Probabilistic Graphical Models

Exponential family & Variational Inference I

Siamak Ravanbakhsh

Fall 2019

Learning objectives

- entropy
- exponential family distribution
  - duality in exponential family
- relationship between
  - two parametrizations
  - inference and learning as mapping between the two
  - relative entropy and two types of projections

A measure of information

a measure of information $I(X = x)$

- observing a **less probable** event gives **more information**
- information is non-negative and $I(X = x) = 0 \Leftrightarrow P(X = x) = 1$
- information from **independent events** is **additive**

$$A = a \perp B = b \Rightarrow I(A = a, B = b) = I(A = a) + I(B = b)$$

A measure of information

a measure of information $I(X = x)$

- observing a **less probable** event gives **more information**
- information is non-negative and $I(X = x) = 0 \Leftrightarrow P(X = x) = 1$
- information from **independent events** is **additive**

$$A = a \perp B = b \Rightarrow I(A = a, B = b) = I(A = a) + I(B = b)$$

definition follows from these characteristics:

$$I(X = x) \triangleq \log(\frac{1}{P(X=x)}) = -\log(P(X = x))$$

Entropy: information theory

information in obs. $X = x$ is $I(X = x) \triangleq -\log(P(X = x))$

entropy: expected amount of information

$$H(P) \triangleq \mathbb{E}[I(X)] = -\sum_{x \in Val(X)} P(X = x) \log(P(X = x))$$

Entropy: information theory

information in obs. $X = x$ is $I(X = x) \triangleq -\log(P(X = x))$

entropy: expected amount of information

$$H(P) \triangleq \mathbb{E}[I(X)] = -\sum_{x \in Val(X)} P(X = x) \log(P(X = x))$$

- achieves its maximum for uniform distribution $0 \leq H(P) \leq \log(|Val(X)|)$

Entropy: information theory

alternatively

expected (optimal) message length in reporting observed X

e.g., using Huffman coding

Entropy: information theory

alternatively

expected (optimal) message length in reporting observed X

e.g., using Huffman coding

$$Val(X) = \{a, b, c, d, e, f\}$$

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{16}, P(e) = P(f) = \frac{1}{32}$$

an optimal code for transmitting X:

Entropy: information theory

alternatively

expected (optimal) message length in reporting observed X

e.g., using Huffman coding

$$Val(X) = \{a, b, c, d, e, f\}$$

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{16}, P(e) = P(f) = \frac{1}{32}$$

an optimal code for transmitting X:

$a \rightarrow 0$

$b \rightarrow 10$

$c \rightarrow 110$

$d \rightarrow 1110$

$e \rightarrow 11110$

$f \rightarrow 11111$

average length?

$$H(P) = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{4}\log(\frac{1}{4}) - \frac{1}{8}\log(\frac{1}{8}) - \frac{1}{16}\log(\frac{1}{16}) - \frac{1}{16}\log(\frac{1}{32}) = 1\frac{15}{16}$$

$$\frac{1}{2} \qquad \frac{1}{2} \qquad \frac{3}{8} \qquad \frac{1}{4} \qquad \frac{5}{16}$$

$\downarrow$

contribution to the average length from X=a

Relative entropy: information theory

what if we used a code designed for *q*?

average cod length when transmitting $X \sim p$

is $H(p,q) \triangleq -\sum_{x \in Val(X)} p(x) \log(q(x))$

cross entropy

negative of the optimal code length for X=x according to q

Relative entropy: information theory

what if we used a code designed for *q*?

average cod length when transmitting $X \sim p$

is $H(p,q) \triangleq -\sum_{x \in Val(X)} p(x) \log(q(x))$

cross entropy

negative of the optimal code length for X=x according to q

the **extra** amount of information transmitted:

$$D(p\|q) \triangleq \sum_{x \in Val(X)} p(x)(\log(p(x)) - \log(q(x)))$$

Kullback-Leibler divergence or relative entorpy

Relative entropy: information theory

Kullback-Leibler divergence

$$D(p\|q) \triangleq \sum_{x\in Val(X)} p(x)(\log(q(x) - \log(p(x)))$$

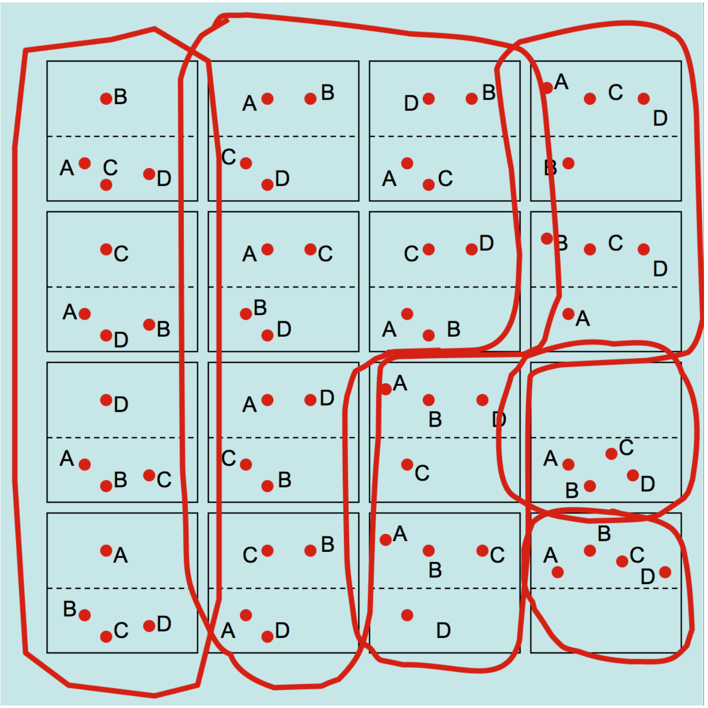
some properties:

non-negative and zero iff p=q

asymmetric

$$D(p\|u) = \sum_x p(x)(\log(p(x)) - \log(\frac{1}{N})) = \log(N) - H(p)$$

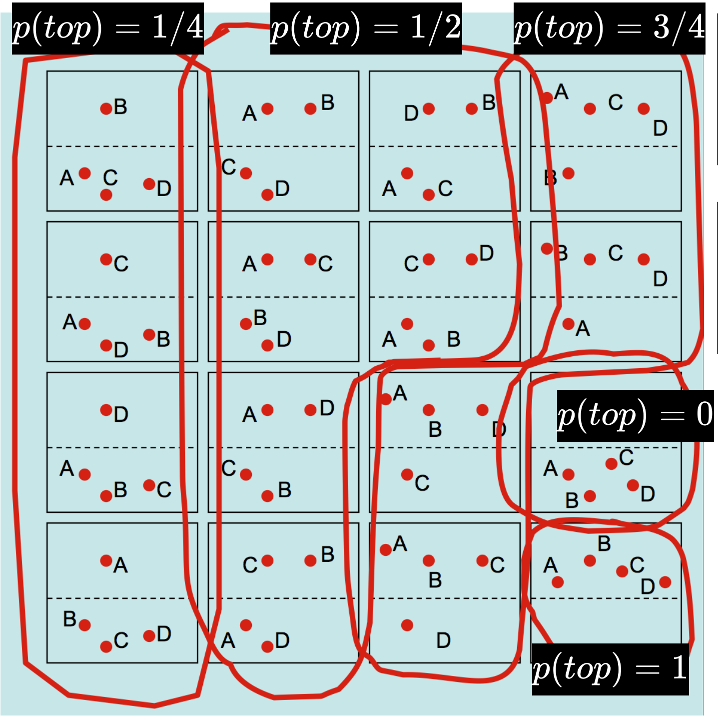
Entropy: physics



16 microstates: position of 4 particles in top/bottom box

5 macrostates: indistinguishable states assuming exchangeable particles

Entropy: physics



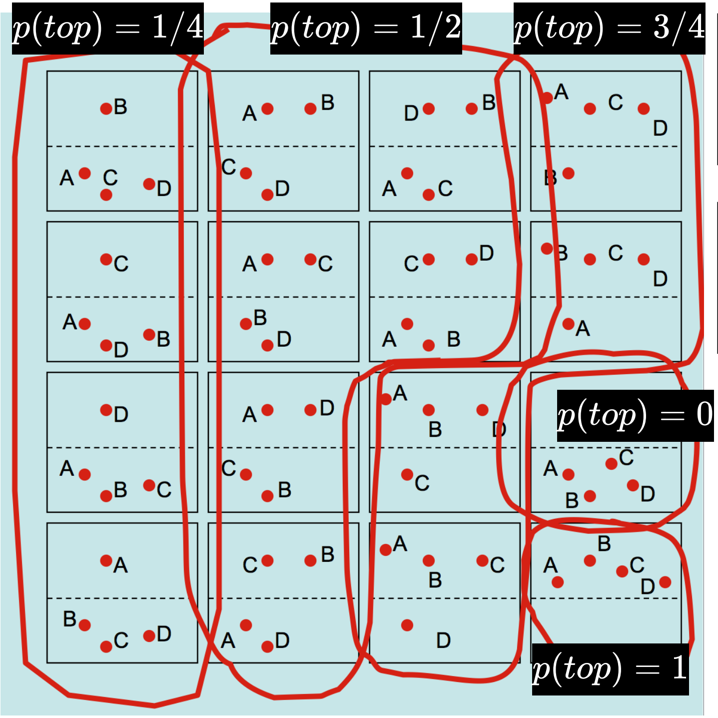
16 microstates: position of 4 particles in top/bottom box

5 macrostates: indistinguishable states assuming exchangeable particles

with $Val(X) = \{top, bottom\}$ we can assume

5 different distributions

Entropy: physics



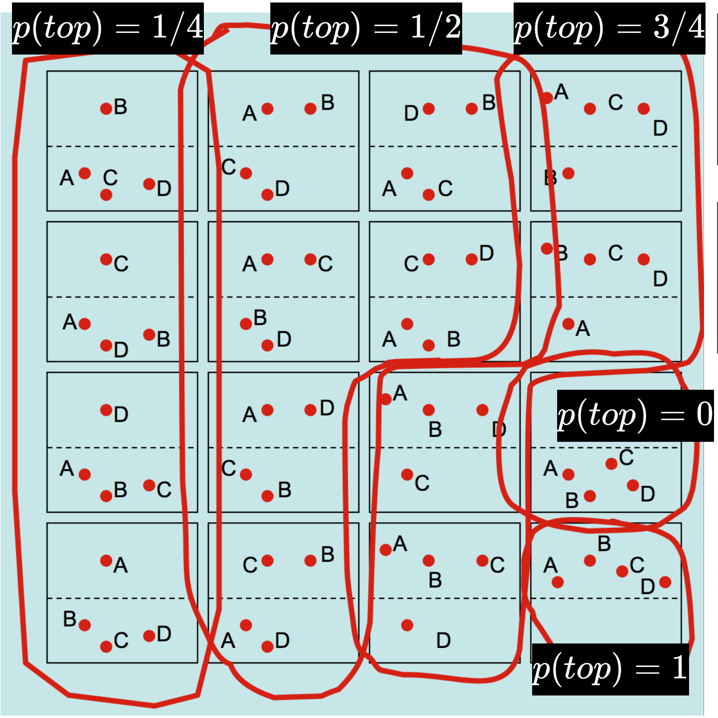
16 microstates: position of 4 particles in top/bottom box

5 macrostates: indistinguishable states assuming exchangeable particles

with $Val(X) = \{top, bottom\}$ we can assume 5 different distributions

each macrostate is a distribution

Entropy: physics



16 microstates: position of 4 particles in top/bottom box

5 macrostates: indistinguishable states assuming exchangeable particles

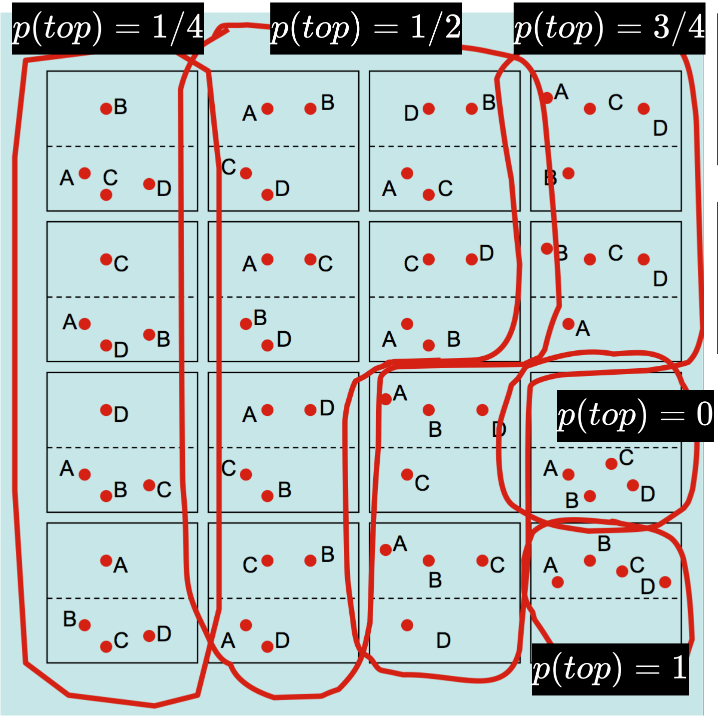
with $Val(X) = \{top, bottom\}$ we can assume

5 different distributions

each macrostate is a distribution

which distribution is more likely?

Entropy: physics



16 microstates: position of 4 particles in top/bottom box

5 macrostates: indistinguishable states assuming exchangeable particles

with $Val(X) = \{top, bottom\}$ we can assume

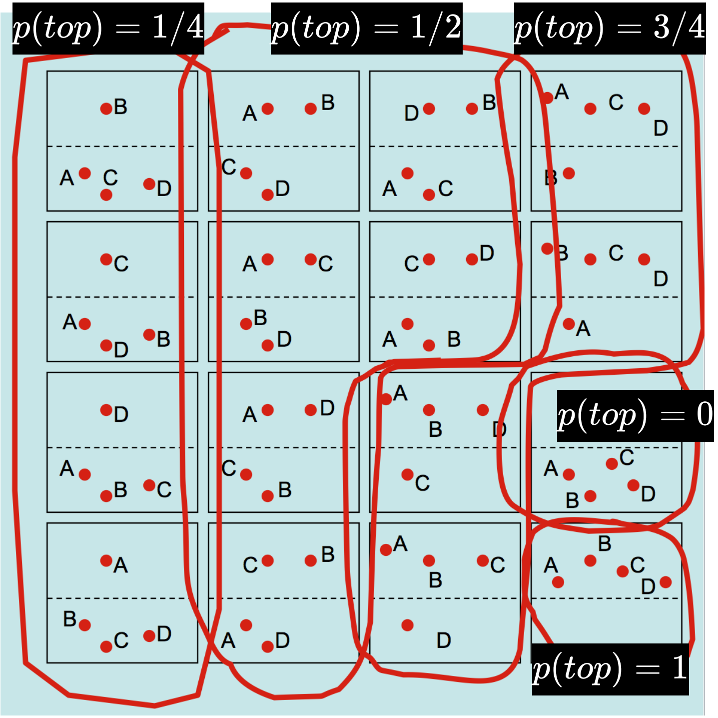
5 different distributions

each macrostate is a distribution

which distribution is more likely?

entropy of a macrostate: (normalized) log number of its microstates

Entropy: physics

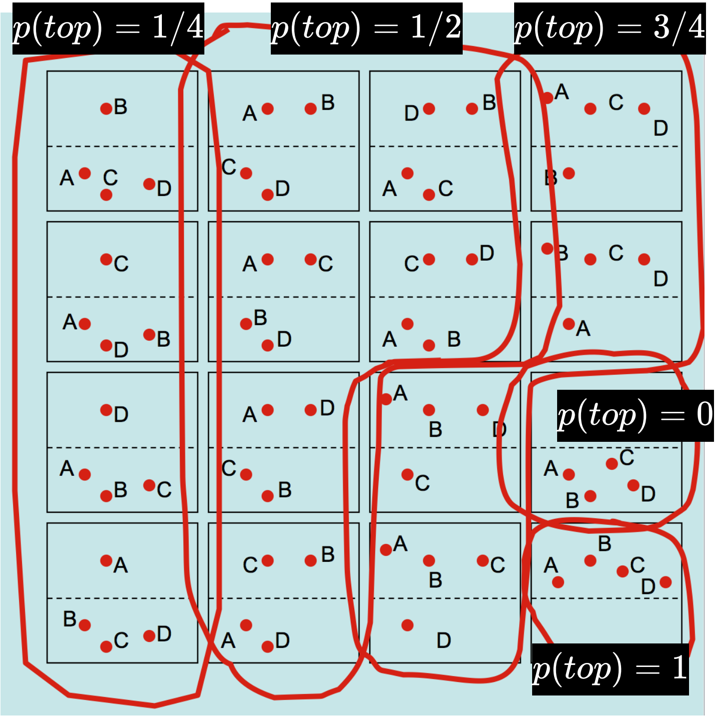


entropy of a macrostate: (normalized) log number of its microstates

assume a large number of particles N

$$H_{\text{macrostate}} = \frac{1}{N}\ln(\frac{N!}{N_t!N_b!}) = \frac{1}{N}\left(\ln(N!) - \ln(N_t!) - \ln(N_b!)\right))$$

Entropy: physics



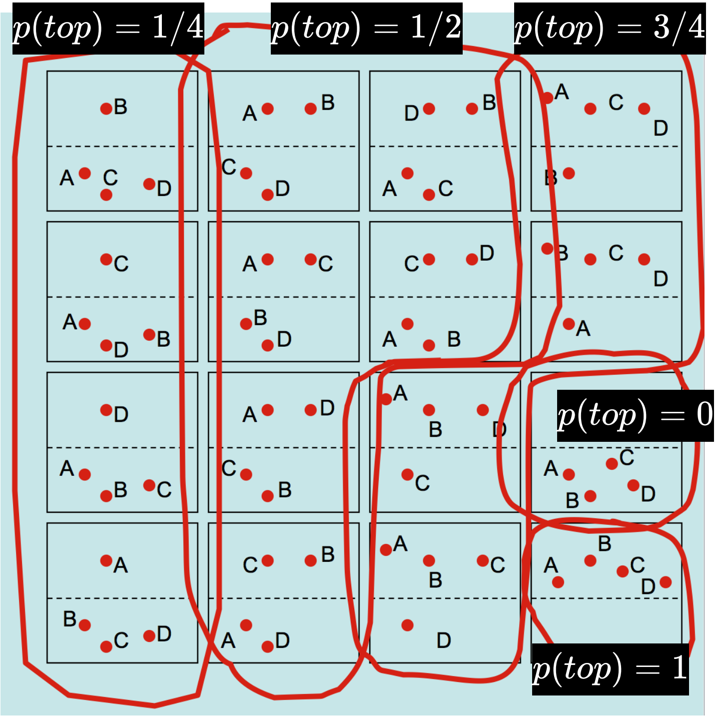
entropy of a macrostate: (normalized) log number of its microstates

assume a large number of particles N

$$H_{\text{macrostate}} = \frac{1}{N}\ln(\frac{N!}{N_t!N_b!}) = \frac{1}{N}(\ln(N!) - \ln(N_t!) - \ln(N_b!)))$$

$$\ln(N!) \simeq N\ln(N) - N$$

Entropy: physics



entropy of a macrostate: (normalized) log number of its microstates

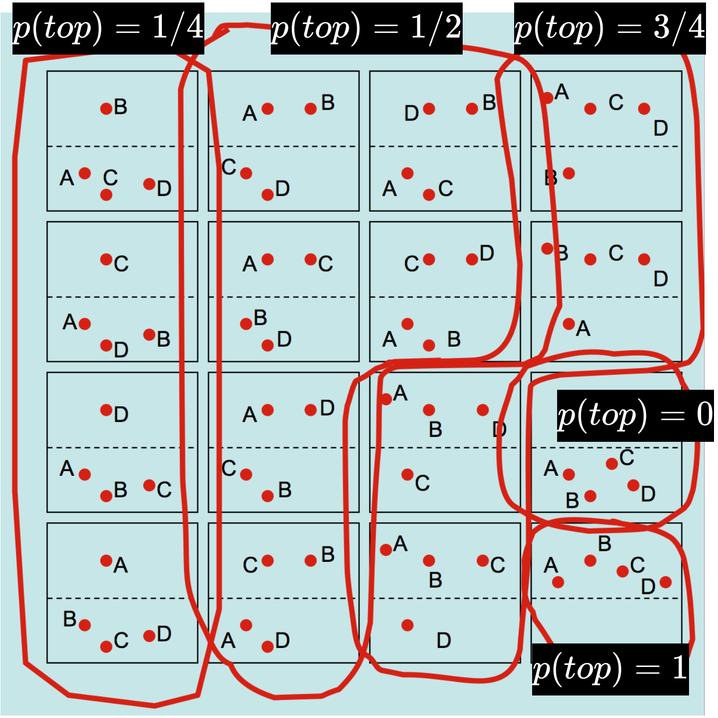
assume a large number of particles N

$$H_{\text{macrostate}} = \frac{1}{N}\ln\left(\frac{N!}{N_t!N_b!}\right) = \frac{1}{N}\left(\ln(N!) - \ln(N_t!) - \ln(N_b!)\right)$$

$$\ln(N!) \simeq N\ln(N) - N$$

$$= c - \frac{N_t}{N}\ln\left(\frac{N_t}{N}\right) - \frac{N_b}{N}\ln\left(\frac{N_b}{N}\right)$$

Entropy: physics



entropy of a macrostate: (normalized) log number of its microstates

assume a large number of particles N

$$H_{\text{macrostate}} = \frac{1}{N} \ln\left(\frac{N!}{N_t! N_b!}\right) = \frac{1}{N} \left(\ln(N!) - \ln(N_t!) - \ln(N_b!)\right)$$

$$\simeq N \ln(N) - N$$

$$= c - \frac{N_t}{N} \ln\left(\frac{N_t}{N}\right) - \frac{N_b}{N} \ln\left(\frac{N_b}{N}\right)$$

$$P(X = top)$$

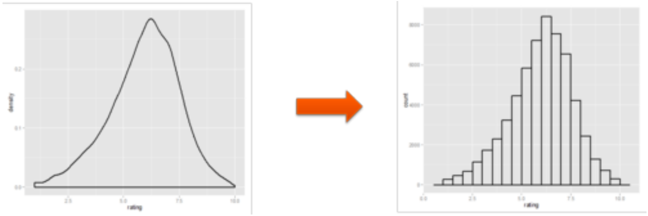
$$= - \sum_{x \in \{top, bottom\}} p(x) \ln(p(x))$$

Differential entropy for continuous domains

divide the domain $Val(X)$ using small bins of width Δ

$\exists x_i \in (\Delta i, \Delta(i+1))$

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\mathrm{d}x = p(x_i)\Delta$$



$$H_\Delta(p) = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\ln(\Delta) - \sum_i p(x_i)\Delta \ln(p(x_i))$$

ignore

take the limit $\Delta \to 0$ to get $H(p) \triangleq \int_{Val(x)} p(x)\ln(p(x))\mathrm{d}x$

max-entropy distribution

High entropy distribution:

- more information in observing X ~ p
- it's a more likely "macrostate"
- the least amount of assumption about p

max-entropy distribution

High entropy distribution:

- more information in observing X ~ p
- it's a more likely "macrostate"
- the least amount of assumption about p

when optimizing for p(x) subject to constrains, maximize the entropy

$$\arg\max_p H(p)$$

$$p(x) > 0 \quad \forall x$$

$$\int_{Val(X)} p(x) \mathrm{d}x = 1$$

$$\mathbb{E}_p[\phi_k(X)] = \mu_k \quad \forall k$$

max-entropy distribution

High entropy distribution:

- more information in observing X ~ p
- it's a more likely "macrostate"
- the least amount of assumption about p

when optimizing for p(x) subject to constrains, maximize the entropy

$$\arg\max_p H(p)$$

$$p(x) > 0 \quad \forall x$$

$$\int_{Val(X)} p(x)\mathrm{d}x = 1$$

$$\mathbb{E}_p[\phi_k(X)] = \mu_k \quad \forall k$$

$$p(x) \propto \exp(\textstyle\sum_k \theta_k \phi_k(x))$$

Lagrange multipliers

Exponential family

an exponential family has the following form

$$p(x;\theta) = h(x) \exp(\langle \eta(\theta), \phi(x) \rangle - A(\theta))$$

base measure

the inner product of two vectors

sufficient statistics

log-partition function

$$A(\theta) = \ln(\int_{Val(X)} h(x) exp(\sum_k \theta_k \phi_k(x)) dx)$$

with a convex parameter space $\theta \in \Theta = \{\theta \in \Re^D \mid A(\theta) < \infty\}$

Example: univariate Gaussian

moment form: $p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$

$$p(x;\theta) = h(x)\exp(\langle \eta(\theta), \phi(x)\rangle - A(\theta))$$

$\theta \to [\mu, \sigma^2]$

$h(x) \to 1$

$\eta(\theta) \to \eta(\mu,\sigma^2) = [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]$

$\phi(x) \to [x, x^2]$

$A(\theta) \to \frac{1}{2}(\ln(2\pi\sigma^2) + \frac{\mu^2}{\sigma^2})$

for $\mu, \sigma^2 \in \Re \times \Re^+$

Example: Bernoulli

conventional form (mean parametrization) $p(x;\mu) = \mu^{x}(1-\mu)^{1-x}$

$$p(x;\mu) = h(x)\exp(\langle \eta(\theta), \phi(x)\rangle - A(\theta))$$

$h(x) = 1$

$\eta(\mu) = [\ln(\mu), \ln(1-\mu)]$

$\phi(x) = [\mathbb{I}(x=1), \mathbb{I}(x=0)]$

$A(\theta) = 1$

for $\mu \in (0,1)$

Linear exponential family

when using natural parameters

$$p(x;\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

natural parameters

simply define $\eta(\theta)$ to be the new θ ?

natural parameter-space needs to be convex

$$\theta \in \Theta = \{\theta \in \Re^D \mid A(\theta) < \infty\}$$

Linear exponential family

when using natural parameters

$$p(x;\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

can absorb it as a sufficient stat. with $\theta = 1$

natural parameters simply define $\eta(\theta)$ to be the new θ ?

natural parameter-space needs to be convex

$$\theta \in \Theta = \{\theta \in \Re^D \mid A(\theta) < \infty\}$$

Example: univariate Gaussian

take 2

natural parameters in the univariate Gaussian

$$p(x;\theta) = \exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

$\theta \rightarrow [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]$

$\phi(x) \rightarrow [x, x^2]$

$A(\theta) \rightarrow \frac{-1}{2}(\ln(\theta_2/\pi) + \frac{\theta_1^2}{2\theta_2})$?

where $\theta \in \Re \times \Re^-$ is a convex set

Example: Bernoulli

take 2

conventional form (mean parametrization) $p(x;\mu) = \mu^x (1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

θ → $[\ln(\mu), \ln(1-\mu)]$

$\phi(x)$ → $[\mathbb{I}(x=1), \mathbb{I}(x=0)]$

Example: Bernoulli

take 2

conventional form (mean parametrization) $p(x;\mu) = \mu^x (1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

θ: $[\ln(\mu), \ln(1-\mu)]$

$\phi(x)$: $[\mathbb{I}(x=1), \mathbb{I}(x=0)]$

however Θ is not a convex set ☹

Example: Bernoulli

take 3

conventional form (mean parametrization) $p(x;\mu) = \mu^x(1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle\theta,\phi(x)\rangle - A(\theta))$$

$\theta \in \Re^2$

$\phi(x) = [\mathbb{I}(x=1), \mathbb{I}(x=0)]$

Example: Bernoulli

take 3

conventional form (mean parametrization) $p(x;\mu) = \mu^x(1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle\theta,\phi(x)\rangle - A(\theta))$$

$\theta \in \Re^2$

$\phi(x) = [\mathbb{I}(x=1), \mathbb{I}(x=0)]$

this parametrization is redundant or **overcomplete**

$$p(x,[\theta_1,\theta_2]) = p(x,[\theta_1+c,\theta_2+c])$$

redundant iff $\quad \exists\theta \quad \text{s.t.} \quad \forall x \quad \langle\theta,\phi(x)\rangle = c$

Example: Bernoulli

take 4

conventional form (mean parametrization) $p(x;\mu) = \mu^x (1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$$

$\theta \rightarrow [\ln \frac{\mu}{1-\mu}]$

$\phi(x) \rightarrow [\mathbb{I}(x=1)]$

$A(\theta) \rightarrow \log(1+e^\theta)$

Example: Bernoulli

take 4

conventional form (mean parametrization) $p(x;\mu) = \mu^x (1-\mu)^{1-x}$

$$p(x;\theta) = h(x)\exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

$\theta \rightarrow [\ln \frac{\mu}{1-\mu}]$

$\phi(x) \rightarrow [\mathbb{I}(x=1)]$

$A(\theta) \rightarrow \log(1+e^{\theta})$

Θ is convex and this parametrization is **minimal** ☺

Example: categorical distribution

more generally $p(x;\mu) = \prod_d \mu_d^{\mathbb{I}(x=d)}$

has a minimal linear exp-family form

$$p(x;\theta) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

$\theta \rightarrow [\ln \frac{\mu_2}{\mu_1}, \ldots, \ln \frac{\mu_D}{\mu_1}]$

$\phi(x) \rightarrow [\mathbb{I}(x=2), \ldots, \mathbb{I}(x=D)]$

Example: Beta distribution

motivation: when discussing Bayesian inference

for shape parameters $\alpha, \beta \in \Re^+ \times \Re^+$

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

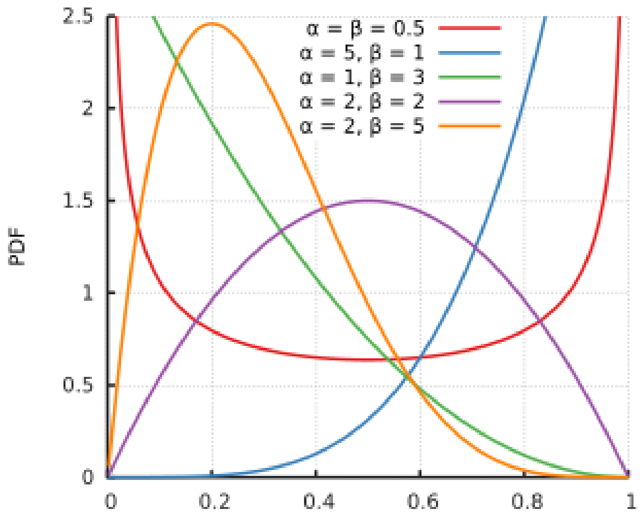


image: wikipedia

linear exp-family form

$$p(x; \theta) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

$\theta \rightarrow [\alpha - 1, \beta - 1]$

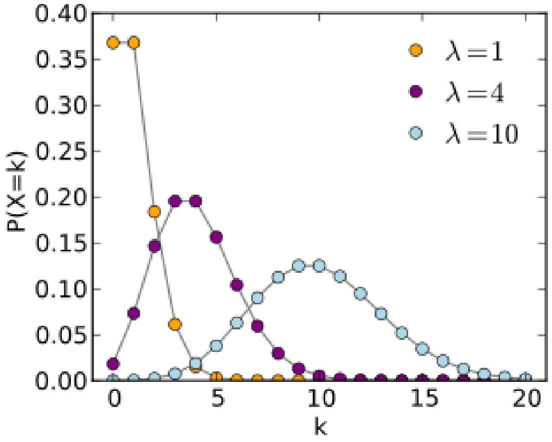
$\phi(x) \rightarrow [\ln(x), \ln(1-x)]$

where $\theta \in (-1, +\infty) \times (-1, +\infty)$

Example: Poisson distribution

Poisson: $p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ where $\lambda > 0$ is the *mean frequency*
(rate parameter)

- probability of x events happening in a fixed period
- events happen independently with the rate λ
- similar to binomial with large number of trials $(\lambda \approx n\mu)$



Example: Poisson distribution

for the rate parameter $\lambda \in \Re^+$

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

linear exp-family form

$$p(x;\theta) = h(x)\exp(\langle\theta,\phi(x)\rangle - A(\theta))$$

$\theta \to \ln(\lambda)$, $h(x) \to \frac{1}{x!}$, $\phi(x) \to x$, $A(\theta) \to \exp(\theta)$

where $\theta \in \Re$

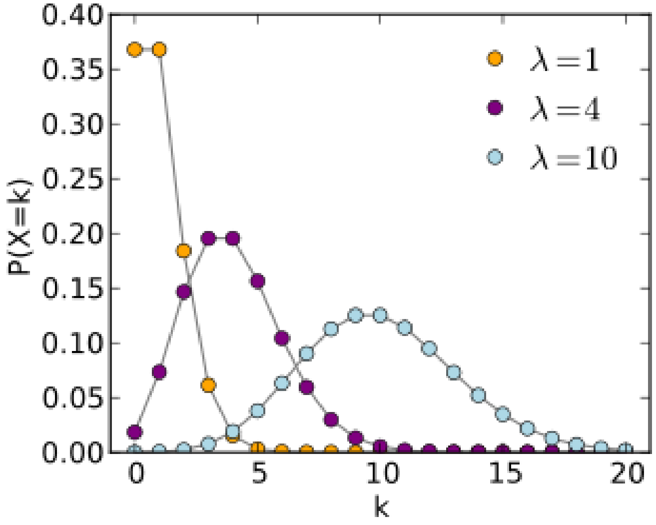


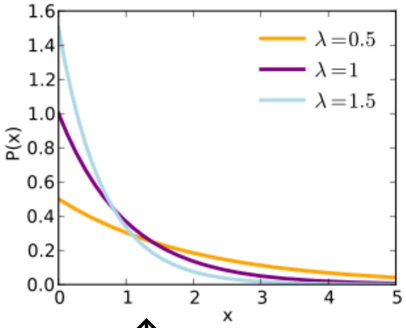
image: wikipedia

Example: exponential distribution

Exponential: $p(x; \lambda) = \lambda e^{-\lambda x} \quad \text{where} \quad \lambda > 0$

- time between events in Poisson dist.
- memoryless property

$Val(X) = \mathbb{R}^+$

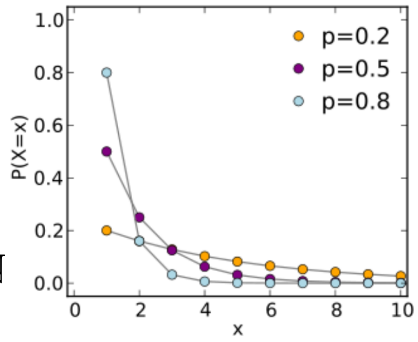


$(1 - \mu) \equiv e^{-\lambda}$

Geometric: $p(x, k; \mu) = (1 - \mu)^{k-1} \mu \quad \text{where} \quad 0 < \mu < 1$

- number of Bernoulli trials until success
- memoryless property

$Val(X) = \mathbb{N}$



Example: exponential distribution

for the rate parameter $\lambda \in \Re^+$

$$p(x; \lambda) = \lambda e^{-\lambda x}$$

linear exp-family form

$$p(x; \theta) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$

$\theta \rightarrow -\lambda$, $h(x) \rightarrow 1$, $\phi(x) \rightarrow x$, $A(\theta) \rightarrow -\ln(-\theta)$

where $\theta \in \Re$

max-entropy interpretation?

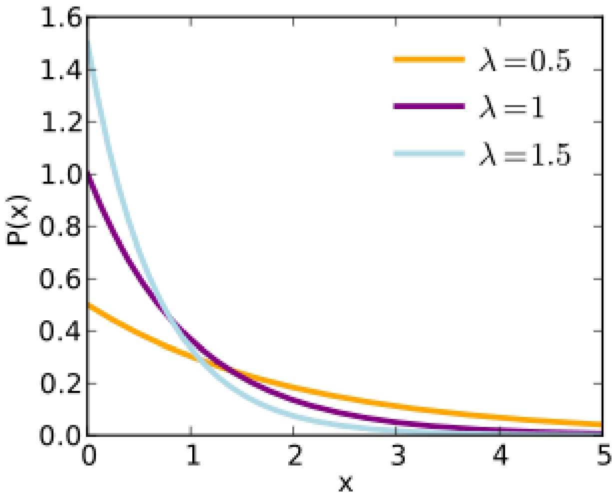


image: wikipedia

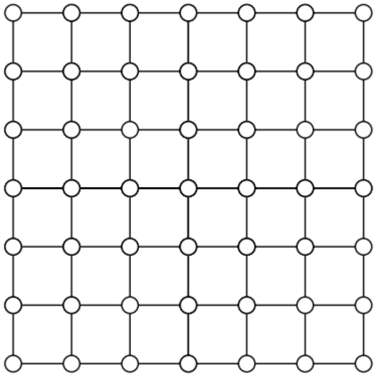
Example: Ising model

pairwise MRF with binary variables $x_i \in \{0, 1\}$

$$p(x;\theta) = \exp(-\sum_{i,j\leq i} \theta_{i,j} x_i x_j - A(\theta))$$

for i = j this encodes the local field

where $\theta \in \Re$



2D Ising grid

image: wainwright&jordan

Example: mixture models

X is discrete and $p(x, y) = p(x)p(y \mid x)$

$[y; y^2]$

for mixture of Gaussians

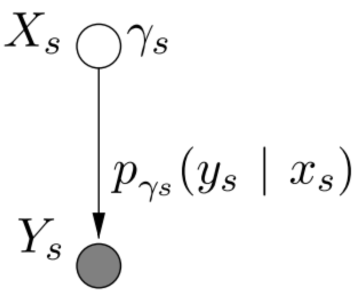
sufficient statistics: $[\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = D)]$

natural parameters:

$$\theta = [\theta_1, \ldots, \theta_D, \frac{\mu_1}{\sigma_1^2}, \ldots, \frac{\mu_D}{\sigma_D^2}, \frac{-1}{\sigma_1^2}, \ldots, \frac{-1}{\sigma_D^2}]$$

overcomplete parametrization for p(x)

natural params for each component in the mixture



more genral forms

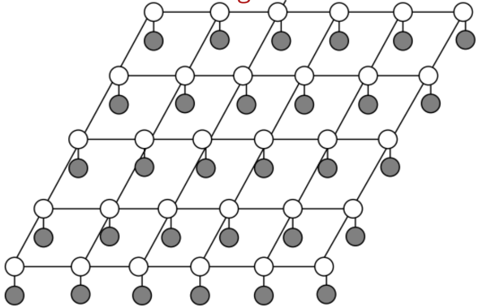


image: wainwright&jordan

Example: general Markov networks

log-linear form for **positive dists.**

$$p(x;\theta) = \exp(\sum_k \theta_k \phi_k(\mathbf{D}_k) - A(\theta))$$

cliques in the the undirected graph

where $\theta \in \Re$

$$\ln(\sum_{x \in Val(X)} \exp(-\sum_k \theta_k \phi_k(\mathbf{D}_k)))$$

familiar log-sum-exp form

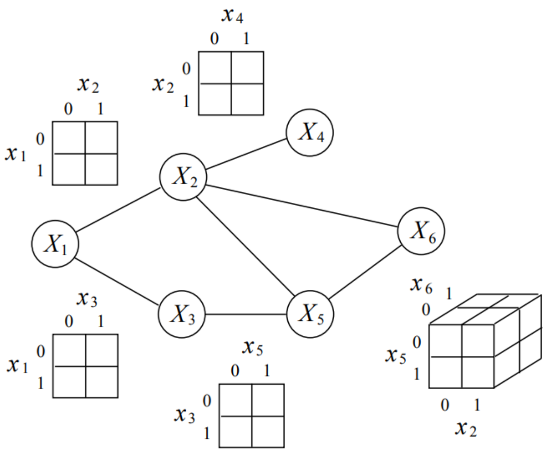


image: Michael Jordan's draft

Markov networks as exponential family

Discrete distributions

$$p(x;\theta) = \exp(\textstyle\sum_k \theta_k \phi_k(\mathbf{D}_k) - A(\theta))$$

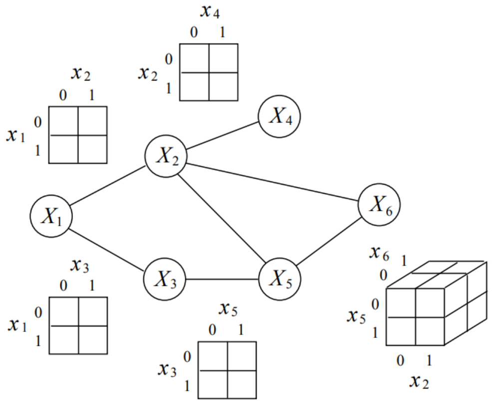


image: Michael Jordan's draft

Mean parameters are the marginals

mean parameters		natural params.		sufficient statistics
$\mu_{1,2,0,0} = P(X_1 = 0, X_2 = 0)$	$\longleftrightarrow$	$\theta_{1,2,0,0}$	—	$\mathbb{I}(X_1 = 0, X_2 = 0)$
$\mu_{1,2,1,0} = P(X_1 = 1, X_2 = 0)$	$\longleftrightarrow$	$\theta_{1,2,1,0}$	—	$\mathbb{I}(X_1 = 1, X_2 = 0)$
$\mu_{1,2,0,1} = P(X_1 = 0, X_2 = 1)$	$\longleftrightarrow$	$\theta_{1,2,0,1}$	—	$\mathbb{I}(X_1 = 0, X_2 = 1)$
$\mu_{1,2,1,1} = P(X_1 = 1, X_2 = 1)$	$\longleftrightarrow$	$\theta_{1,2,1,1}$	—	$\mathbb{I}(X_1 = 1, X_2 = 1)$

Mean parametrization

natural parameter θ $\Rightarrow$ mean parameter $\mu = \mathbb{E}_{p_\theta}[\phi(x)]$

one-to-one mapping $\Leftarrow$ if *minimal* sufficiant statistics

$$\theta \in \Theta \quad \Leftrightarrow \quad \mu \in \mathcal{M} = \{\mathbb{E}_p[\phi(x)] \quad \forall p\}$$

any distribution p

mean parameter space

$\mathcal{M}$ is also convex why?

Mean parametrization: example

Multivariate Gaussian

natural parameter θ $\Longrightarrow$ mean parameter $\mu = \mathbb{E}_{p_\theta}[\phi(x)]$

$$\eta = \Sigma^{-1}\mu, \quad \Lambda = \Sigma^{-1} \quad \Longleftrightarrow \quad \mu = \Lambda^{-1}\eta, \quad \Sigma - \mu\mu^T$$

$$\downarrow \qquad \downarrow$$

sufficient statistics: $\phi_1(X) = X, \phi_2(X) = X^2$

Mean parametrization: example

Multivariate Gaussian

natural parameter θ $\Longrightarrow$ mean parameter $\mu = \mathbb{E}_{p_\theta}[\phi(x)]$

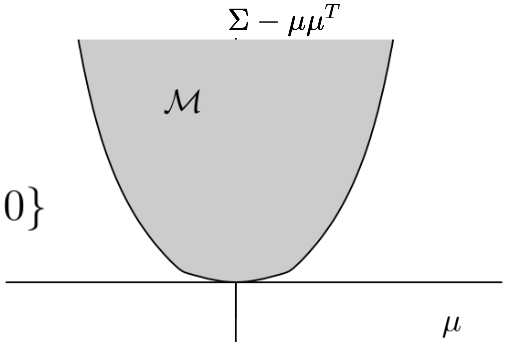
$\eta = \Sigma^{-1}\mu, \quad \Lambda = \Sigma^{-1}$ $\Longleftrightarrow$ $\mu = \Lambda^{-1}\eta, \quad \Sigma - \mu\mu^T$

$\downarrow \qquad \downarrow$

sufficient statistics: $\phi_1(X) = X, \phi_2(X) = X^2$

$\mathcal{M}, \Theta$ are both convex

$$\mathcal{M} = \{(\mu, \Sigma) \in \mathbb{R}^m \times \mathcal{S}_+^m \mid \Sigma - \mu\mu^T \succeq 0\}$$



Marginal polytope

for variables with finite domain: $Val(X)$

mean parameter space is a convex polytope

$$\mathcal{M} = \{\mathbb{E}_p[\phi(x)] \quad \forall p\} = conv\{\phi(x)\ \forall x\}$$

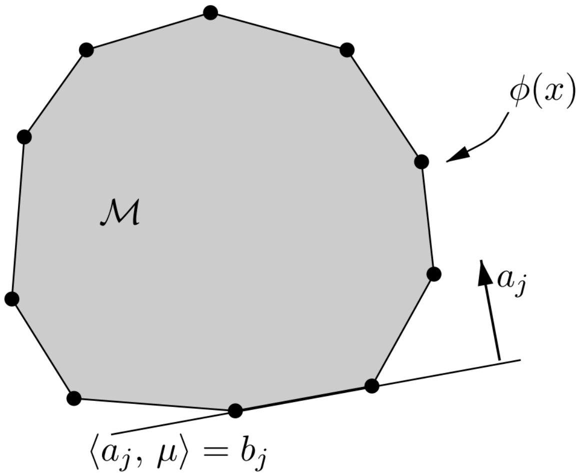
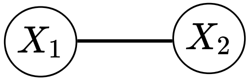
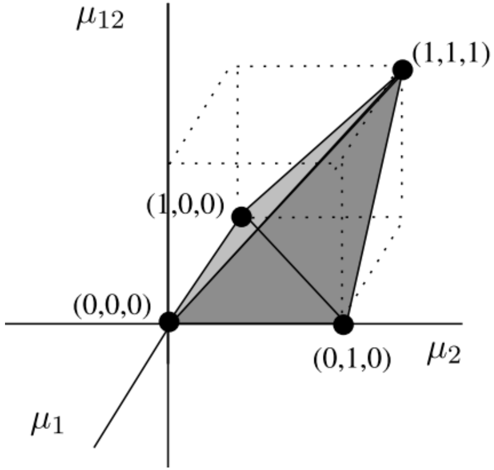


image: wainwright &jordan

Marginal polytope: example

2 variables $X_1, X_2 \in \{0,1\}$





sufficient statistics

$\mathbb{I}[X_1 = 1], \mathbb{I}[X_2 = 1], \mathbb{I}(X_1 = 1, X_2 = 1)$

mean parameters

$\mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2], \mu_{1,2} = \mathbb{E}[X_1 X_2]$

marginal polytope

$\mathcal{M} = \{\mathbb{E}_p[\phi(x)] \quad \forall p\} = conv\{(0,0,0),(0,1,0),(1,0,0),(1,1,1)\}$

image: wainwright &jordan

Summary so far...

- motivate entropy from *physics* and *information theory*
- derivation of exponential family using entropy
- examples:
  - famous univariate distributions
  - minimal & overcomplete discrete MRF
  - multivariate Gaussian
- expected sufficient statistics and natural parameters
  - identify the same distribution

Significance of μ and θ

Inference $\theta \Rightarrow \mu = \mathbb{E}_{p_\theta}[\phi(x)]$

- for $\phi_k(x) = \mathbb{I}(x_i = r, x_j = s)$ mean parameter are marginals

Significance of μ and θ

Inference $\theta \Rightarrow \mu = \mathbb{E}_{p_\theta}[\phi(x)]$

- for $\phi_k(x) = \mathbb{I}(x_i = r, x_j = s)$ mean parameter are marginals

Learning $\mu \Rightarrow \theta \quad s.t. \quad \mathbb{E}_{p_\theta}[\phi(x)] = \mu$

- given samples $X_1, X_2, \ldots, X_n \sim p_\theta$
- calculate expected sufficient statistics $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$
- find $\theta \quad s.t. \quad \mathbb{E}_{p_\theta}[\phi(x)] = \hat{\mu}$

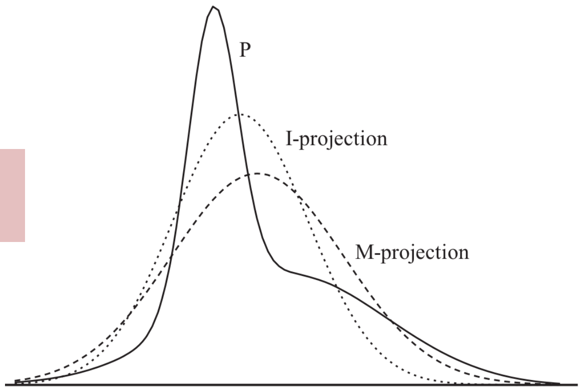
Projections

Project p into a convex set of dists. $\mathcal{Q}$

I-projection $q^I \triangleq \arg\min_{q \in \mathcal{Q}} D(q \| p)$

(information projection)

$$-H(q) + \mathbb{E}_q[-\ln(p)]$$



Projections

Project p into a convex set of dists. $\mathcal{Q}$

I-projection $q^I \triangleq \arg\min_{q \in \mathcal{Q}} D(q \| p)$

(information projection)

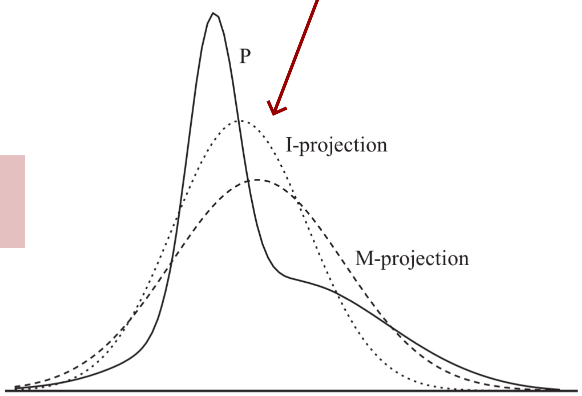
$-H(q) + \mathbb{E}_q[-\ln(p)]$

M-projection $q^M \triangleq \arg\min_{q \in \mathcal{Q}} D(p \| q)$

(moment projection)

$-\mathbb{E}_p[\ln q]$

mode-seeking behavior



Projections: example

$p(a^0, b^0) = .45$

$p(a^0, b^1) = .05$

$p(a^1, b^0) = .05$

$p(a^1, b^1) = .45$

project into a q with **factorized** form $q(a, b) = q(a)q(b)$

Projections: example

$$p(a^0, b^0) = .45$$
$$p(a^0, b^1) = .05$$
$$p(a^1, b^0) = .05$$
$$p(a^1, b^1) = .45$$

project into a q with **factorized** form $q(a, b) = q(a)q(b)$

M-projection:

$$q^M(a^0) = q^M(a^1) = .5$$
$$q^M(b^0) = q^M(b^1) = .5$$

Projections: example

$$p(a^0, b^0) = .45$$
$$p(a^0, b^1) = .05$$
$$p(a^1, b^0) = .05$$
$$p(a^1, b^1) = .45$$

project into a q with **factorized** form $q(a, b) = q(a)q(b)$

M-projection:

$$q^M(a^0) = q^M(a^1) = .5$$
$$q^M(b^0) = q^M(b^1) = .5$$

I-projection:

$$q^I(a^0) = q^I(b^0) = .25$$
$$q^I(a^1) = q^I(b^1) = .75$$

mode-seeking behavior

M-Projection

M-projection of p into a q with **factorized** form $q(x) = \prod_k q(x_k)$

and otherwise unrestricted

gives $q^M(x) = \prod_k p(x_k)$

M-Projection

M-projection of p into a q with **factorized** form $q(x) = \prod_k q(x_k)$

and otherwise unrestricted

gives $q^M(x) = \prod_k p(x_k)$

Proof $D(p\|q) = \mathbb{E}_p[\ln p(x)] - \sum_k \mathbb{E}_p[\ln q(x_k)]$

M-Projection

M-projection of p into a q with **factorized** form $q(x) = \prod_k q(x_k)$

and otherwise unrestricted

gives $q^M(x) = \prod_k p(x_k)$

Proof

$$D(p\|q) = \mathbb{E}_p[\ln p(x)] - \sum_k \mathbb{E}_p[\ln q(x_k)]$$

$$= \mathbb{E}_p[\ln \frac{p(x)}{\prod_k p(x_k)}] + \sum_k \mathbb{E}_p[\ln \frac{p(x_k)}{q(x_k)}]$$

M-Projection

M-projection of p into a q with **factorized** form $q(x) = \prod_k q(x_k)$

and otherwise unrestricted

gives $q^M(x) = \prod_k p(x_k)$

Proof

$$D(p\|q) = \mathbb{E}_p[\ln p(x)] - \sum_k \mathbb{E}_p[\ln q(x_k)]$$

$$= \mathbb{E}_p[\ln \frac{p(x)}{\prod_k p(x_k)}] + \sum_k \mathbb{E}_p[\ln \frac{p(x_k)}{q(x_k)}]$$

$$= D(p\|q^M) + \sum_k D(p(x_k)\|q(x_k))$$

minimized when this is zero! $q = q^M$

M-Projection: exponential family

M-projection of p into a $q_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$

is given by moment-matching $\mathbb{E}_{q_\theta}[\phi(x)] = \mathbb{E}_p[\phi(x)]$

M-Projection: exponential family

M-projection of p into a $q_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$

is given by moment-matching $\mathbb{E}_{q_\theta}[\phi(x)] = \mathbb{E}_p[\phi(x)]$

Proof

consider two distributions:

- q_θ has the same moments as p
- $q_{\theta'}$ has different moments

M-Projection: exponential family

M-projection of p into a $q_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$

is given by moment-matching $\mathbb{E}_{q_\theta}[\phi(x)] = \mathbb{E}_p[\phi(x)]$

Proof

consider two distributions:

- q_θ has the same moments as p
- $q_{\theta'}$ has different moments

$$D(p \| q_{\theta'}) - D(p \| q_\theta) = \langle \mathbb{E}_p[\phi(x)], \theta - \theta' \rangle - A(\theta) + A(\theta')$$

M-Projection: exponential family

M-projection of p into a $q_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$

is given by moment-matching $\mathbb{E}_{q_\theta}[\phi(x)] = \mathbb{E}_p[\phi(x)]$

Proof consider two distributions:

- q_θ has the same moments as p
- $q_{\theta'}$ has different moments

$$D(p \| q_{\theta'}) - D(p \| q_\theta) = \langle \mathbb{E}_p[\phi(x)], \theta - \theta' \rangle - A(\theta) + A(\theta')$$

$$= \langle \mathbb{E}_{q_\theta}[\phi(x)], \theta - \theta' \rangle - A(\theta) + A(\theta') = D(q_\theta \| q_{\theta'}) \geq 0$$

so q_θ is the projection

M-Projection: exponential family

M-projection of p into a $q_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$

is given by moment-matching $\mathbb{E}_{q_\theta}[\phi(x)] = \mathbb{E}_p[\phi(x)]$

Proof

consider two distributions:

- q_θ has the same moments as p
- $q_{\theta'}$ has different moments

$$D(p \| q_{\theta'}) - D(p \| q_\theta) = \langle \mathbb{E}_p[\phi(x)], \theta - \theta' \rangle - A(\theta) + A(\theta')$$

$$= \langle \mathbb{E}_{q_\theta}[\phi(x)], \theta - \theta' \rangle - A(\theta) + A(\theta') = D(q_\theta \| q_{\theta'}) \geq 0$$

so q_θ is the projection

M-projection produces a distribution with the same moments

(note that p can have any form)

Projections, inference & learning

Information projection $\arg\min_{q\in\mathcal{Q}} D(q\|p) = \arg\min_{q\in\mathcal{Q}} \mathbb{E}_q[-\ln(p)] - H(q)$

exponential family form: $A(\theta) = \max_{\mu\in\mathcal{M}} \langle\mu,\theta\rangle - A^*(\mu)$

negative energy negative entropy

Projections, inference & learning

Information projection $\arg\min_{q\in\mathcal{Q}} D(q\|p) = \arg\min_{q\in\mathcal{Q}} \mathbb{E}_q[-\ln(p)] - H(q)$

exponential family form: $A(\theta) = \max_{\mu\in\mathcal{M}} \langle\mu,\theta\rangle - A^*(\mu)$

negative energy negative entropy

variational inference: inference as divergence optimization

but we saw that M-projection gives correct marginals, why use I-projection?

maximum likelihood learning of parameters from data

ideas based on moment-matching are also applied to inference

Projections, inference & learning

Information projection

$$\arg\min_{q\in\mathcal{Q}} D(q\|p) = \arg\min_{q\in\mathcal{Q}} \mathbb{E}_q[-\ln(p)] - H(q)$$

exponential family form: $A(\theta) = \max_{\mu\in\mathcal{M}} \langle\mu,\theta\rangle - A^*(\mu)$

negative energy — negative entropy

variational inference: inference as divergence optimization

but we saw that M-projection gives correct marginals, why use I-projection?

Moment projection

$$\arg\min_{q\in\mathcal{Q}} D(p\|q) = \mathbb{E}_p[-\ln(q)]$$

aka moment matching

$$A^*(\mu) = \max_{\theta\in\Theta} \langle\mu,\theta\rangle - A(\theta)$$

likelihood

maximum likelihood learning of parameters from data

ideas based on moment-matching are also applied to inference

Summary

- intuition for **entropy** & relative entropy
- examples of **linear** exponential family
- mean & natural **parametrization**
- **inference** and **learning** as a mapping between the two
    - relation to information and moment **projections**

bonus slides

Duality in exponential family

- consider log-partition function $A(\theta) = \log \int_{Val(X)} \exp(\langle \theta, \phi(x) \rangle) dx$
- its derivative gives the mean parameter

$$\nabla_\theta A(\theta) = \int_{Val(X)} p_\theta(x) \phi(x) dx = \mu$$

Duality in exponential family

- consider log-partition function $A(\theta) = \log \int_{Val(X)} \exp(\langle \theta, \phi(x) \rangle) dx$
- its derivative gives the mean parameter

$$\nabla_\theta A(\theta) = \int_{Val(X)} p_\theta(x)\phi(x)dx = \mu$$

- it is **convex** and its **conjugate dual** is negative entropy

$$-H(p_{\theta(\mu)}) = A^*(\mu) = \max_{\theta \in \Theta} \langle \mu, \theta \rangle - A(\theta)$$

$$A(\theta) = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu)$$

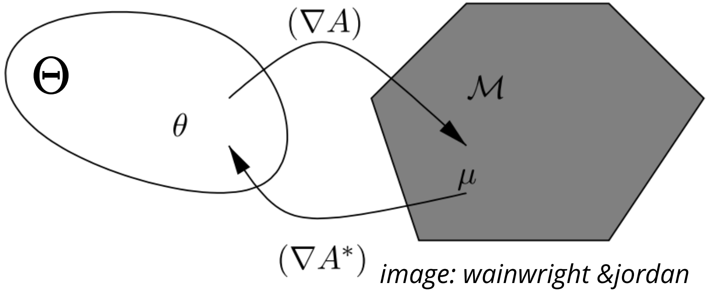


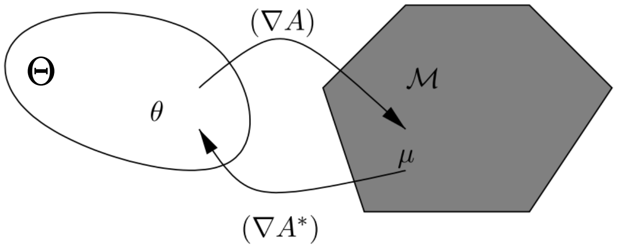
image: wainwright &jordan

Conjugate duality: example

Bernoulli

$$p(x,\theta) = \exp(\theta x - \underbrace{\ln(1 + exp(\theta))}_{A(\theta)})$$

$$\Theta = \Re$$

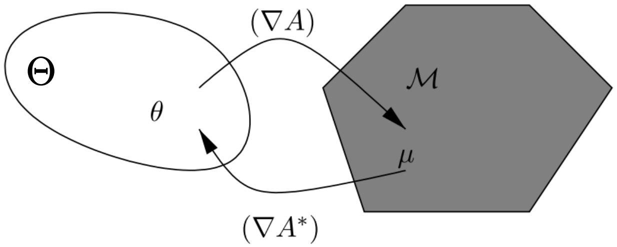


Conjugate duality: example

Bernoulli $p(x, \theta) = \exp(\theta x - \ln(1 + exp(\theta)))$ $\Theta = \Re$

$A(\theta)$

forward mapping: $\nabla_\theta A(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$ mean parameter



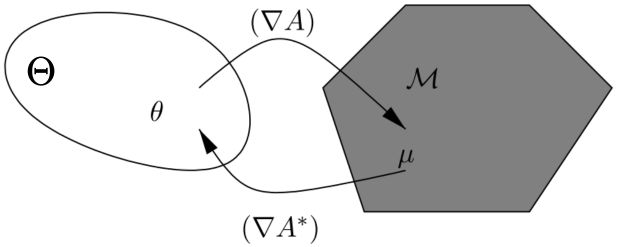
Conjugate duality: **example**

Bernoulli $p(x,\theta) = \exp(\theta x - \ln(1 + exp(\theta)))$ $\Theta = \Re$

$A(\theta)$

forward mapping: $\nabla_\theta A(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$ mean parameter

conjuage dual: $A^*(\mu) = \max_{\theta \in \Re} \langle \mu, \theta \rangle - \ln(1 + \exp(\theta))$



Conjugate duality: example

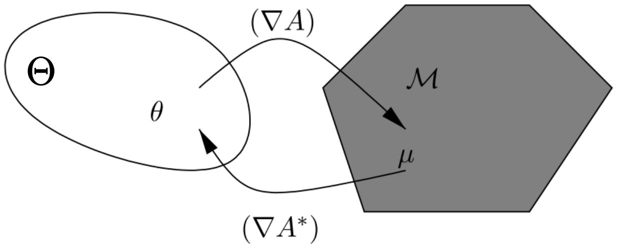
Bernoulli $p(x, \theta) = \exp(\theta x - \ln(1 + exp(\theta)))$ $\Theta = \Re$

$A(\theta)$

forward mapping: $\nabla_\theta A(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$ mean parameter

conjuage dual: $A^*(\mu) = \max_{\theta \in \Re} \langle \mu, \theta \rangle - \ln(1 + \exp(\theta))$

substitute $\theta = \frac{\ln(\mu)}{\ln(1-\mu)}$ *backward mapping*



Conjugate duality: example

Bernoulli $p(x,\theta) = \exp(\theta x - \underbrace{\ln(1 + exp(\theta))}_{A(\theta)})$ $\Theta = \Re$

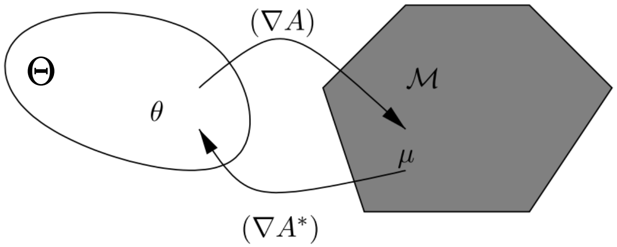
forward mapping: $\nabla_\theta A(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$ mean parameter

conjuage dual: $A^*(\mu) = \max_{\theta\in\Re}\langle\mu,\theta\rangle - \ln(1+\exp(\theta))$

substitute $\theta = \frac{\ln(\mu)}{\ln(1-\mu)}$ *backward mapping*

$A^*(\mu) = \mu\ln(\mu) + (1-\mu)\ln(1-\mu)$

negative entropy!

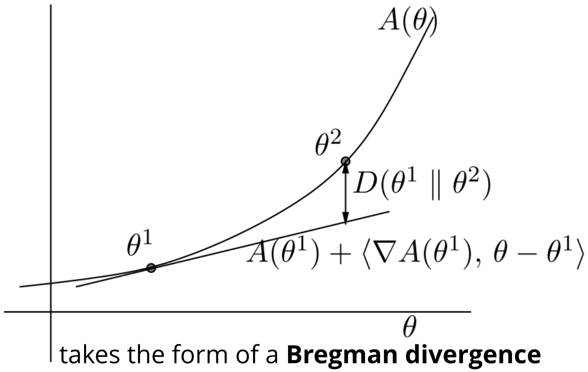


Relative entropy & inference

relative entropy of $p(x, \theta_1)$ and $p(x, \theta_2)$

$$D(\theta_1 \| \theta_2) = \langle \mu_1, \theta_1 - \theta_2 \rangle - A(\theta_1) + A(\theta_2)$$

where $\mu_1 = \nabla_\theta A(\theta_1)$



takes the form of a **Bregman divergence**

alternative form:

$$\min_{\mu_1 \in \mathcal{M}} D(\mu_1 \| \theta_2) = \max_{\mu_1 \in \mathcal{M}} \langle \mu_1, \theta_2 \rangle - A^*(\mu_1) - A(\theta_2)$$

familiar optimization!

does not depend on μ_1

so mapping $\theta \to \mu$ is minimizing the KL-divergence

- not symmetric, which one to use? is this the "right" one?

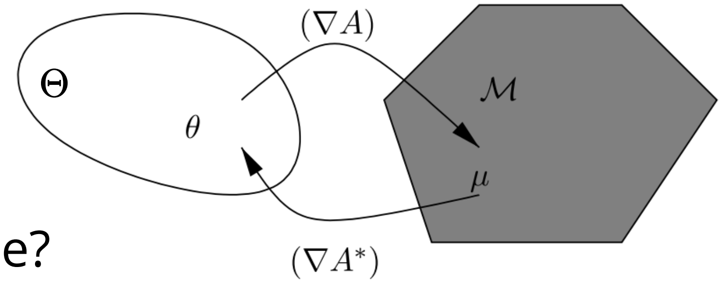
image: wainwright &jordan

Difficulty of inference

$A(\theta) = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu)$

e.g., gives us marginals in the Ising model

- isn't convex optimization tractable?



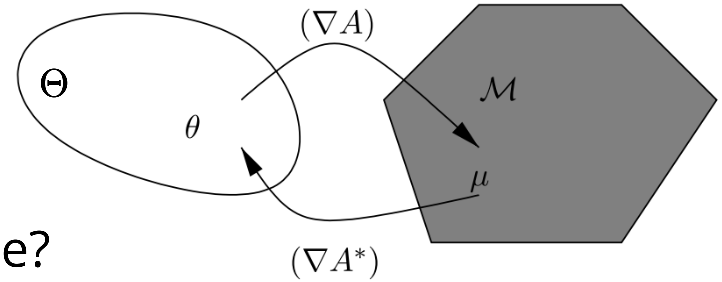
$\mathcal{M}$

image: wainwright &jordan

Difficulty of inference

$A(\theta) = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu)$

e.g., gives us marginals in the Ising model



- isn't convex optimization tractable?
- easy in the univariate case
  - closed form mapping $\nabla_\theta A(\theta)$

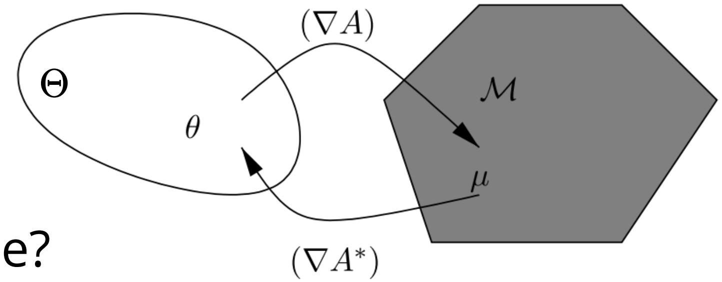
$\mathcal{M}$

image: wainwright &jordan

Difficulty of inference

$A(\theta) = \max_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu)$

e.g., gives us marginals in the Ising model



- isn't convex optimization tractable?
- easy in the univariate case
  - closed form mapping $\nabla_\theta A(\theta)$
- in (high-dimensional) graphical models:
  - $\mathcal{M}$ is difficult to specify (exponential #facets)
  - entropy doesn't have a simple form (approximate)

image: wainwright &jordan