Lecture 7 - Exponential Families

1. Exponential Families

→ A general way to describe many prob. distr. of interest
→ Well suited for inference / learning algorithms

Prelude: Exponential distributions from Max. Entropy

→ Recall: Gaussian is Max. entropy distr. subject to $\begin{cases} \mathbb{E}[X] = \mu \\ \mathrm{Cov}[X] = \Sigma \end{cases}$

→ More generally, consider a family $(\varphi_\alpha(x))_{\alpha \in I}$ of "potentials" / "energies" / "sufficient statistics"

- Observe some empirical values (e.g. on a sample)

$$\hat{\mu}_\alpha = \hat{\mathbb{E}}[\varphi_\alpha(X)] = \frac{1}{n} \sum_{i=1}^{n} \varphi_\alpha(x_i)$$

- Goal: Find a density $p(x)$ that matches those $\hat{\mu}_\alpha$:

$$\mathbb{E}_p[\varphi_\alpha(X)] = \hat{\mu}_\alpha$$

- Impose some regularity on $p(x)$, through entropy

$$H(p) = -\int p(x) \log p(x) \, dx$$

$\Rightarrow$ Solve the following optimization problem:

(Max-entropy)

$$\max_{p} \; H(p)$$
$$\text{s.t.} \quad \mathbb{E}_p[\varphi_\alpha(X)] = \hat{\mu}_\alpha, \quad \forall \alpha \in I$$

Fact: The solution to the problem (Max-entropy) is given by

$$p(x) \propto \exp\left(\sum_{\alpha \in I} \theta_\alpha \varphi_\alpha(x)\right),$$

where $(\theta_\alpha)_{\alpha \in I}$ are optimal Lagrange multipliers.

proof: Define the Lagrangian (Lagrange multipliers)

$$\mathcal{L}(p, \theta) = H(p) + \mathbb{E}_p\left[\sum_{\alpha \in I} \theta_\alpha (\varphi_\alpha(X) - \mu_\alpha)\right]$$

We want (p, θ) s.t.

$$\frac{\delta}{\delta p(x)} \mathcal{L}(p, \theta) = 0 \quad \forall x$$
$$\frac{\partial}{\partial \theta_\alpha} \mathcal{L}(p, \theta) = 0 \quad \forall \alpha$$

$$\frac{\delta}{\delta p(x)} \mathcal{L}(p, \theta) = 0 \Rightarrow p(x) \propto \exp\left(\sum_{\alpha} \theta_\alpha \varphi_\alpha(x)\right)$$

(calculus of variations, see lecture 2)

θ chosen to satisfy the constraints.

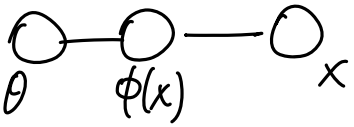
Exponential families : definitions

Def. (sufficient statistics)
Let $\mathcal{P}_\theta = \{ p_\theta(x), \theta \in \Theta \}$ be a class of models.
A function $\phi : x \mapsto \phi(x) \in \mathbb{R}^d$ is a sufficient statistic for $\mathcal{P}_\theta$ if we may write

$$p(x;\theta) = h(x)\, q(\phi(x);\theta)$$

Remarks:
- $\phi(x)$ carries all information needed to estimate θ
- Ex: Gaussian: $\phi(x) = (x, x^2)$ is sufficient
- In Bayesian terms, this encodes that

$$X \perp \theta \mid \phi(X)$$



Def. (exponential family)
A distribution is an exponential family if it has a density

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

w.r.t. some measure $d\lambda(x)$, where
- $\phi(x) = (\varphi_\alpha(x))_{\alpha \in I}$ are sufficient statistics
- $\theta = (\theta_\alpha)_{\alpha \in I}$ are canonical parameters (or natural params)
- $h(x)$ is the ancillary statistic

- $d\nu(x) = h(x)\, d\lambda(x)$ is the **base measure**
- $A(\theta)$ is the **log-partion function**, or cumulant function

$$A(\theta) := \log \int_x \exp\left(\langle \theta, \phi(x) \rangle\right) d\nu(x)$$

Domain: $\Omega := \left\{ \theta \in \mathbb{R}^d,\ A(\theta) < +\infty \right\}$

Def:

- **Regular** exp. family: Ω is an open set
- **Minimal** representation: there is no $(u_\alpha)_\alpha$ s.t.

$$\langle u, \phi(x) \rangle = \sum u_\alpha \varphi_\alpha(x) = \text{cst} \quad (\nu\text{-a.s.})$$

→ this makes parameters θ unique.

- **Overcomplete** representation: non-minimal representation

→ often more convenient to manipulate

Examples:

- **Bernoulli**:

$$p(x;\pi) = \pi^x (1-\pi)^{1-x}$$
$$= e^{x \log \pi + (1-x)\log(1-\pi)}$$
$$= e^{x \log \frac{\pi}{1-\pi} + \log(1-\pi)}$$

$$\phi(x) = x$$
$$\theta = \log \frac{\pi}{1-\pi}$$
$$A(\theta) = \log \sum_{x \in \{0,1\}} e^{x\theta} = \log(1 + \exp\theta)$$
$$= \log\left(1 + \frac{\pi}{1-\pi}\right) = -\log(1-\pi)$$

→ minimal, regular

- **Multinomial** (over-complete) $\quad x \in \{0,1\}^K, \ \sum_{i=1}^K x_i = 1$

$$p(x;\pi) = \prod_{i=1}^{K} \pi_i^{x_i} \qquad \sum \pi_i = 1.$$

$$= \exp\left(\sum_i x_i \cdot \log \pi_i\right)$$

$$= \exp\left(\langle x, \theta \rangle\right)$$

with $\theta_i := \log \pi_i$.

Here, $A(\theta)$ doesn't appear since π is normalized.

For general θ:

$$A(\theta) = \log \sum_i \exp \theta_i$$

$\Rightarrow$

$$p(x;\theta) = \exp\left(\sum_i \theta_i x_i - \log \sum_i \exp \theta_i\right)$$

$$= \exp\left(\sum_i \left(\log \frac{\exp \theta_i}{\sum_j \exp \theta_j}\right) x_i\right)$$

$$\pi_i !$$

- **Gaussian**

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

$$= \exp\left\{x \cdot \frac{\mu}{\sigma^2} + x^2\left(\frac{-1}{2\sigma^2}\right) - \left\{\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2\right\}\right\}$$

$$= \exp\left(\langle \phi(x), \theta \rangle - A(\theta)\right)$$

with: $\phi(x) = (x, x^2)$

$$\theta = (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$

$$A(\theta) = \frac{1}{2}\log\left|-\frac{\pi}{\theta_2}\right| - \frac{\theta_1^2}{4\theta_2}$$

- Other ex: Poisson, exponential, Beta, Binomial, etc...

- **Ising model**: $x_i \in \{\pm 1\}$ $G=(V,E)$

$$p(x;\theta) = \exp\left(\sum_{i\in V}\theta_i x_i + \sum_{(i,j)\in E}\theta_{ij}x_i x_j - A(\theta)\right)$$

$$\phi(x) = \left((x_i)_{i\in V}, (x_i x_j)_{(i,j)\in E}\right) \in \mathbb{R}^{|V|+|E|}$$

$$\theta = \left((\theta_i), (\theta_{ij})\right)$$

$$A(\theta) = \log \sum_{x\in\mathcal{X}} \exp\left(\sum_i \theta_i x_i + \sum_{i,j}\theta_{ij} x_i x_j\right)$$

- **Mixture model**:

Assume $p(x|z=i) = \exp\left(\langle \theta_i, \phi(x)\rangle - a_2(\theta_i)\right)$

$$p(z) = \prod_{i=1}^{K} \pi_i^{\mathbb{1}\{z=i\}} = \exp\left(\sum_i \mathbb{1}\{z=i\}\eta_i - a_2(\eta)\right)$$

$$p(x,z;\theta,\eta) = \exp\left\{\sum_i \mathbb{1}\{z=i\}\eta_i + \sum_i \langle \mathbb{1}\{z=i\}\phi(x), \theta_i\rangle - A(\theta,\eta)\right\}$$

$$\phi(x,z) = \left((\mathbb{1}\{z=i\})_i, (\mathbb{1}\{z=i\}\varphi(x))_i\right)$$

params: $(\eta_1,\dots,\eta_K,\theta_1,\dots,\theta_K)$

Remark: **Mean parameterization**

→ in many cases, there is an alternative meaningful parameterization given by

$$\mu_\alpha = \mathbb{E}_p[\varphi_\alpha(X)]$$

e.g. $\pi_i = \mathbb{E}[X_i]$ in Bernoulli / Multinomial

$\left.\begin{aligned} \mu &= \mathbb{E}[X] \\ \Sigma &= \mathbb{E}[(X-\mu)(X-\mu)^T] \end{aligned}\right)$ for Gaussians

→ going from one to the other is often difficult !!
(esp. in high dim.)

→ Note: $\theta \rightsquigarrow \mu$: inference

$\mu \rightsquigarrow \theta$: learning from empirical moments
(e.g. max-ent
also: MLE! see below)

■ Properties of $A(\theta)$

Prop: (Differentiability and convexity)
Let $A(\theta)$ be the log-partition fnct. of a regular exp. family

$$A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle) \, d\nu(x)$$

(i) A admits derivatives at all orders.

(ii) We have

$$\nabla A(\theta) = \mathbb{E}_\theta[\phi(X)] =: \mu_\theta$$

$$\nabla^2 A(\theta) = \mathbb{E}_\theta[\phi(X)\phi(X)^T] - \mu_\theta \mu_\theta^T$$

(iii) A is convex on its domain Ω

Proof:

(i)(ii) We only show $\nabla A(\theta) = \mu_\theta$, higher-order derivatives are analogous.

$$\nabla A(\theta) = \nabla_\theta \left\{ \log \int \exp(\langle \theta, \phi(x) \rangle) d\nu(x) \right\}$$

$$= \frac{\int \phi(x) \exp(\langle \theta, \phi(x) \rangle) \, d\nu(x)}{\int \exp(\langle \theta, \phi(x) \rangle) d\nu(x)}$$

$$= \int \phi(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \, d\nu(x)$$

$$= \int \phi(x) \, p(x;\theta) \, d\lambda(x) = \mathbb{E}_\theta[\phi(X)]$$

(iii) note that $\nabla^2 A(\theta)$ is the covariance matrix of $\phi(x)$ under $p(x;\theta)$. Thus, it is p.s.d $\Rightarrow$ A is convex. □

Corollary: (MLE as Moment Matching)

In a regular exponential family, maximum likelihood estimation is equivalent to finding θ such that

$$\mu_\theta = \hat{\mu}$$

(i.e. $\mathbb{E}_\theta[\phi(X)] = \hat{\mathbb{E}}[\phi(X)]$)

proof: $\ell(\theta) = \sum_{i=1}^{n} \log p(x_i, \theta)$

$= \langle \theta, \sum_i \phi(x_i) \rangle - n A(\theta)$

$= n \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right)$

$\ell(\theta)$ is concave, so the MLE is given by $\nabla_\theta \ell(\theta) = 0$

i.e. $\hat{\mu} = \nabla A(\theta) = \mu_\theta$ □

Remark: . Solving MLE then corresponds to "inverting" ∇A

. Recall that moment-matching is also used for maximum entropy -

- Q: → When is ∇A invertible and one-to-one?
  → Links with entropy?

$\theta \leftrightarrow \mu$ mappings, conjugate duality

- Forward mapping: $\theta \xrightarrow{\nabla A} \mu$

Define $\mathcal{M} = \{ \mu \in \mathbb{R}^d ; \exists p, \mathbb{E}_p[\phi(x)] = \mu \}$

(realizable mean parameters)

p is arbitrary

Q: Is $\nabla A : \Omega \to \mathcal{M}$ one-to-one?

Fact: Yes, if and only if we have a minimal representation

Also, for any $\mu \in \mathcal{M}^\circ$ (interior of $\mathcal{M}$), there exists $\theta = \theta(\mu) \in \Omega$ s.t. $\mathbb{E}_\theta[\phi(x)] = \mu$

proof: see W&J '08, Chap 3.

- Duality with maximum entropy

→ Conjugate duality (Fenchel/Legendre duality) [Rockafellar, 1997]

If, $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ convex

we may define

$$f^*(y) = \sup_{x \in \Omega} \langle x, y \rangle - f(x)$$ (dual function)

We then also have $f(x) = \sup_y x^T y - f^*(y)$

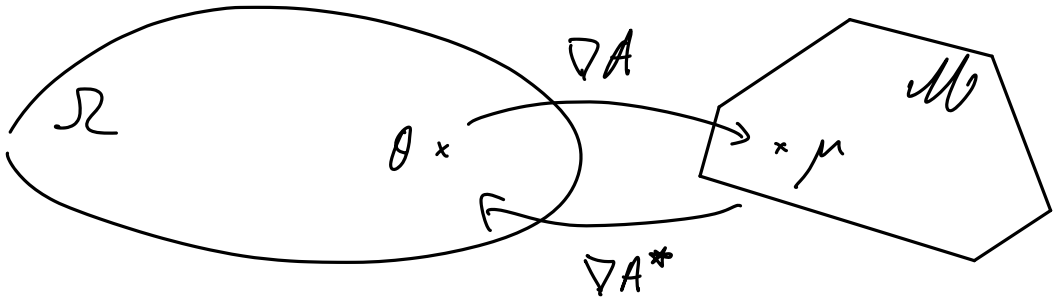
If f is also differentiable, we have (under some cond.)

$$\nabla f^* = (\nabla f)^{-1}$$

$$y = \nabla f(x) \iff x = \nabla f^*(y)$$

→ Duality for $A(\theta)$

Define $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$



Theorem: For $\mu \in \mathcal{M}^\circ$, let $\theta(\mu)$ be s.t. $\mu = \nabla A(\theta(\mu)) = \mathbb{E}_{\theta(\mu)}[\phi(x)]$

We have

$$A^*(\mu) = -H(p_{\theta(\mu)})$$

Remark:

- $A^*(\mu)$ corresponds to the optimum of the max-entropy problem over distributions that obey the moment constraint

- We have a variational representation of $A(\theta)$ as

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

The optimum corresponds to inference ($\mu = \mathbb{E}_\theta[\phi(x)]$)

$\Rightarrow$ "Variational" inference uses this to view inference as an optimization problem.

Note: representing $\mu \in \mathcal{M}$ may be intractable

$\Rightarrow$ consider subsets of $\mathcal{M}$ that lead to more tractable problems