have 2 ‘heads’ together.
- Can Alice now tell if there now 2 or 3 ‘tails’ in total?

A H T I E

Exercise: Separoid Axioms

- Prove the following claims for random variables X, Y, Z, W:
- Redundancy: $X \perp\!\!\!\perp Y | X$ always holds.
- Symmetry: $X \perp\!\!\!\perp Y | Z \iff Y \perp\!\!\!\perp X | Z$
- Decomposition: $X \perp\!\!\!\perp (Y, W) | Z \implies X \perp\!\!\!\perp Y | Z$
- Weak Union: $X \perp\!\!\!\perp (Y, W) | Z \implies X \perp\!\!\!\perp W | (Y, Z)$
- Contraction: $X \perp\!\!\!\perp W | (Y, Z) \wedge X \perp\!\!\!\perp Y | Z \implies X \perp\!\!\!\perp (Y, W) | Z$
- So we have the equivalence:

$$X \perp\!\!\!\perp (Y, W) | Z \iff X \perp\!\!\!\perp W | (Y, Z) \wedge X \perp\!\!\!\perp Y | Z$$

Summary

- Conditional Independence: $X \perp\!\!\!\perp Y \mid Z$ defined to hold if:

$$\forall x, y, z: \; p(x, y \mid z) = p(x \mid z) \cdot p(y \mid z)$$

- Example 1 shows: $X \perp\!\!\!\perp Y \not\Longrightarrow X \perp\!\!\!\perp Y \mid Z$

- Example 2 shows: $X \perp\!\!\!\perp Y \mid Z \not\Longrightarrow X \perp\!\!\!\perp Y$

Machine Learning 2

Information Theory
- Entropy

Patrick Forré

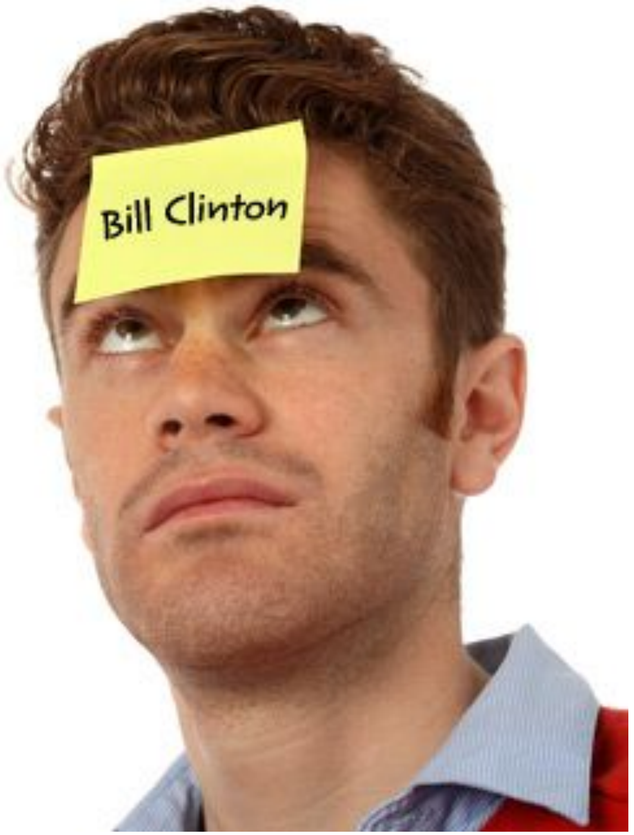
Let's Play: "Who am I?"

- Find out who you represent in a minimal number of YES/NO-questions.

- What is the minimal number of YES/NO-questions required averaged over all possible characters?

→ Entropy

- need to know whole space of possible characters (with frequencies)



Entropy - Informal Definition

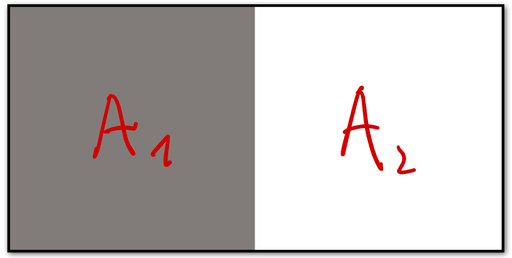
- Let X be a random variable.

- Entropy of X := Minimal bits (=YES/NO-questions) needed to encode state X averaged over all possible states that X can take weighted their probability/frequency.

- $$H(X) := \mathbb{E}[\#\text{bits}(X)]$$

Entropy - Examples

- Alice selects field.
- How many YES/NO-questions does Bob need to ask to find that field?



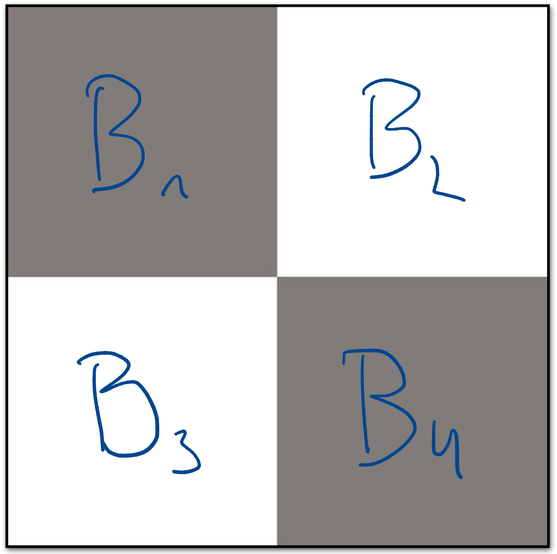
$\#bits(A_1) = 1$

$p(A_1) = \frac{1}{2}$

$\#bits(A_2) = 1$

$p(A_2) = \frac{1}{2}$

$H(A) = 1$



$\#bits(B_i) = 2$

$p(B_i) = \frac{1}{4} = \frac{1}{2^2}$

$$H(B) = \mathbb{E}[\#bits(B)]$$

$$= \sum_{i=1}^{4} p(B_i) \cdot \#bits(B_i)$$

$$= \sum_{i=1}^{4} p(B_i) \cdot 2$$

$$= 2$$

Entropy - Examples

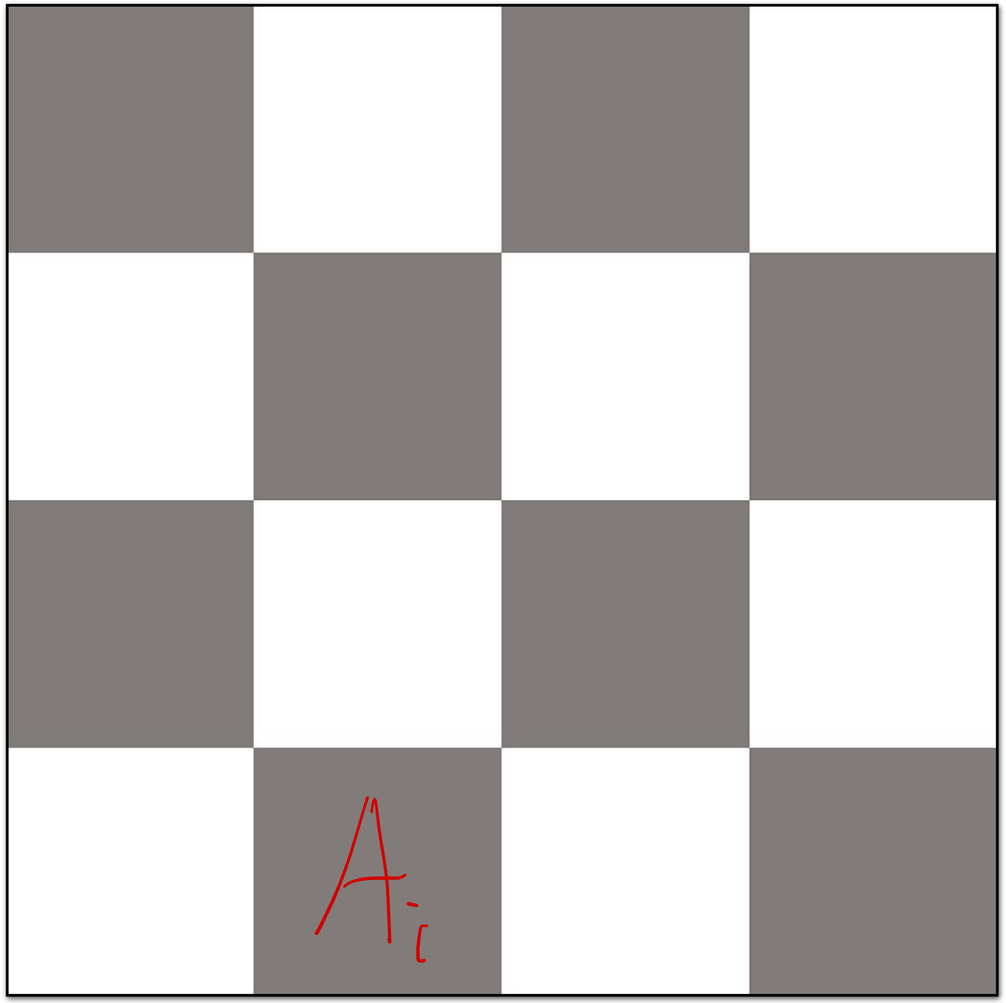
- Alice selects field.
- How many YES/NO-questions does Bob need to ask to find that field?

$\#bits(A_i) = 4$, $H(A) = 4$

$p(A_i) = \frac{1}{2^4}$

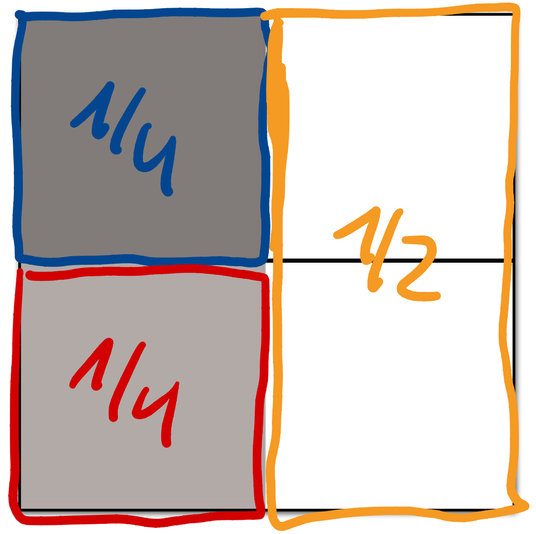
$p(A_i) = \frac{1}{2^{\#bits}}$

$\rightsquigarrow$ $\boxed{\#bits(A_i) = -\log_2 p(A_i)}$



Entropy - Examples

- Alice selects field.
- How many YES/NO-questions does Bob need to ask to find that field?



$\#bits(A_1) = 2$, $p(A_1) = \frac{1}{4} = \frac{1}{2^{\#bits}}$

$\#bits(A_2) = 2$, $p(A_2) = \frac{1}{4} = \frac{1}{2^{\#bits}}$

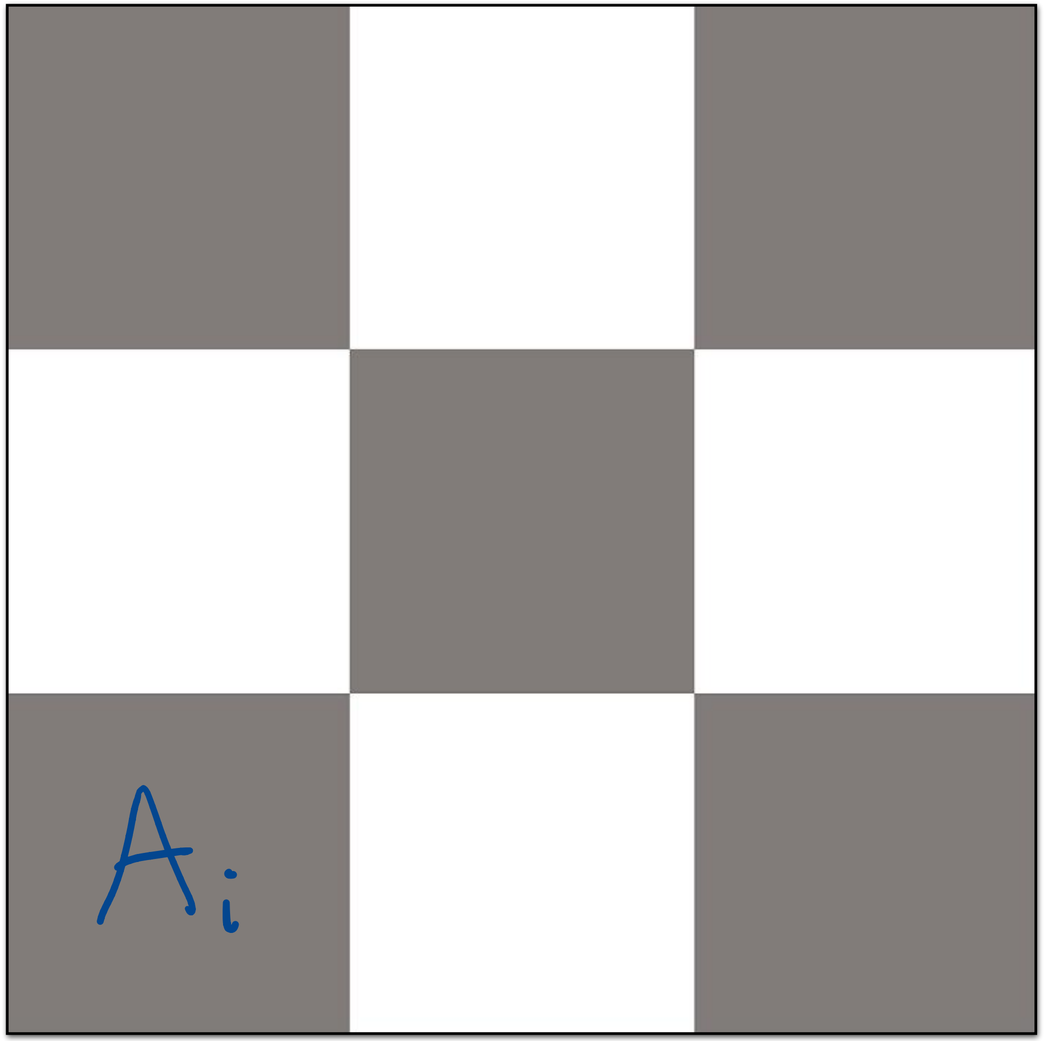
$\#bits(A_3) = 1$, $p(A_3) = \frac{1}{2} = \frac{1}{2^{\#bits}}$

$$H(A) = \sum_{i=1}^{3} p(A_i) \cdot \#bits(A_i) = \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 1$$

$$= 3/2 \quad \in [1, 2]$$

Entropy - Examples

- Alice selects field.
- How many YES/NO-questions does Bob need to ask to find that field?



$p(A_i) = \frac{1}{9}$

$$\#\text{bits}(A_i) = -\log_2 p(A_i)$$

Fano - Shannon formula / code

$$H(A) = \mathbb{E}[\#\text{bits}(A)] = \mathbb{E}[-\log p(A)]$$
$$= -\log_2 \tfrac{1}{9} = \log_2 9 \approx 3.17$$

Entropy - Definition

- Let X be a random variable with distribution $p(x)$.

Shannon Entropy
(X discrete)

- Entropy of X: $H(X) := \mathbb{E}[-\log p(X)]$ — Differential Entropy (X continuous)

- The choice of the basis of $\log$ does not really matter, but one should be consistent with it. Usually one uses basis 2 to measure 'bits' or basis e to measure 'nats'.

Entropy - Remarks

- The Entropy measures spread/uncertainty in the distribution similar to the Variance, but does not need a center of mass $\mathbb{E}[X]$ as a reference point like the Variance does.
- Entropy is also meaningful in the multimodal case.

→ high Variance
→ low Entropy

- $H(X)$ small -> X concentrated in small area.
- $H(X)$ big -> X is scattered around.
- $H(X) \geq 0$ for discrete variables, but not in general.

Entropy - Exponential Families

- $$p(x \mid \eta) = h(x) \cdot \exp\left(\eta^{\top} T(x) - A(\eta)\right)$$

$$H(X) = \mathbb{E}_{\eta}[-\log_e p(x|\eta)] = \mathbb{E}\left[-\log_e h(x) - (\eta^T T(x) - A(\eta))\right]$$

$$= -\mathbb{E}[\log_e h(x)] - \eta^T \underbrace{\mathbb{E}[T(x)]}_{\nabla_\eta A(\eta)} + A(\eta)$$

$$= A(\eta) - \eta^T \cdot \nabla_\eta A(\eta) - \mathbb{E}[\log_e h(x)]$$

(in nats)

Entropy - D-dimensional Gaussians

- $$p(x \mid \mu, \Sigma) = \det(2\pi\Sigma)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

$$H(X) = \mathbb{E}[-\log p(X \mid \mu, \Sigma)]$$

$$= \frac{1}{2}\log\det(2\pi\Sigma) + \frac{1}{2}\mathbb{E}\left[(X-\mu)^T \Sigma^{-1}(X-\mu)\right] \cdot \log e$$

$$(X-\mu)^T \Sigma^{-1}(X-\mu) = \mathrm{Tr}\left[\Sigma^{-1}(X-\mu)(X-\mu)^T\right]$$

$$= \sum$$

$$= \frac{1}{2}\log\det(2\pi\Sigma) + \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\mathbb{E}\left[(X-\mu)(X-\mu)^T\right]\right] \cdot \log e$$

$$= \frac{1}{2}\log\det(2\pi\Sigma) + \frac{1}{2} D \cdot \log e$$

$$= \frac{1}{2}\log\det(2\pi \cdot e \cdot \Sigma)$$

Machine Learning 2

Information Theory
- Maximum Entropy Principle

Patrick Forré

Maximum Entropy Principle

- "Whenever you don't know what probability distribution to pick, then take a distribution expressing the highest amount of uncertainty (=Entropy), given all (known) constraints.":

$$p^* \in \operatorname*{argmax}_{p \in \mathcal{C}} H(p)$$

where $H(p) = \mathbb{E}[-\log p(X)]$ is the entropy for $X \sim p$ and $\mathscr{C}$ the constraint set.

- Used for choosing a prior in Bayesian statistics.
- Used in classical statistical mechanics in physics.

Exercise: Deriving Exponential Families

- We want to derive a distribution $p(x)$ for X that satisfies the constraints:
  $\mathbb{E}_p[T_j(X)] = \tau_j \quad$ for $j = 1, \ldots, s$ for given functions T_j and numbers τ_j.
- The Lagrangian for the maximum entropy principle is then:

$$\mathscr{L} = H(p) + \sum_{j=1}^{s} \eta_j \cdot \left(\mathbb{E}_p[T_j(X)] - \tau_j \right) + (1 - A) \cdot \left(\int p(x)\, dx - 1 \right)$$

  with suggestively chosen Lagrange multipliers η_j, and $(1 - A)$
- The solution of the maximum entropy principle is then of exponential family form with the suggested sufficient statistics and natural parameters, $h(x) = 1$, etc.

Machine Learning 2

Information Theory
- Relative Entropy

Patrick Forré

Recap: Jennsen's Inequality

- Let X be a real-valued random variable and φ a convex function (e.g. if $\varphi''(x) \geq 0$ for all x). Then we have the inequality:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

e.g. $|\mathbb{E}[X]| \leq \mathbb{E}|X|$

- In case φ is strictly convex (e.g. if $\varphi''(x) > 0$ for all x) then: equality holds in the above inequality if and only if X is (almost-surely) a constant random variable (with value $\mathbb{E}[X]$).

Kullback-Leibler Divergence

- Let $p(x)$ and $q(x)$ be two probability distribution on the same space. Then the relative entropy or Kullback-Leibler divergence of p and q is:

$$\mathrm{KL}(p\|q) := \begin{cases} \int p(x) \cdot \log \frac{p(x)}{q(x)} \, dx & , \text{ continuous case} \\ \sum_x p(x) \log \frac{p(x)}{q(x)} & , \text{ discrete case} \\ \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right] & \end{cases}$$

$$= -\mathbb{E}_p[\log q] - (\mathbb{E}_p[-\log p]) = \underbrace{C(p,q)}_{\text{cross-entropy}} - \underbrace{H(p)}_{\text{entropy}}$$

- Conventions: $0 \cdot \infty := 0,\ 1/\infty := 0,\ 1/0 := \infty,\ 0/0 := 0.$

Fundamental Inequality of Information Theory

- Let $p(x)$ and $q(x)$ be two probability distribution on the same space. Then we always have the inequality:

$$\mathrm{KL}(p\|q) \geq 0$$

- Furthermore, equality holds if and only if $p(x) = q(x)$ for (p-almost) all x.

- Proof: $-\log$ is strictly convex. By Jennsen's inequality we get:

$$KL(p\|q) = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathbb{E}_p\left[-\log \frac{q}{p}\right] \geq -\log \mathbb{E}_p\left[\frac{q}{p}\right] = -\log \int p(x) \frac{q(x)}{p(x)}\, dx$$

$$= -\log 1 = 0$$

$$= \Leftrightarrow \frac{q(x)}{p(x)} = 1 \ \forall x \Leftrightarrow q(x) = p(x) \ \forall x.$$

Kullback-Leibler Divergence - Remarks

- KL is not symmetric in the arguments and can take value ∞.
- KL measures the average additional bits needed to encode the location of x sampled from “true” distribution $p(x)$, but using “proposal” distribution $q(x)$ (instead of $p(x)$)
- Minimizing the KL divergence between some distributions can be described as the principle of minimal relative entropy.

Deriving Maximum Likelihood Estimation

- Let $x_1, \ldots, x_N$ be i.i.d. data with true distribution $q(x)$.
- Let $\{\, p(x \mid \theta) \mid \theta \in \Theta \,\}$ be a statistical model aimed to describe $q(x)$.
- The principle of minimal relative entropy tells us to pick:

$$\theta^* \in \arg\min_{\theta \in \Theta} KL(q(x) \,\|\, p(x|\theta)) = \mathbb{E}_q\left[\log \frac{q(x)}{p(x|\theta)}\right]$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \log \frac{q(x_n)}{p(x_n|\theta)} = -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\theta) + \text{const.}$$

Where we used the law of large numbers.

- Maximum likelihood estimation:

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\theta)$$

Example: Exponential Families

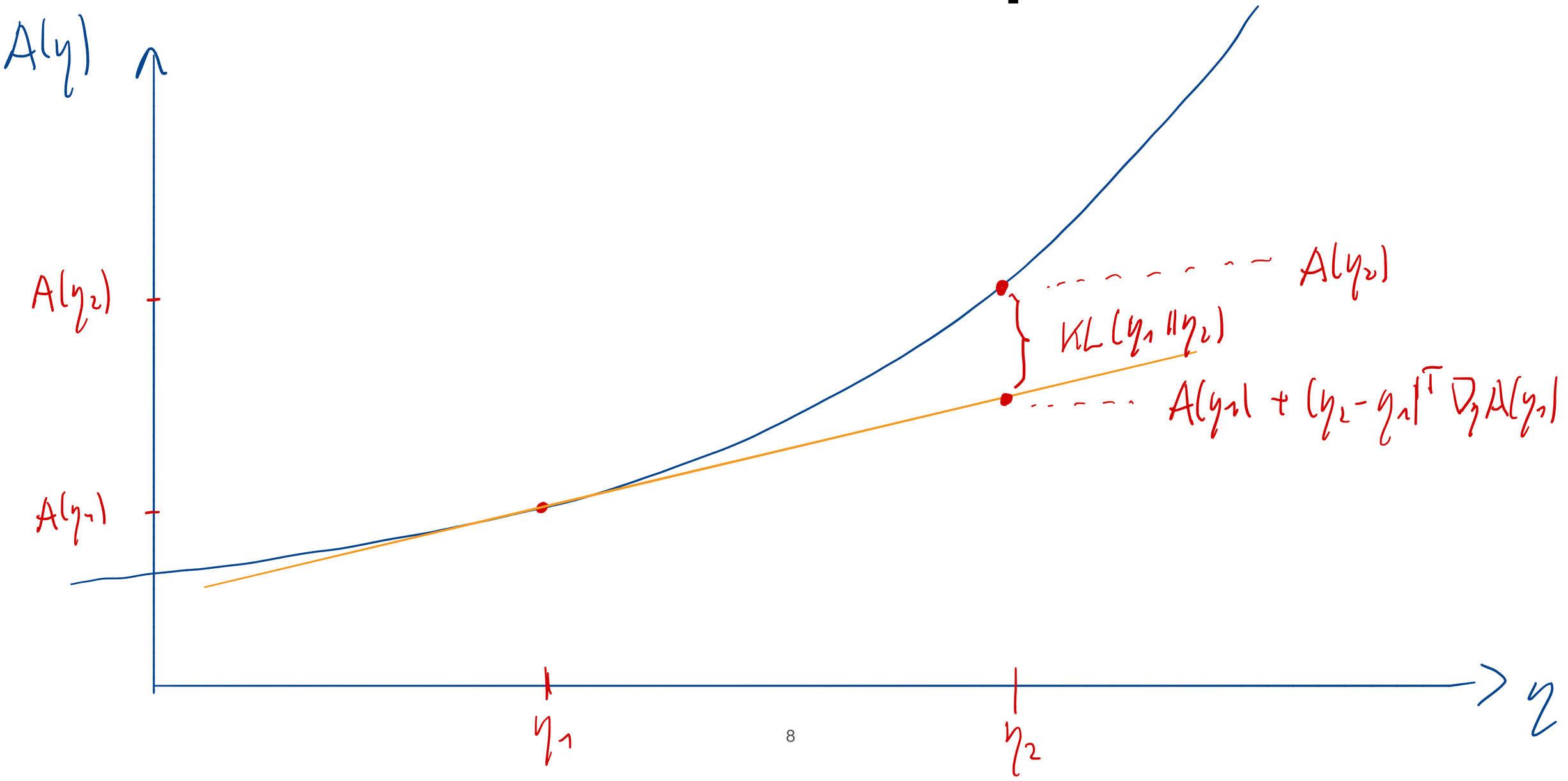
- Exponential family: $p(x \mid \eta) = h(x) \cdot \exp\left(\eta^\top T(x) - A(\eta)\right)$
- Pick two natural parameters η_1, η_2.

$$\mathrm{KL}(p(X \mid \eta_1) \,\|\, p(X \mid \eta_2)) = \mathbb{E}_{\eta_1}\left[\log \frac{p(x|\eta_1)}{p(x|\eta_2)}\right]$$

$$= \mathbb{E}_{\eta_1}\left[(\eta_1 - \eta_2)^\top T(x) - A(\eta_1) + A(\eta_2)\right] = (\eta_1 - \eta_2)^\top \underbrace{\mathbb{E}_{\eta_1}[T(x)]}_{\nabla_\eta A(\eta_1)} - A(\eta_1) + A(\eta_2)$$

$$= A(\eta_2) - \left[(\eta_2 - \eta_1)^\top \cdot \nabla_\eta A(\eta_1) + A(\eta_1)\right]$$

Information Geometric Interpretation of KL



Machine Learning 2

Information Theory
- Conditional Mutual Information and more

Patrick Forré

Joint Entropy & Conditional Entropy

- Let X, Y be two random variables with joint distribution $p(x, y)$.
- The Joint Entropy of X, Y is then:

$$H(X, Y) := H((X,Y)) := H(p(X,Y)) = \mathbb{E}_p[-\log p(X,Y)]$$

- The Conditional Entropy of X given Y is:

$$H(X \mid Y) := \mathbb{E}_y[H(p(X|Y=y))] = \int p(y)\left[\int p(x|y)\cdot(-\log p(x|y))\,dx\right]dy$$

$$= H(X,Y) - H(Y) = \int p(x,y)\,(-\log p(x|y))\,dx\,dy$$

$$= H(Y|X) + H(X) - H(Y)$$

(Conditional) Mutual Information

- Let X, Y, Z be three random variables with joint distribution $p(x, y, z)$.
- The Mutual Information between X and Y is:

$$I(X;Y) := KL(p(x,y) \,\|\, p(x) \cdot p(y))$$
$$= H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

- The Conditional Mutual Information between X and Y given Z is:

$$I(X;Y|Z) := \mathbb{E}_z[\, KL(p(x,y|z=z) \,\|\, p(x|z=z) \cdot p(y|z=z)\,]$$
$$= H(X|Z) - H(X|Y,Z)$$
$$= H(Y|Z) - H(Y|X,Z)$$

Shannon's Theorem

- Let X, Y, Z be three random variables with joint distribution $p(x, y, z)$.

  We always have the inequality:

$$I(X; Y \mid Z) \geq 0$$

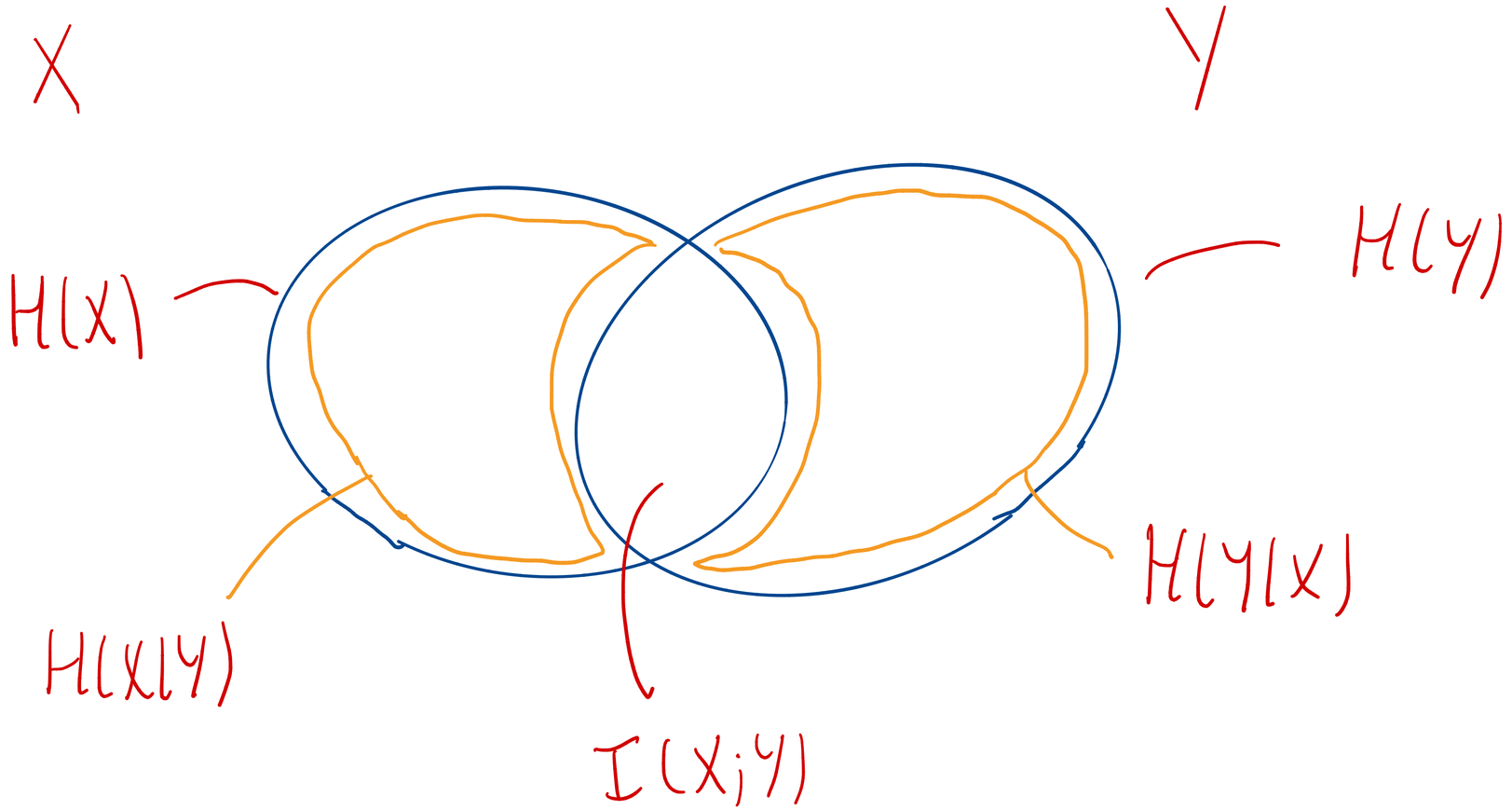
- Furthermore, equality holds if and only if: $X \perp\!\!\!\perp Y \mid Z$

- The statement also holds for the (unconditional) mutual information and (unconditional) independence (i.e. with $Z = \varnothing$).

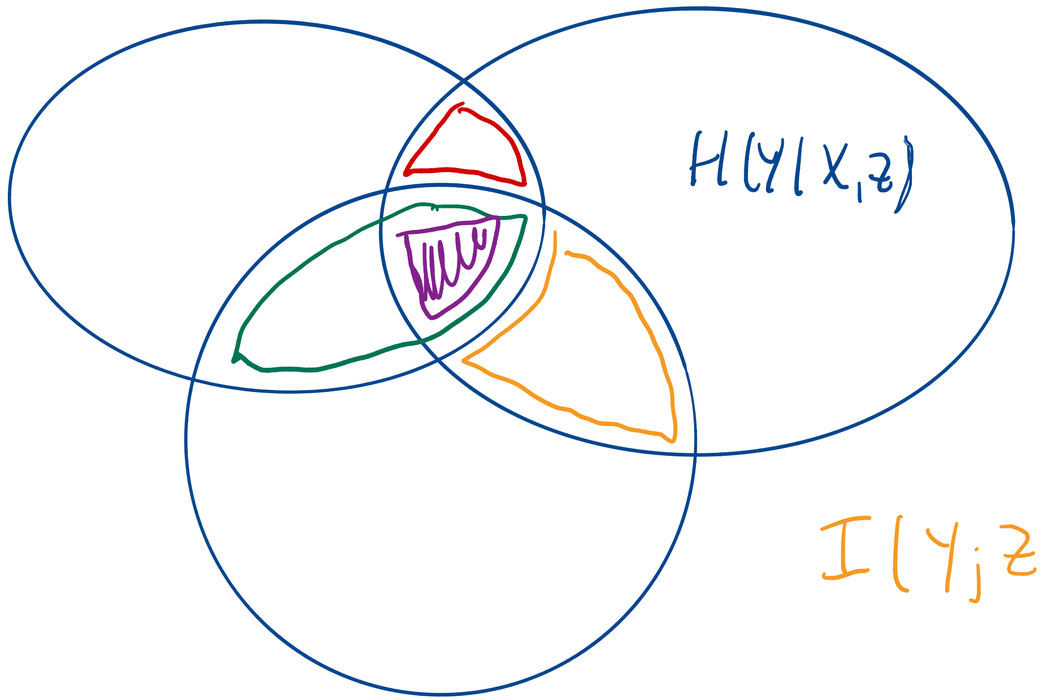
- Proof: $I(X;Y|Z) = \mathbb{E}_Z[KL(p(X,Y|Z) \,\|\, p(X|Z)\,p(Y|Z))] \geq 0$

$= 0 \iff \forall z : KL(\cdots) = 0 \iff \forall x,y,z \; p(x,y|z) = p(x|z) \cdot p(y|z)$

Information Diagram for 2 Variables



Information Diagram for 3 Variables



Chain Rule

X, Y, Z, W

$$I(X; (Y,W) \mid Z) = I(X; Y \mid Z) + I(X; W \mid Y, Z).$$

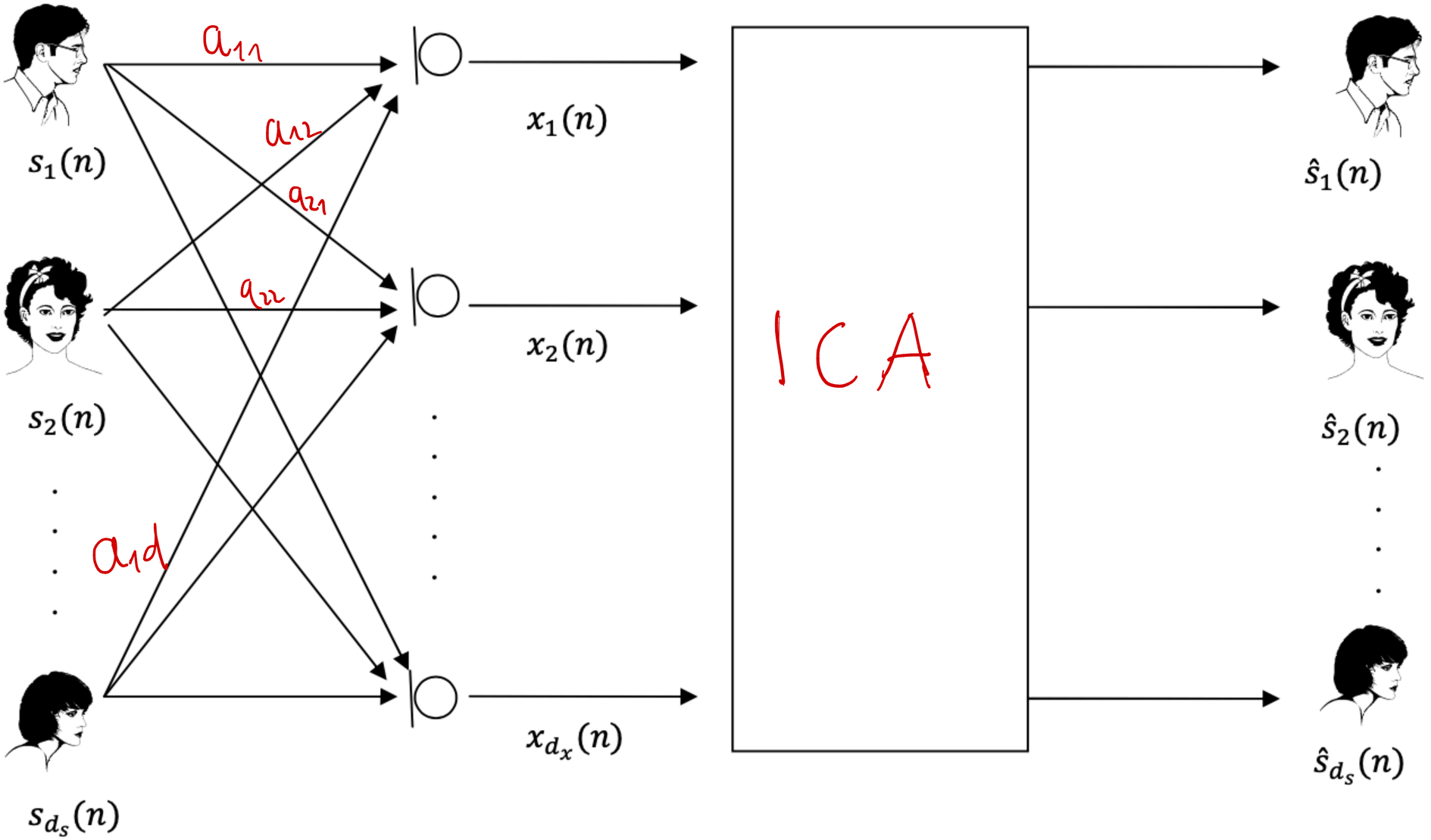
≥ 0 ≥ 0 ≥ 0

Machine Learning 2

Independent Component Analysis (ICA)
- Objective and Assumptions

Patrick Forré

The Cocktail Party Problem



Source: S.R. Lakmal: Blind source separation in real time

ICA - Assumptions

- Independent Components: We have K random sources:
  $S(t) := [s_1(t), \ldots, s_K(t)]^\top$ over time steps $t = 1, \ldots, T$; (K x T)-matrix S;
  satifying: $s_k(t) \perp\!\!\!\perp s_{-k}(t)$ for all $k = 1, \ldots, K$, $t = 1, \ldots, T$.
- Linear and noiseless measurements: We measure M signals:
  $X(t) := [x_1(t), \ldots, x_M(t)]^\top$; (M x T)-matrix X; such that:
  $X = A\,S$ with time-independent (M x K)-matrix A.
- Completeness: A is invertible, $M = K$, with invers $W = A^{-1}$.
- Non-Gaussianity: None (or at most one) of the sources $s_k(t)$ for $k = 1, \ldots, K$ is Gaussian distributed, e.g. if all sources are super-Gaussian distributed.

ICA - Identifiability Theorem

- Problem: Recover both mixing matrix A and sources S from the data X.
- Identifiability Theorem: Under the mentioned assumptions, one can recover the sources $s_k(t)$ up to sign, scale and permutation and corresponding A.

- W.l.o.g. we fix the scale by assuming $\mathrm{var}(s_k(t)) = 1$ for all k, t.
- Together with the independence we then have: $\mathrm{cov}(s_i(t), s_j(t)) = \delta_{i,j}$

ICA - helpful pre-processing steps

- In most ICA algorithms the following pre-processing steps of the data are helpful (but for the algorithm from this lecture not necessary):

- Center the data by subtracting mean:
  $X = A\,S$ implies: $\mathbb{E}[X] = A\,\mathbb{E}[S]$ implies: $X - \mathbb{E}[X] = A\,(S - \mathbb{E}[S])$.
  So w.l.o.g.: $\mathbb{E}[S] = 0$ and $\mathbb{E}[X] = 0$.

- "Whiten" the data by using full-rank PCA and rescaling of the PCs:
  So we get matrix B with $BB^\top$ diagonal s.t. BX has covariance matrix I.
  So we write X for BX and A for BA again. Then we have:
  $I = \mathbb{E}[XX^\top] = A\mathbb{E}[SS^\top]A^\top = AA^\top$, thus A is orthogonal.
  So we reduced parameters from K^2 to $K(K-1)/2$. "PCA is half-ICA".

Machine Learning 2

Independent Component Analysis (ICA)
- Maximum Likelihood Derivation of ICA with Natural Gradients

Patrick Forré

ICA - Maximum Likelihood Objective

- Under the mentioned ICA assumptions: $X = A\,S$ or equiv.: $S = W\,X$
- By independence we have: $p_S(s_1, \ldots, s_K) = p_{S_1}(s_1) \cdots p_{S_K}(s_K)$
- Parameters are the unknowns: $\theta = (W, p_{S_1}, \ldots, p_{S_K})$
- Statistical model for x: $p(x\,|\,\theta) = p_S(W \cdot x) \cdot \left|\frac{\partial S}{\partial x}\right| = p_S(W \cdot x) \cdot |W|$

$$= p_{S_1}(w_1^T x) \cdots p_{S_K}(w_K^T x) \cdot |W|$$

- Log-likelihood: $L(\theta) = \hat{\mathbb{E}}_x[\log p(x\,|\,\theta)] = \hat{\mathbb{E}}\left[\sum_{k=1}^{K} \log p_{S_k}(w_k^T x)\right] + \log|W|$
- MLE objective: $\arg\max_{\theta} L(\theta)$

Recap: Natural Gradient

- Objective: Minimize $L(w)$
- Gradient Descent: $w^{\text{new}} := w - \alpha \cdot \nabla_w L(w),$
- Newton-Raphson: $w^{\text{new}} := w - \alpha \cdot \left(\nabla_w^2 L(w) \right)^{-1} \cdot \nabla_w L(w),$
- Natural Gradient: $w^{\text{new}} := w - \alpha \cdot M \cdot \nabla_w L(w),$

  where M is a so called pre-conditioner, usually a symmetric, positive-definite matrix reflecting the geometry of the space.

ICA - Gradient of Log-Likelihood

- Maximize: $L(\theta) = \hat{\mathbb{E}}_x \left[\sum_{k=1}^{K} \log p_{S_k}(w_k^\top x) + \log |W| \right]$

- $\frac{\partial L}{\partial w_{ij}} = \hat{\mathbb{E}}_x \left[\sum_{u=1}^{u} \frac{\partial}{\partial w_{ij}} \log p_{S_u}(w_u^\top x) \right] + \frac{\partial}{\partial w_{ij}} \log |W|$

$= \hat{\mathbb{E}}_x \left[\sum_{u=1}^{u} \phi_u(w_u^\top x) \cdot (\delta_{ui} \cdot x_j) \right] + (W^{-1})_{ji}$

$= \hat{\mathbb{E}}_x \left[\phi_i(w_i^\top x) \cdot x_j \right] + (W^{-1})_{ji}$

$\phi_u(s) := \frac{\partial}{\partial s} \log p_{S_u}(s)$

ICA - Activation Functions Approximation

- Approximation of the quantities related to p_{S_k} :

- $\phi_k(s) = \frac{\partial}{\partial s} \log p_{S_k}(s) \approx$ $\begin{cases} -\tanh(s), & S_k \text{ super-Gaussian} \\ \tanh(s) - s, & S_k \text{ sub-Gaussian} \end{cases}$

ICA - Natural Gradient

- A good data-independent preconditioner for $L(\theta)$ turns out to be:

$$M_{(lm),(ij)} = \sum_{\nu} W_{\nu m} W_{\nu j} \cdot \delta_{ij}$$

- Natural gradient: $\left(M \cdot \nabla_W L\right)_{(lm)} = \sum_{i,j,\nu} W_{\nu m} W_{\nu j} \delta_{ie} \left(\nabla_W L\right)_{(ij)}$

$K^2 \times K^2 \quad K^2 \times 1$

$$= \sum_{j,\nu} W_{\nu m} W_{\nu j} \cdot \left[\hat{\mathbb{E}}_x\left[\phi_e(w_e^T x) \cdot x_j\right] + (W^{-1})_{je}\right]$$

$$= \sum_{\nu} W_{\nu m} \hat{\mathbb{E}}_x\left[\phi_e(w_e^T x)(w_\nu^T \cdot x)\right] + \underbrace{\sum_{\nu} W_{\nu m} \delta_{\nu e}}_{W_{em}}$$

$$= W_{em} + \sum_{\nu} W_{\nu m} \hat{\mathbb{E}}_x\left[\phi_e(w_e^T x) \cdot (w_\nu^T x)\right].$$

ICA - Update Rule

$x(t)$

- $W^{\text{new}} = W + \alpha \cdot \left[W + \phi\left(W \cdot x(t) \right) \cdot x(t)^T W^T W \right]$

$\hat{W} \rightsquigarrow$ MLE (after convergence)

$\hat{A} \simeq \hat{W}^{-1}$

- After convergence reconstruct:

$$\hat{s}_k(t) := \left(\hat{W} \cdot x(t) \right)_k$$

$$\hat{s}(t) := \hat{W} \cdot x(t)$$

Machine Learning 2

Independent Component Analysis (ICA)
- Covariant Online ICA Algorithm
- Summary

Patrick Forré

ICA - Assumptions for Identifiability

- Independent sources: $s_k(t) \perp\!\!\!\perp s\_k(t)$.
- Linear and noiseless measurements: $X = A\,S$ with matrix A.
- Completeness: A is invertible, $W := A^{-1}$.
- Non-Gaussianity: Sources $s_k(t)$ are non-Gaussian (or at most one is).
- Theorem: Under those assumptions both mixing matrix A and sources $s_k(t)$ can be identified up to sign, scale and permutation of the sources.

Covariant Online ICA Algorithm

- Chose learning rate α, activation functions ϕ_k (usually $\phi_k(s) = -\tanh(s)$ for super-Gaussian and $\phi_k(s) = \tanh(s) - s$ for sub-Gaussian), initialize W
- Until convergence take new data point $x(t)$ and do:
  - $c(t) := W \cdot x(t)$
  - $z(t) := \phi(c(t))$
  - $y(t) := W^\top \cdot c(t)$
  - $W^{\text{new}} := W + \alpha \cdot [W + z(t) \cdot y(t)^\top]$
- Reconstruct sources: $\hat{s}(t) := \hat{W} \cdot x(t)$