Machine Learning 2

Exponential Families
- Basics

Patrick Forré

Exponential Families

- Exponential Families are sets of probability distributions that have a certain convenient structure relating the (data) arguments to the parameters.
- Most families of distributions we have seen so far can be written in exponential-family-form.
- Examples are Gaussian, Gamma, Poisson, Multinomial (with fixed number of trials), Bernoulli, Categorical, Dirichlet distributions, etc. etc.
- Counter examples are Student-t, Cauchy and Uniform distributions (with parameterized bounds).

Exponential Families - Definition

- A set of probability distributions $\{p(x \mid \theta) \mid \theta \in \Theta\}$ forms an Exponential Family, if it can be written in the form:

$$p(x \mid \theta) = h(x) \cdot \exp[\eta^\top \cdot T(x) - A(\eta)]$$

where:

- $h(x)$ is the base measure,
- $\eta = (\eta_1(\theta), \ldots, \eta_s(\theta))^\top$ are the natural parameters,
- $T(x) = (T_1(x), \ldots, T_s(x))^\top$ are the sufficient statistics,
- $A(\eta)$ is log-partition function, aka log-normalizer, aka cumulant function.

Exponential Families - Remarks

- We usually aim at a representation with a minimal number s of natural parameters: $\eta = (\eta_1(\theta), \ldots, \eta_s(\theta))$.
- $A(\eta) = \log\left(\int \exp\left(\eta^\top T(x)\right) \cdot h(x)\, dx\right)$ as log-normalizer, as $Z(\eta) := \exp(A(\eta))$ is the normalizing constant.
- Exponential Families are NOT the same as exponential distributions, but:

Example: Exponential distributions

- $p(x \mid \theta) = \theta \cdot e^{-\theta \cdot x}$ for $x \geq 0$, otherwise: 0. Parameters: $\theta \in \mathbb{R}_{>0}$

$= \underbrace{\mathbb{1}_{[0,\infty]}(x)}_{h(x)} \cdot \exp(\underbrace{-\theta}_{\eta} \cdot \underbrace{x}_{T(x)} + \underbrace{\log(\theta)}_{-A(\eta)})$

$p(x|\eta) = \mathbb{1}_{[0,\infty]}(x) \cdot \exp(\eta \cdot x - (-\log(-\eta)))$

- $h(x) = \mathbb{1}_{[0,\infty]}(x)$
- $\eta(\theta) = -\theta$
- $T(x) = x$
- $A(\eta) = -\log(-\eta)$

$\theta(\eta) = -\eta$

How to find an exponential family structure?

- Let the set of probability distributions $\{p(x\,|\,\theta) \mid \theta \in \Theta\}$ be given.
- Be clear what are parameters θ and what possible fixed hyper-parameters α.
- Find support: $S = \{x \,|\, p(x\,|\,\theta) > 0 \forall \theta\}$ and try to write for $x \in S$:

$$\log p(x\,|\,\theta) = \eta_1(\theta) \cdot T_1(x) + \cdots + \eta_s(\theta) \cdot T_s(x) + B(x) + C(\theta),$$

  by factoring out everything and separating θ-terms and x-terms, and gathering them into a minimal number of functions: $\eta_i(\theta)$, $T_i(x)$, $B(x)$, $C(\theta)$.

- Put $h(x) := \mathbb{1}_S(x) \cdot \exp(B(x))$.
- Express θ as a function $\theta(\eta)$ of η and put $A(\eta) := -C(\theta(\eta))$.

Counter-Example: Student-t distributions

- $p(x \mid \mu, \Sigma, \nu) = \frac{\Gamma(\frac{D}{2} + \frac{\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{(\pi\nu)^{D/2} (\det \Sigma)^{1/2}} \left[1 + \nu^{-1}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]^{-\frac{D}{2} - \frac{\nu}{2}}$

NOT an exponential family

- can't be written as exponential family (not so easy to prove).
- looks like Gaussian, but with heavier tails, polynomial decay (instead of exponential decay)

Machine Learning 2

Exponential Families
- Example: Multinomial distributions (fixed K, M)

Patrick Forré

Example: Multinomial distribution (fixed M, K)

$K+1 := \#\text{classes}$

$M := \#\text{trial}$

$$\Delta^K := \left\{ \pi = (\pi_0, \dots, \pi_K) \in \mathbb{R}^{K+1} \;\middle|\; \pi_k \geq 0 \;,\; \sum_{k=0}^{K} \pi_k = 1 \right\}$$

$$x = (x_0, \dots, x_K) \in \mathbb{N}^{K+1}$$

$$\sum_{k=0}^{K} x_k = M$$

$$p(x|\pi) = \binom{M}{x} \cdot \pi^x$$

$$= \begin{cases} \dfrac{M!}{x_0! \cdots x_K!} & \text{if } \sum_{k=0}^{K} x_k = M \\ 0 & \text{o.Hemite} \end{cases} \cdot \pi_0^{x_0} \cdot \dots \cdot \pi_K^{x_K}$$

$$S = \left\{ x \in \mathbb{N}^{K+1} \;\middle|\; \sum_{k=0}^{K} x_k = M \right\}$$

Example: Multinomial distribution (fixed M, K)

$$\log p(x|\pi) = \log\binom{M}{x} + \log \pi^x = \log\binom{M}{x} + \sum_{k=0}^{K} x_k \cdot \log \pi_k$$

$$= \log\binom{M}{x} + x_0 \log \pi_0 + \sum_{k=1}^{K} x_k \log \pi_k$$

$$= \log\binom{M}{x} + \left(M - \sum_{k=1}^{K} x_k\right) \cdot \log\left(1 - \sum_{k=1}^{K} \pi_k\right) + \sum_{k=1}^{K} x_k \log(\pi_k)$$

$$= \underbrace{\log\binom{M}{x}}_{B(x)} + \underbrace{M \cdot \log\left(1 - \sum_{k=1}^{K} \pi_k\right)}_{C(\pi)} + \sum_{k=1}^{K} \underbrace{x_k}_{T_k(x)} \cdot \underbrace{\log\left(\frac{\pi_k}{1 - \sum_{j=1}^{K} \pi_j}\right)}_{\eta_k}$$

$$x_0 = M - \sum_{k=1}^{K} x_k$$

$$\pi_0 = 1 - \sum_{k=1}^{K} \pi_k$$

Example: Multinomial distribution (fixed M, K)

$$\eta_k = \log\left(\frac{\pi_k}{1 - \sum_{j=1}^{k} \pi_j}\right) \qquad k = 1, \dots, K$$

$$1 + \sum_{k=1}^{K} \exp(\eta_k) = 1 + \frac{\sum_{k=1}^{K} \pi_k}{1 - \sum_{j=1}^{K} \pi_j} = \frac{1 - \sum_{j=1}^{K} \pi_j + \sum_{k=1}^{K} \pi_k}{1 - \sum_{j=1}^{K} \pi_j} = \left(1 - \sum_{j=1}^{K} \pi_j\right)^{-1}$$

$$\Rightarrow \exp(\eta_k) = \frac{\pi_k}{1 - \sum_{j=1}^{K} \pi_j} = \pi_k \left(1 + \sum_{j=1}^{K} \exp(\eta_j)\right)$$

$$\Rightarrow \pi_k = \frac{\exp(\eta_k)}{1 + \sum_{j=1}^{K} \exp(\eta_j)} = \text{softmax}_k(\eta) = \sigma_k(\eta)$$

Example: Multinomial distribution (fixed M, K)

$$A(\eta) = -C(\pi(\eta)) = -C(\theta(\eta)) = -M \log\left(1 - \sum_{j=1}^{K} \pi_j\right)$$

$$= +M \cdot \log\left(1 + \sum_{j=1}^{K} \exp(\eta_j)\right)$$

$$\left(1 + \sum_{j=1}^{K} \exp(\eta_j)\right)^{-1}$$

$$p(x|\eta) = \binom{M}{x} \cdot \exp\left[\sum_{k=1}^{K} \eta_k \cdot x_k - \left(M \cdot \log\left(1 + \sum_{j=1}^{K} \exp(\eta_j)\right)\right)\right]$$

↳ exponential family with minimum number of natural parameters η_i sufficient statistics T_i.

Example: Binomial distribution (fixed N)

$$p(x|\pi) = \binom{N}{x} \pi^x (1-\pi)^{(N-x)}$$

$$= \text{multinomial}(\,|\,M=N,\ K=1)$$

↳ exponential family, same quantities as before

Example: Bernoulli distributions

- $p(x \mid \pi) =$ Multinomial (K=1, N=1)

  ↳ exponential family

  same quantities as before

- $h(x) =$
- $\eta(\theta) =$ $\theta(\eta) =$
- $T(x) =$
- $A(\eta) =$

Machine Learning 2

Exponential Families
- Example: Gaussian distributions

Patrick Forré

Example: D-dim Gaussian distributions

- $p(x \mid \mu, \Sigma) =$ $(2\pi)^{-D/2} \det(\Sigma)^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$

$$\log p(x \mid \mu, \Sigma) = \underbrace{-\frac{D}{2}\log(2\pi)}_{B(x)} + \underbrace{\mu^T\Sigma^{-1}}_{\eta_1^T}\underbrace{x}_{T_1(x)} \underbrace{-\frac{1}{2}x^T\Sigma^{-1}x}_{??} \underbrace{-\frac{1}{2}\mu^T\Sigma^{-1}\mu - \frac{1}{2}\log\det(\Sigma)}_{C(\mu,\Sigma)}$$

(1×1)

$$-\frac{1}{2}x^T\Sigma^{-1}x = \mathrm{Tr}\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) = \mathrm{Tr}\left(\underbrace{-\frac{1}{2}\Sigma^{-1}}_{\eta_2}\underbrace{xx^T}_{T_2(x)}\right)$$

$$= \mathrm{vec}\left(-\tfrac{1}{2}\Sigma^{-1}\right)^T \mathrm{vec}(xx^T)$$

Example: D-dim Gaussian distributions

- $p(x \mid \mu, \Sigma) =$

$$\eta_1^T = \mu^T \Sigma^{-1}$$

$$\eta_2 = -\frac{1}{2}\Sigma^{-1}$$

$$\Rightarrow \Sigma = -\frac{1}{2}\eta_2^{-1} \qquad (\Sigma^{-1} = -2\eta_2)$$

$$\mu^T = \eta_1^T \Sigma = -\frac{1}{2}\eta_1^T \eta_2^{-1}$$

$$\mu = -\frac{1}{2}\eta_2^{-T}\eta_1$$

$$A(\eta) = -C(\mu, \Sigma) = \frac{1}{2}\mu^T \Sigma^{-1}\mu + \frac{1}{2}\log\det(\Sigma)$$

$$= \frac{1}{2}\eta_1^T\left(-\frac{1}{2}\eta_2^{-T}\eta_1\right) + \frac{1}{2}\log\det\left(-\frac{1}{2}\eta_2^{-1}\right)$$

$$= -\frac{1}{4}\eta_1^T\eta_2^{-T}\eta_1 - \frac{1}{2}\log\det(-2\eta_2)$$

Example: D-dim Gaussian distributions

- $p(x \mid \mu, \Sigma) =$

$$p(x|\eta) = (2\pi)^{-D/2} \exp\left[\eta_1^T x + \mathrm{Tr}(\eta_2 \cdot (xx^T)) - A(\eta)\right]$$

$$A(\eta) = -\frac{1}{4}\eta_1^T \eta_2^{-T} \eta_1 - \frac{1}{2}\log\det(-2\eta_2)$$

Example: 1-dim Gaussian distributions

- $p(x \mid \mu, \sigma^2) =$

$\Sigma = \sigma^2$

$\rightsquigarrow D = 1$

$\rightsquigarrow$ same equations hold

- $h(x) =$
- $\eta(\mu, \sigma^2) =$
- $T(x) =$
- $A(\eta) =$

Machine Learning 2

Exponential Families
- Conjugate Prior, Bayesian Update

Patrick Forré

Conjugate prior - Definition

- Let $M = \{p(x \mid \theta) \mid \theta \in \Theta\}$ be a statistical model and D some data.
- A conjugate prior w.r.t. M is a family of priors $F = \{p(\theta \mid \alpha) \mid \alpha \in A\}$ such that all possible posteriors $p(\theta \mid \alpha, D)$ also lie in F.
- Remember:

$$\text{posterior} \quad p(\theta \mid \alpha, D) = \frac{\overbrace{p(D \mid \theta)}^{\text{likelihood}} \cdot \overbrace{p(\theta \mid \alpha)}^{\text{prior}}}{\underbrace{p(D \mid \alpha)}_{\text{evidence}}} \propto p(D \mid \theta) \cdot p(\theta \mid \alpha)$$

$$p(D \mid \alpha) = \int p(D \mid \theta)\, p(\theta \mid \alpha)\, d\theta$$

Conjugate prior for Exponential Families

- Consider an Exponential Familie in a natural parameterization:

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

- Then a conjugate prior is given by:

$$p(\eta \mid \tau, \nu) \propto \exp\left[\eta^\top \tau - \nu \cdot A(\eta)\right]$$

$$p(\eta \mid \tau, \nu) \propto \exp\left(\eta^T \tau - \nu \cdot A(\eta)\right)$$

$$p(\eta \mid \tau, \nu, X) \propto \prod_{n=1}^{N} p(x^{(n)} \mid \eta) \cdot p(\eta \mid \tau, \nu)$$

$$\propto \prod_{n=1}^{N} \exp\left[\eta^T T(x^{(n)}) - A(\eta)\right] \cdot \exp\left[\eta^T \tau - \nu \cdot A(\eta)\right]$$

$$= \exp\Big[\eta^T \underbrace{\Big(\sum_{n=1}^{N} T(x^{(n)}) + \tau\Big)}_{\tilde{\tau}} - \underbrace{(N+\nu)}_{\tilde{\nu}} \cdot A(\eta)\Big]$$

$$\propto p(\eta \mid \tilde{\tau}, \tilde{\nu}) \Rightarrow \text{conjugate.}$$

Bayesian update rule

prior → posterior

$\tau \longmapsto \tau + \sum_{n=1}^{N} T(x^{(n)})$

$\nu \longmapsto \nu + N$

Example: Multinomial distributions (fixed M)

$$p(x|\pi) = \binom{M}{x} \cdot \pi^x$$

$$p(x|\eta) = \binom{M}{x} \exp\left[\sum_{k=0}^{K} \eta_k \cdot x_k\right]$$

$$p(\eta|\tau) \propto \exp\left[\sum_{k=0}^{K} \eta_k \cdot \tau_k\right]$$

$\eta_k = \log \pi_k$, $k = 0, \dots, K$

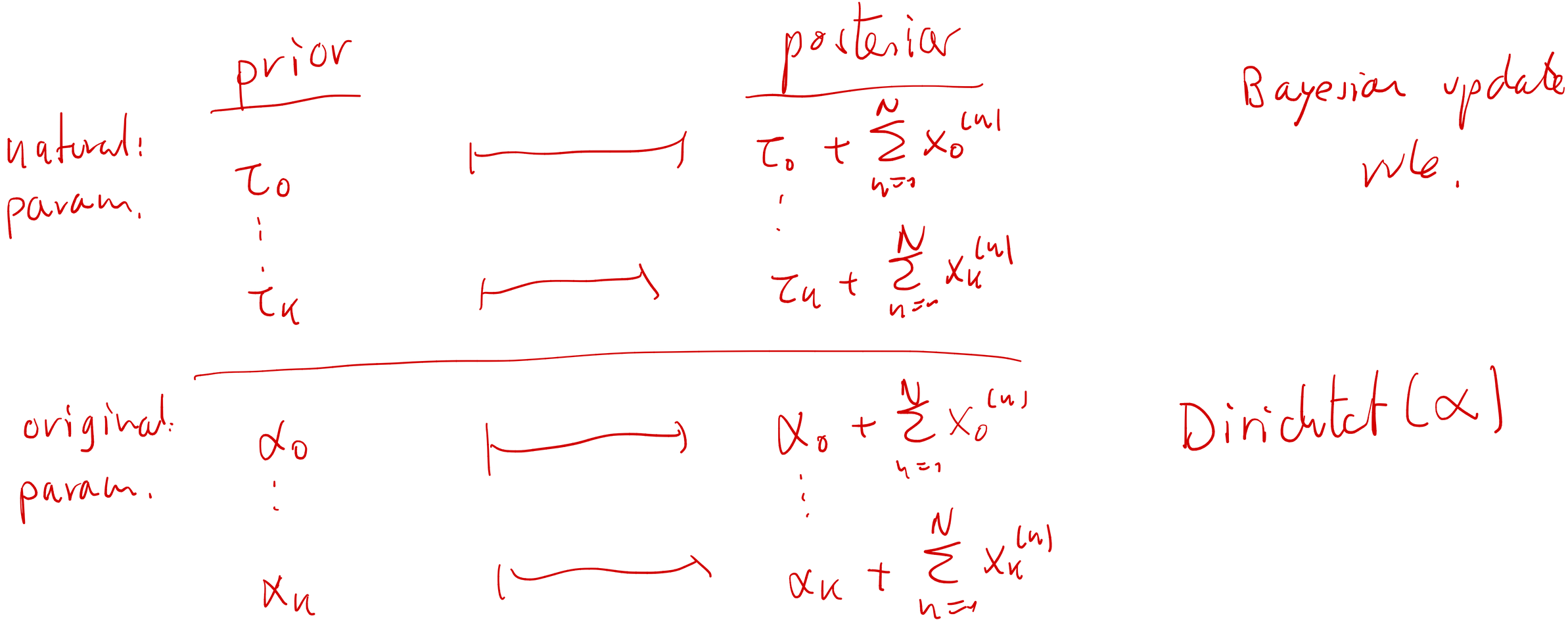
$A(\eta) = 0$

τ_k , $k = 0, \dots, K$

↳ go back to original parametrization $\alpha_k := \tau_k + 1$

$\rightsquigarrow p(\pi|\alpha) \propto \prod_{k=0}^{K} \pi_k^{\alpha_k - 1}$ Dirichlet distribution

Example: Multinomial distributions (fixed M)



Machine Learning 2

Exponential Families
- Deriving Moments

Patrick Forré

Computing Moments for Exponential Families

$A(\eta)$ (minimal number of natural parameters)

$$A(\eta) = \log \int h(x) \exp(\eta^T T(x)) \, dx$$

1.) $\nabla_\eta A(\eta) = \mathbb{E}_\eta[T(x)]$

2.) $\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta) = \mathrm{Cov}_\eta(T_i(x), T_j(x))$

proof: $\frac{\partial}{\partial \eta_k} A(\eta) = \frac{1}{\int h(x) \cdot \exp(\eta^T T(x)) \, dx} \cdot \int T_k(x) \, h(x) \exp(\eta^T T(x)) \, dx$

$= \int T_k(x) \underbrace{\exp(-A(\eta)) \cdot h(x) \exp(\eta^T T(x)) \, dx}_{p(x|\eta)} = \mathbb{E}_\eta[T_k]$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta) = \frac{\partial}{\partial \eta_i} \int T_j(x)\, h(x) \exp[\eta^T T(x) - A(\eta)]\, dx$$

$$= \int T_j(x) \cdot \underbrace{\left(T_i(x) - \mathbb{E}_\eta[T_i(x)] \right) \cdot h(x) \cdot \exp\left(\eta^T T(x) - A(\eta) \right)}_{p(x|\eta)} dx$$

$$= \mathbb{E}_\eta\left[T_j(x) \cdot \left(T_i(x) - \mathbb{E}_\eta[T_i(x)] \right) \right]$$

$$= \mathbb{E}_\eta[T_j(x) \cdot T_i(x)] - \mathbb{E}_{\eta_1}[T_j(x)] \cdot \mathbb{E}_{\eta_2}[T_i(x)]$$

$$= \mathrm{cov}_\eta(T_i(x), T_j(x))$$

Example: Exponential distributions

$$p(x|\lambda) = \mathbb{1}_{[0,\infty)}(x) \cdot \lambda \cdot \exp(-\lambda \cdot x)$$

$$p(x|\eta) = \mathbb{1}_{[0,\infty)}(x) \exp(\eta \cdot x - (-\log(-\eta)))$$

$\eta = -\lambda$, $T(x) = x$

$A(\eta) = -\log(-\eta)$

$$\mathbb{E}[T(x)] = A'(\eta) = +\frac{1}{-\eta} = \frac{1}{\lambda} \quad (\checkmark)$$

$$Var(T(x)) = A''(\eta) = \frac{1}{\eta^2} = \frac{1}{\lambda^2} \quad (\checkmark)$$

Example: Gaussian distributions

$N(x|\mu, \sigma^2)$, $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$

$\eta_1 = \frac{\mu}{\sigma^2}$

$\eta_2 = -\frac{1}{2\sigma^2}$

$T_1(x) = x$, $T_2(x) = x^2$

- $\mathbb{E}[X] = \mathbb{E}[T_1(X)] = \partial_1 A(\eta) = -\frac{\eta_1}{2\eta_2}$ $\left(= \left(\frac{\mu}{\sigma^2}\right) \cdot (\sigma^2) = \mu \right)$ ✓
- $\mathrm{Var}(X) = \mathrm{Var}(T_1(X)) = \partial_{1,1}^2 A(\eta) = -\frac{1}{2\eta_2}$ $\left(= \sigma^2 \right)$ ✓
- $\mathbb{E}[X^2] = \mathbb{E}[T_2(X)] = \partial_2 A(\eta) = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2}$ $\left(= +\mu^2 + \sigma^2 \right)$ ✓
- $\mathrm{Var}(X^2) = \mathrm{Var}[T_2(X)] = \partial_{2,2}^2 A(\eta) = (-2)\frac{\eta_1^2}{4\eta_2^3} + \frac{1}{2\eta_2^2}$ $\left(= 4\mu^2 + 2\sigma^2 \; ? \right)$

Machine Learning 2

Exponential Families
- Summary

Patrick Forré

Summary

- Exponential Family: $p(x \mid \eta) = h(x) \cdot \exp\left(\eta^\top T(x) - A(\eta)\right)$
- How to transform typical distributions into exponential-family-form.
- Conjugate prior: $p(\eta \mid \tau, \nu) \propto \exp(\eta^\top \tau - \nu \cdot A(\eta))$
- Baysian update: $(\tau, \nu) \mapsto \left(\tau + \sum_{n=1}^{N} T(x^{(n)}), \nu + N\right)$
- Moments: $\mathbb{E}[T_k(X)] = \frac{\partial}{\partial \eta_k} A(\eta), \qquad \operatorname{cov}(T_i(X), T_j(X)] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta)$
- Counter-example: Student-t distributions