Information Theory and Applications in Vision

Ying Nian Wu
UCLA Department of Statistics

July 31, 2013
IPAM Summer School

Goal: A *gentle* introduction to the basic concepts in information theory and a couple of applications in vision

Emphasis: understanding and interpretations of these concepts

Reference: *Elements of Information Theory* by Cover and Thomas

Topics

- Entropy and Kullback-Leibler
- Asymptotic equipartition property
- Large deviation
- Image labeling
- Image modeling

Entropy

Randomness or uncertainty of a probability distribution

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K

$$\Pr(X = a_k) = p_k \qquad\qquad p(x) = \Pr(X = x)$$

Example

Ω	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$

Entropy

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K

$$\Pr(X = a_k) = p_k \qquad p(x) = \Pr(X = x)$$

Definition

$$H(X) = -\sum_k p_k \log p_k = -\sum_{x \in \Omega} p(x) \log p(x)$$

Entropy

Example

Ω	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$

$$H(X) = -\sum_{k} p_k \log p_k = -\sum_{x \in \Omega} p(x) \log p(x)$$

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4}$$

Entropy

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K

$$\Pr(X = a_k) = p_k \qquad\qquad p(x) = \Pr(X = x)$$

$$H(X) = -\sum_k p_k \log p_k = -\sum_{x \in \Omega} p(x) \log p(x)$$

$$H(p) = \mathbf{E}[-\log p(X)]$$

Definition for both discrete and continuous

Recall $\mathbf{E}[f(X)] = \sum_x f(x)p(x)$

Interpretation 1: cardinality

Uniform distribution $X \sim \mathbf{Unif}[A]$

There are $|A|$ elements in A
All these choices are equally likely

$$H(X) = \mathbf{E}[-\log p(X)] = \mathbf{E}[-\log(1/|A|)] = \log|A|$$

Entropy can be interpreted as log of volume or size
n dimensional cube has 2^n vertices
can also be interpreted as dimensionality

What if the distribution is not uniform?

Asymptotic equipartition property

Any distribution is essentially a uniform distribution in *long run repetition*

Recall if $X \sim \mathbf{Unif}[A]$ then $p(X) = 1/|A|$ a constant

$$X_1, X_2, \ldots, X_n \sim p(x) \quad \text{independently}$$

$$p(X_1, X_2, \ldots, X_n) = p(X_1)p(X_2)\ldots p(X_n) \quad \text{Random?}$$

But in some sense, it is essentially a constant!

Law of large number

$$X_1, X_2, \ldots, X_n \sim p(x) \quad \text{independently}$$

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbf{E}[f(X)] = \sum_{x} f(x)p(x)$$

Long run average converges to expectation

Asymptotic equipartition property

$$p(X_1, X_2, \ldots, X_n) = p(X_1)p(X_2)\ldots p(X_n)$$

$$\frac{1}{n}\log p(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \to \mathbf{E}[\log p(X)] = -H(p)$$

Intuitively, in the long run, $p(X_1, \ldots, X_n) \approx 2^{-nH(p)}$

Asymptotic equipartition property

$$p(X_1,\ldots,X_n) \approx 2^{-nH(p)}$$

Recall if $X \sim \mathbf{Unif}[A]$ then $p(X) = 1/|A|$ a constant

Therefore, as if $(X_1,\ldots,X_n) \sim \mathbf{Unif}[A]$,with $|A| = 2^{nH(p)}$

So the dimensionality per observation is $H(p)$

We can make it more rigorous

Weak law of large number

$X_1, X_2, \ldots, X_n \sim p(x)$ independently

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbf{E}[f(X)] = \sum_{x} f(x)p(x) = \mu$$

for $n \ge n_0$

$$\Pr(|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mu| < \varepsilon) > 1 - \delta$$

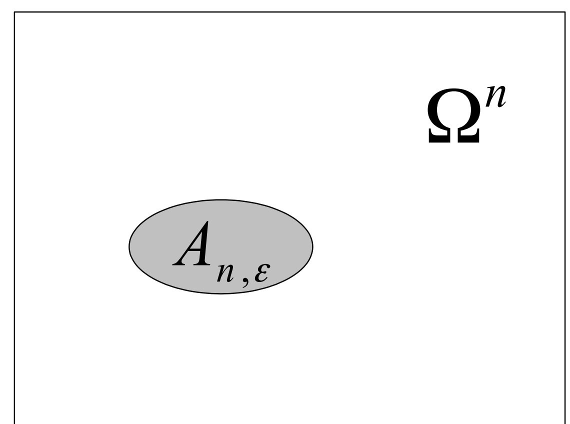
Typical set

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K

$$p(X_1,\ldots,X_n) \approx 2^{-nH(p)}$$

$$(X_1,\ldots,X_n) \sim \textbf{Unif}[A] \quad \text{,with } |A| = 2^{nH(p)}$$

Typical set $\quad A_{n,\varepsilon} \in \Omega^n$



Typical set

$$p(X_1,\ldots,X_n) \approx 2^{-nH(p)}$$

$$(X_1,\ldots,X_n) \sim \mathbf{Unif}[A] \quad \text{,with } |A| = 2^{nH(p)}$$

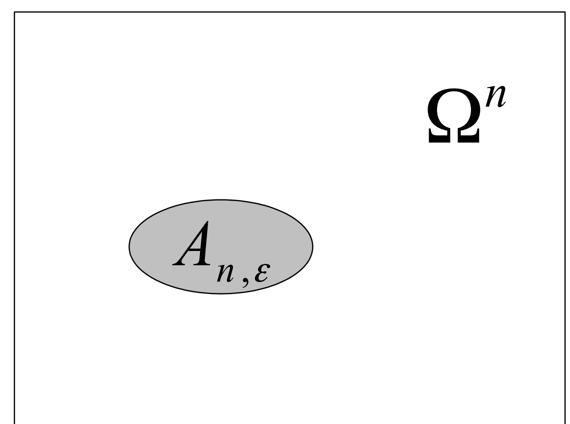
$A_{n,\varepsilon}$: The set of sequences $(x_1, x_2,\ldots, x_n) \in \Omega^n$

$$2^{-n(H(p)+\varepsilon)} \le p(x_1, x_2,\ldots, x_n) \le 2^{-n(H(p)-\varepsilon)}$$

for n sufficiently large

$$p(A_{n,\varepsilon}) > 1-\delta$$

$$(1-\delta)2^{n(H(p)-\varepsilon)} \le |A_{n,\varepsilon}| \le 2^{n(H(p)+\varepsilon)}$$



Interpretation 2: coin flipping

Flip a fair coin → {Head, Tail}
Flip a fair coin twice independently
→ {HH, HT, TH, TT}

……

Flip a fair coin n times independently
→ 2^n equally likely sequences

We may interpret entropy as the number of flips

Interpretation 2: coin flipping

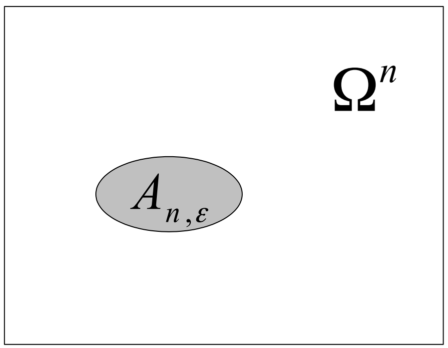
Example

	a	b	c	d
Pr	$1/4$	$1/4$	$1/4$	$1/4$
flips	HH	HT	TH	TT

The above uniform distribution amounts to 2 coin flips

Interpretation 2: coin flipping

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K



$$p(X_1,\ldots,X_n) \approx 2^{-nH(p)}$$

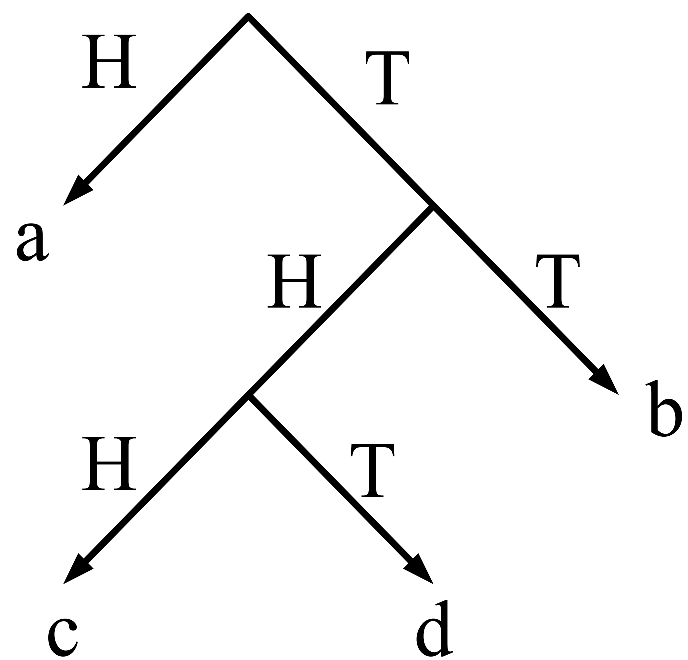
$$(X_1,\ldots,X_n) \sim \mathbf{Unif}[A] \quad \text{,with } |A| = 2^{nH(p)}$$

$(X_1,\ldots,X_n)$ amounts to $nH(p)$ flips

$$\boxed{X \sim p(x) \text{ amounts to } H(p) \text{ flips}}$$

Interpretation 2: coin flipping

	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$
flips	*H*	*TT*	*THH*	*THT*



Interpretation 2: coin flipping

Ω	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$
$-\log p$	1	2	3	3

$$H(X)=\frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3=\frac{7}{4}\,\textbf{flips}$$

	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$
flips	H	TT	THH	THT

Interpretation 3: coding

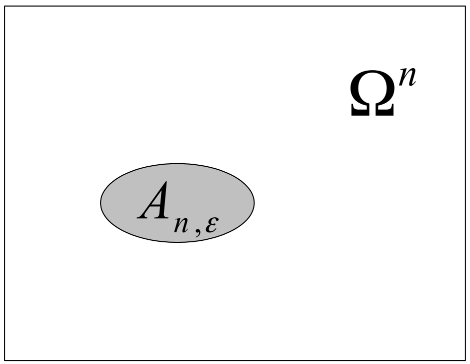
Example

A	a	b	c	d
Pr	$1/4$	$1/4$	$1/4$	$1/4$
code	11	10	01	00

$$\textbf{length} = 2 = \log 4 = -\log(1/4)$$

Interpretation 3: coding

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K



$$p(X_1,\ldots,X_n) \approx 2^{-nH(p)}$$

$$(X_1,\ldots,X_n) \sim \textbf{Unif}[A] \quad \text{,with } |A| = 2^{nH(p)}$$

How many bits to code elements in A ?

$nH(p)$ bits

Can be made more formal using typical set

Prefix code

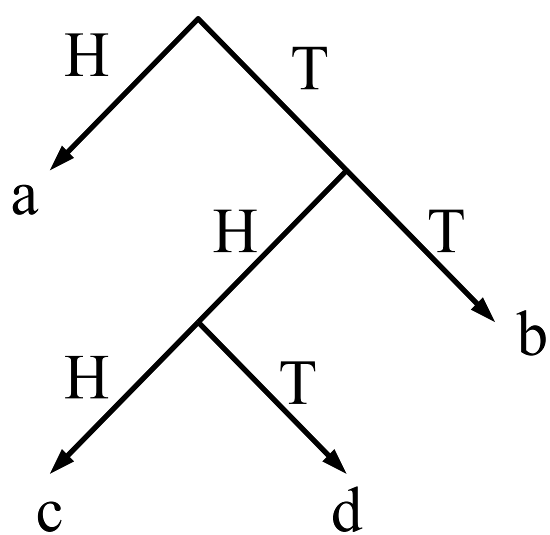
	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$
code	1	00	011	010

100101100010→abacbd

$$\mathbf{E}[l(X)] = \sum_x l(x)p(x) = 1\times\frac{1}{2} + 2\times\frac{1}{4} + 3\times\frac{1}{8} + 3\times\frac{1}{8} = \frac{7}{4}\,\textbf{bits}$$

Optimal code

	a	b	c	d
Pr	$1/2$	$1/4$	$1/8$	$1/8$
code	1	00	011	010



100101100010→abacbd

Sequence of coin flipping
A completely random sequence
Cannot be further compressed

$$l(x) = -\log p(x)$$

$$\mathbf{E}[l(X)] = H(p)$$

e.g., two words I, probability

Optimal code

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K
l	l_1	l_2	$\cdots$	l_k	$\cdots$	l_K

Kraft inequality for prefix code

$$\sum_k 2^{-l_k} \le 1$$

Minimize $L = \mathbf{E}[l(X)] = \sum_k l_k p_k$

Optimal length $l_k^* = -\log p_k$

$$L^* = H(p)$$

Wrong model

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
True	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K
Wrong	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K

Optimal code $L^* = \mathbf{E}[l^*(X)] = \sum_k l_k^* p_k = -\sum_k p_k \log p_k$

Wrong code $L = \mathbf{E}[l(X)] = \sum_k l_k p_k = -\sum_k p_k \log q_k$

Redundancy $L - L^* = \sum_k p_k \log \frac{p_k}{q_k}$

Box: All models are wrong, but some are useful

Relative entropy

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K
q	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K

$$D(p \mid q) = \mathbf{E}_p[\log \frac{p(X)}{q(X)}] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Kullback-Leibler divergence

Relative entropy

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
p	p_1	p_2	$\cdots$	p_k	$\cdots$	p_K
q	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K

Jensen inequality

$$D(p \mid q) = \mathbf{E}_p[\log \frac{p(X)}{q(X)}] = -\mathbf{E}_p[\log \frac{q(X)}{p(X)}] \geq -\log(\mathbf{E}_p[\frac{q(X)}{p(X)}]) = 0$$

$$D(p \mid U) = \mathbf{E}_p[\log p(X)] + \log |\Omega| = \log |\Omega| - H(p)$$

Types

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
q	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K
freq	n_1	n_2	$\cdots$	n_k	$\cdots$	n_K

$X_1, X_2, \ldots, X_n \sim q(x)$ independently

n_k = number of times $X_i = a_k$

$f_k = \dfrac{n_k}{n}$ normalized frequency

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
q	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K
freq	n_1	n_2	$\cdots$	n_k	$\cdots$	n_K

$$f_k = \frac{n_k}{n}$$

Law of large number

$$f = (f_1, \ldots, f_K) \rightarrow q = (q_1, \ldots, q_K)$$

Refinement

$$p(x_1, \ldots, x_n) = \prod_i p(x_i)$$

$$\Pr(f) = \binom{n}{n_1, \ldots, n_K} \prod_k q_k^{\,n_k} = \binom{n}{nf_1, \ldots, nf_K} \prod_k q_k^{\,nf_k}$$

Large deviation

Ω	a_1	a_2	$\cdots$	a_k	$\cdots$	a_K
q	q_1	q_2	$\cdots$	q_k	$\cdots$	q_K
freq	n_1	n_2	$\cdots$	n_k	$\cdots$	n_K

$$f_k = \frac{n_k}{n}$$

Law of large number

$$f = (f_1, \ldots, f_K) \to q = (q_1, \ldots, q_K)$$

Refinement

$$-\frac{1}{n} \log \Pr(f) \to D(f \mid q)$$

$$\Pr(f) \sim 2^{-nD(f|q)}$$

Kolmogorov complexity

Example: a string 011011011011…011

Program:
```
for (i =1 to n/3)
    write(011)
end
```

Can be translated to binary machine code

Kolmogorov complexity = length of shortest machine code that reproduce the string

no probability distribution involved

If a long sequence is not compressible, then it has all the statistical properties of a sequence of coin flipping

string = f(coin flippings)

Joint and conditional entropy

$$
\begin{array}{c|cccccc|c}
\Omega & b_1 & b_2 & \cdots & b_j & \cdots & \\
\hline
a_1 & p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1\bullet} \\
a_2 & p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
a_k & p_{k1} & p_{k2} & \cdots & p_{kj} & \cdots & p_{k\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
\hline
 & p_{\bullet 1} & p_{\bullet 2} & \cdots & p_{\bullet j} & \cdots & 1
\end{array}
$$

e.g., eye color & hair color

Joint distribution

$$\Pr(X = a_k \,\&\, Y = b_j) = p_{kj}$$
$$p(x, y) = \Pr(X = x \,\&\, Y = y)$$

Marginal distribution

$$\Pr(X = a_k) = p_{k\bullet}$$
$$\Pr(Y = b_j) = p_{\bullet j}$$
$$p_X(x) = \Pr(X = x)$$
$$p_Y(y) = \Pr(Y = y)$$

Joint and conditional entropy

$$\begin{array}{c|ccccc|c}
\Omega & b_1 & b_2 & \cdots & b_j & \cdots & \\
\hline
a_1 & p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1\bullet} \\
a_2 & p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
a_k & p_{k1} & p_{k2} & \cdots & p_{kj} & \cdots & p_{k\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
\hline
 & p_{\bullet 1} & p_{\bullet 2} & \cdots & p_{\bullet j} & \cdots & 1
\end{array}$$

Conditional distribution

$$\Pr(X = a_k \mid Y = b_j) = p_{kj} / p_{\bullet j}$$

$$\Pr(Y = b_j \mid X = a_k) = p_{kj} / p_{k\bullet}$$

$$p_{X|Y}(x \mid y) = \Pr(X = x \mid Y = y)$$

$$p_{Y|X}(y \mid x) = \Pr(Y = y \mid X = x)$$

Chain rule

$$p(x, y) = p_X(x) p_{Y|X}(y \mid x) = p_Y(y) p_{X|Y}(x \mid y)$$

Joint and conditional entropy

$$H(X,Y) = -\sum_{kj} p_{kj} \log p_{kj} = -\sum_{x,y} p(x,y) \log p(x,y) = \mathbf{E}[-\log p(X,Y)]$$

$$\begin{array}{c|cccccc|c}
\Omega & b_1 & b_2 & \cdots & b_j & \cdots & \\
\hline
a_1 & p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1\bullet} \\
a_2 & p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
a_k & p_{k1} & p_{k2} & \cdots & p_{kj} & \cdots & p_{k\bullet} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
\hline
 & p_{\bullet 1} & p_{\bullet 2} & \cdots & p_{\bullet j} & \cdots & 1
\end{array}$$

$$H(Y \mid X = x) = -\sum_{y} p(y \mid x) \log p(y \mid x)$$

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x) = -\sum_{x} p(x) \sum_{y} p(y \mid x) \log p(y \mid x)$$

$$H(Y \mid X) = -\sum_{x,y} p(x,y) \log p(y \mid x) = \mathbf{E}[-\log p(Y \mid X)]$$

Chain rule

$$p(x,y) = p_X(x)\,p_{Y|X}(y \mid x) = p_Y(y)\,p_{X|Y}(x \mid y)$$

$$H(X,Y) = H(X) + H(Y \mid X)$$

$$H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1},\ldots,X_1)$$

Mutual information

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(X;Y) = D(p(x,y) \mid p(x)p(y)) = \mathbf{E}[\log \frac{p(X,Y)}{p(X)p(Y)}]$$

$$I(X;Y) = H(X) - H(X \mid Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Entropy rate

Stochastic process $(X_1, X_2, \ldots, X_n, \ldots)$ not independent

Entropy rate: $H = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$
(compression)

Stationary process: $H = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1)$

Markov chain:

$$\Pr(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

Stationary Markov chain: $H = H(X_{n+1} \mid X_n)$

Shannon, 1948

1. **Zero-order approximation**

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. **First-order approximation**

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. **Second-order approximation** (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. **Third-order approximation** (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. **First-order word approximation.**

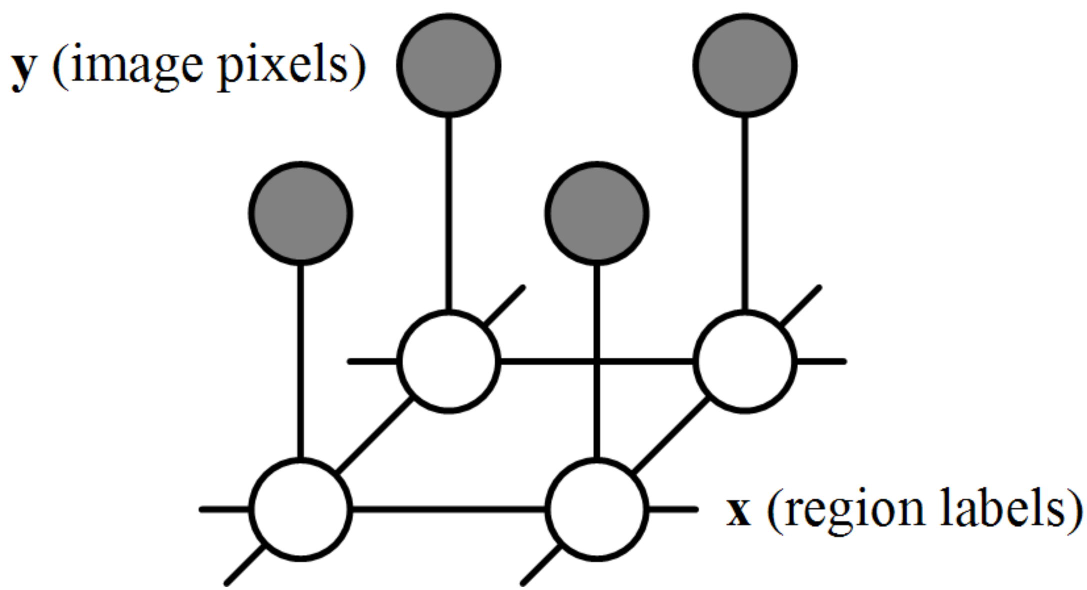
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. **Second-order word approximation.**

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Image labeling

$$P(W \mid \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\{-\sum_{\mu \in \mathcal{V}} \phi_\mu(w_\mu, \mathbf{I}) - \sum_{\mu\nu \in \mathcal{E}} \psi_{\mu\nu}(w_\mu, w_\nu)\}$$

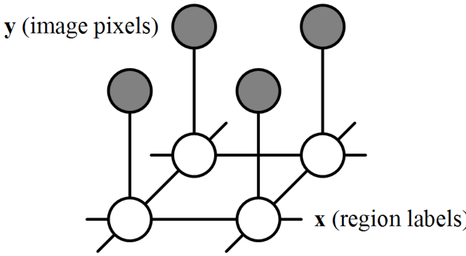


Mean Field Free Energy

$$P(W \mid \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\{-\sum_{\mu \in \mathcal{V}} \phi_\mu(w_\mu, \mathbf{I}) - \sum_{\mu\nu \in \mathcal{E}} \psi_{\mu\nu}(w_\mu, w_\nu)\}$$

$$B(W) = \prod_{\mu \in \mathcal{V}} b_\mu(w_\mu)$$

$$D(B, P) = \mathcal{F}_{\mathrm{MFT}}(B) + \log Z$$

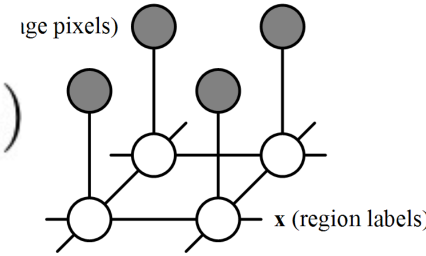


Mean Field Free Energy

$$P(W \mid \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\{-\sum_{\mu \in \mathcal{V}} \phi_\mu(w_\mu, \mathbf{I}) - \sum_{\mu\nu \in \mathcal{E}} \psi_{\mu\nu}(w_\mu, w_\nu)\}$$

$$B(W) = \prod_{\mu \in \mathcal{V}} b_\mu(w_\mu)$$

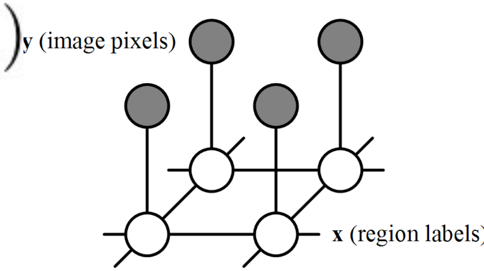
$$\begin{aligned}\mathcal{F}_{\mathrm{MFT}}(B) = & \sum_{\mu\nu \in \mathcal{E}} \sum_{w_\mu, w_\nu} b_\mu(w_\mu) b_\nu(w_\nu) \psi_{\mu\nu}(w_\mu, w_\nu) \\ & + \sum_{\mu \in \mathcal{V}} \sum_{w_\mu} b_\mu(w_\mu) \phi_\mu(w_\mu, \mathbf{I}) \\ & + \sum_{\mu \in \mathcal{V}} \sum_{w_\mu} b_\mu(w_\mu) \log b_\mu(w_\mu)\end{aligned}$$



Bethe Free Energy

$$P(W \mid \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\{-\sum_{\mu \in \mathcal{V}} \phi_\mu(w_\mu, \mathbf{I}) - \sum_{\mu\nu \in \mathcal{E}} \psi_{\mu\nu}(w_\mu, w_\nu)\}$$

$$\begin{aligned}\mathcal{F}_{\text{Bethe}}(B) = & \sum_{\mu\nu \in \mathcal{E}} \sum_{w_\mu, w_\nu} b_{\mu,\nu}(w_\mu, w_\nu)\psi_{\mu\nu}(w_\mu, w_\nu) \\ & + \sum_{\mu \in \mathcal{V}} \sum_{w_\mu} b_\mu(w_\mu)\phi_\mu(w_\mu, \mathbf{I}) \\ & + \sum_{\mu\nu \in \mathcal{E}} \sum_{w_\mu, w_\nu} b_{\mu,\nu}(w_\mu, w_\nu) \log b_{\mu,\nu}(w_\mu, w_\nu) \\ & - \sum_{\mu \in \mathcal{V}} (n_\mu - 1) \sum_{w_\mu} b_\mu(w_\mu) \log b_\mu(w_\mu)\end{aligned}$$

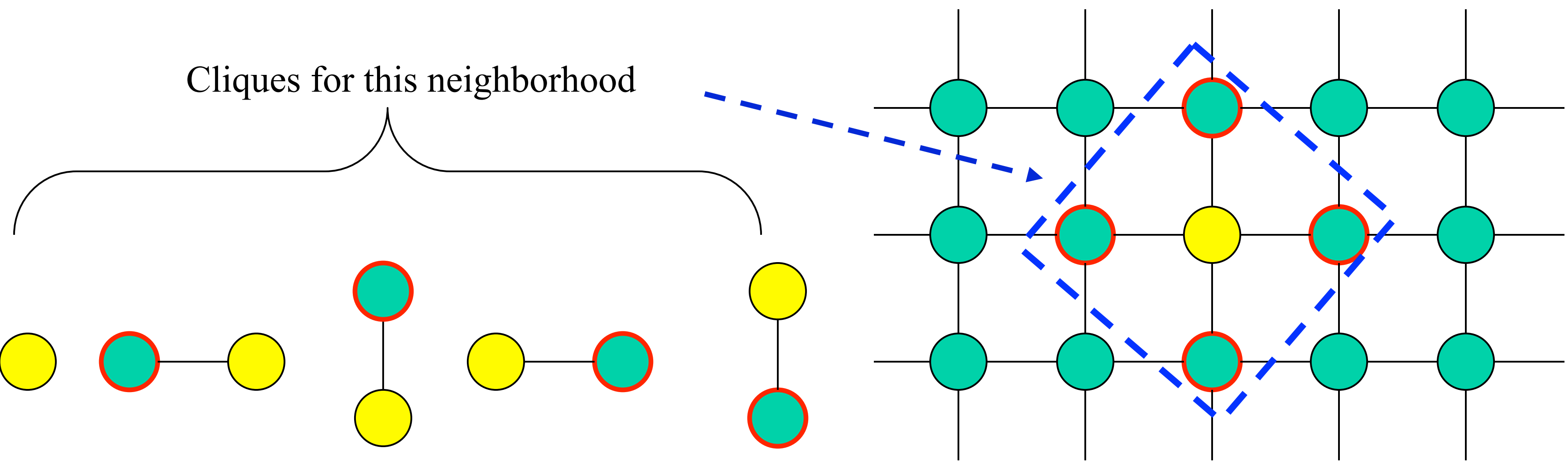


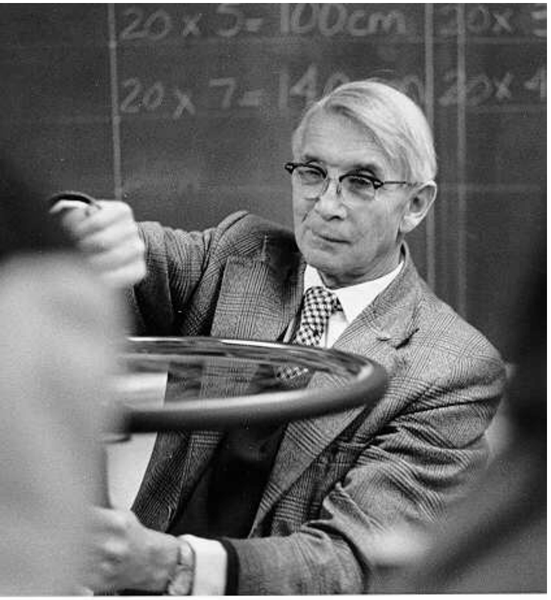
Belief propagation

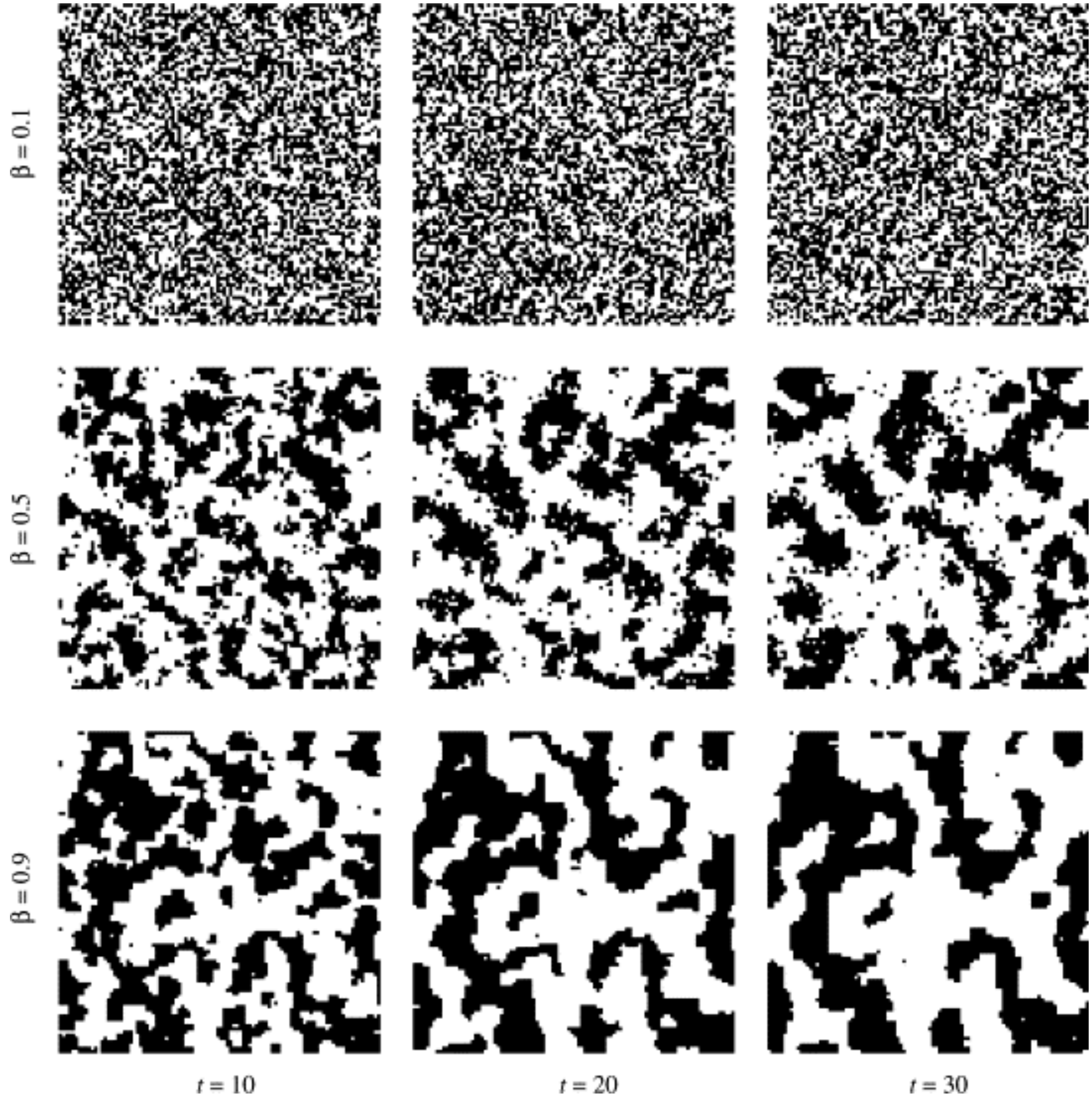
Image Modeling

Ising model

$$P(\mathbf{I}) = \frac{1}{Z}\exp\{\sum_s \alpha \mathbf{I}_s + \sum_{s\sim t} \beta \mathbf{I}_s \mathbf{I}_t\}$$






β = 0.1
β = 0.5
β = 0.9
t = 10
t = 20
t = 30

Energy-based model

Ising model

$$P(\mathbf{I}) = \frac{1}{Z}\exp\{\sum_s \alpha \mathbf{I}_s + \sum_{s\sim t} \beta \mathbf{I}_s \mathbf{I}_t\}$$

Hidden variables, layers, RBM

Exponential family model, log-linear model

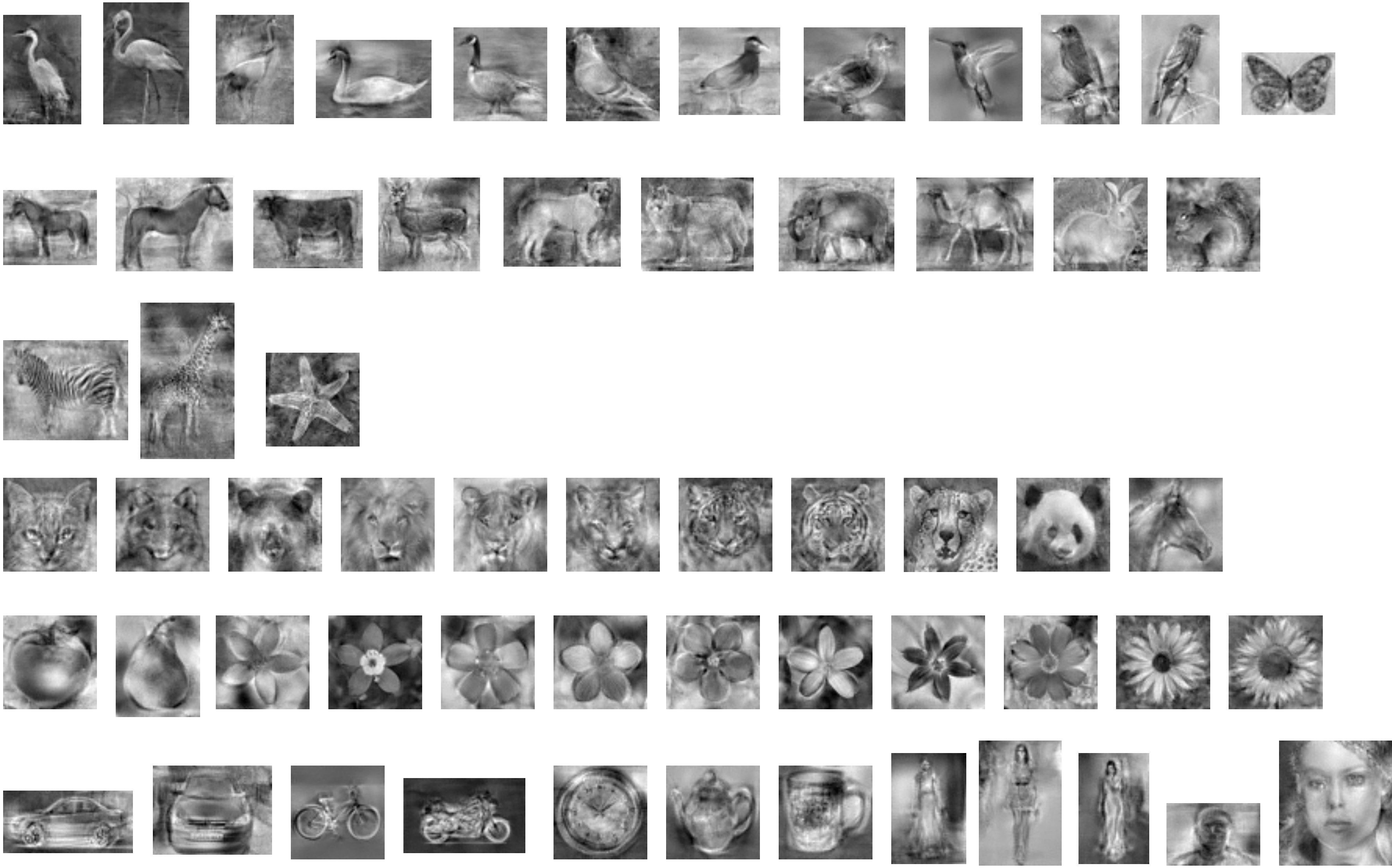
maximum entropy model

$$P(\mathbf{I}) = \frac{1}{Z(\lambda)}\exp\{\sum_i \lambda_i \phi_i(\mathbf{I})\} q(\mathbf{I})$$

$\lambda = (\lambda_i, i = 1, \ldots, n)$ unknown parameters

$\phi_i()$ features (may also need to be learned)

$q()$ reference distribution



Maximum Likelihood

$$\mathbf{I}_m \sim f(\mathbf{I}), m = 1, ..., M$$

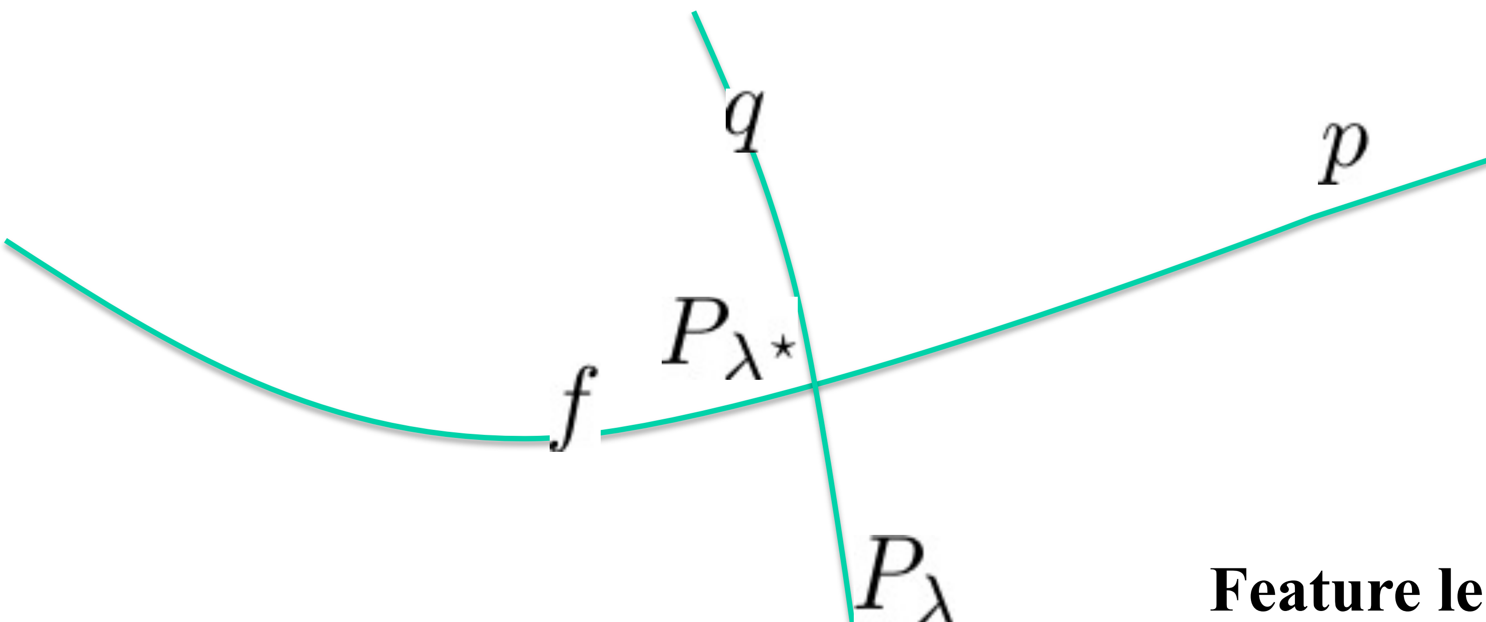
$$P(\mathbf{I}) = \frac{1}{Z(\lambda)} \exp\{\sum_i \lambda_i \phi_i(\mathbf{I})\} q(\mathbf{I})$$

$$\lambda^\star = \arg\min_\lambda D(f, P_\lambda)$$

Maximum Entropy

$$P_\lambda^\star = \arg\min_p D(p, q)$$

$$\text{subject to } \mathrm{E}_p[\phi_k(\mathbf{I})] = \mathrm{E}_f[\phi_k(\mathbf{I})], \forall k.$$



Feature learning?

Summary

Entropy of a distribution
- measures randomness or uncertainty
  - log of the number of equally likely choices
  - average number of coin flips
  - average length of prefix code
  - (Kolmogorov: shortest machine code → randomness)

Relative entropy from one distribution to the other
- measure the departure from the first to the second
  - coding redundancy
  - large deviation

Conditional entropy, mutual information, entropy rate