

PROBABILISTIC MACHINE LEARNING

LECTURE 15

EXPONENTIAL FAMILIES

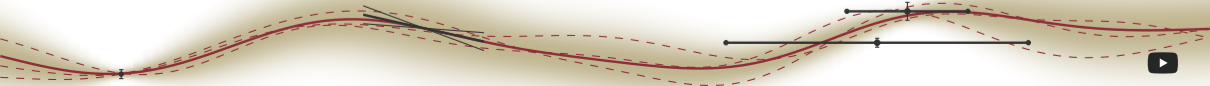
Philipp Hennig

14 June 2021

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



FACULTY OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
CHAIR FOR THE METHODS OF MACHINE LEARNING



#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Generalized Linear Models	
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	8
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	9
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	
6	05.05.	Gaussian Distributions		19	29.06.	Example: Topic Models	10
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	
8	12.05.	Learning Representations		21	06.07.	EM	11
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	
10	19.05.	Understanding Kernels		23	13.07.	Topics	
11	26.05.	Gauss-Markov Models		25	20.07.	Example: Kernel Topic Models	
12	25.05.	An Example for GP Regression	6	24	14.07.	Example: Inferringevision	
13	08.06.	GP Classification	7	26	21.07.	Revision	



Why is this hard?

The computational challenge in Bayesian Inference

$$p(x | y) = \frac{p(y | x)p(x)}{\int p(y | x)p(x) dx}$$

- ▶ the integral $\int p(y | x)p(x) dx$ may be intractable
- ▶ thus, also expectations $\int f(x)p(x | y) dx$ are hard

Practical probabilistic inference is chiefly a *computational* task.

The Toolbox

Framework:

$$\int p(x_1, x_2) dx_2 = p(x_1) \qquad p(x_1, x_2) = p(x_1 | x_2)p(x_2) \qquad p(x | y) = \frac{p(y | x)p(x)}{p(y)}$$

Modelling:

- ▶ graphical models (conditional independence)
- ▶ Gaussian distributions
- ▶ Kernels
- ▶ Markov Chains
- ▶ Exponential Families / Conjugate Priors
- ▶

Computation:

- ▶ Monte Carlo
- ▶ Linear algebra / Gaussian inference
- ▶ maximum likelihood / MAP
- ▶ Laplace approximations
- ▶



Recall from GP regression: How to set parameters θ ? From marginal likelihood $p(Y | \theta)$:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \mathcal{N}(\mathbf{y}; \phi_X^{\theta \top} \mu + b, \phi_X^{\theta \top} \Sigma \phi_X^{\theta} + \Lambda) \\ &= \arg \max_{\theta} \log \mathcal{N}(\mathbf{y}; \phi_X^{\theta \top} \mu + b, \phi_X^{\theta \top} \Sigma \phi_X^{\theta} + \Lambda) \\ &= \arg \min_{\theta} -\log \mathcal{N}(\mathbf{y}; \phi_X^{\theta \top} \mu + b, \phi_X^{\theta \top} \Sigma \phi_X^{\theta} + \Lambda) \\ &= \arg \min_{\theta} \frac{1}{2} \left(\underbrace{(\mathbf{y} - \phi_X^{\theta \top} \mu)^{\top} (\phi_X^{\theta \top} \Sigma \phi_X^{\theta} + \Lambda)^{-1} (\mathbf{y} - \phi_X^{\theta \top} \mu)}_{\text{square error}} + \underbrace{\log |\phi_X^{\theta \top} \Sigma \phi_X^{\theta} + \Lambda|}_{\text{model complexity / Occam factor}} \right) + \frac{N}{2} \log 2\pi\end{aligned}$$

In general, hierarchical inference is not analytically tractable. However, there are special cases...

$$p(\mathbf{x} | \mu) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \Sigma) \quad \text{and} \quad p(\mu | \mu_0, \Sigma_0) = \mathcal{N}(\mu; \mu_0, \Sigma_0)$$

$$p(\mu | \mathbf{x}) = \frac{p(\mathbf{x} | \mu)p(\mu | \mu_0, \Sigma_0)}{p(\mathbf{x})} = \mathcal{N}\left(\mu; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_i x_i), (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}\right)$$



$$p(x | f) = \prod_{i=1}^n f^x \cdot (1 - f)^{1-x} \quad x \in \{0; 1\}$$
$$= f^{n_1} \cdot (1 - f)^{n_0} \quad n_0 := n - n_1$$

$$p(f | \alpha, \beta) = \mathcal{B}(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} f^{\alpha-1} (1 - f)^{\beta-1}$$

$$p(f | x) = \mathcal{B}(\alpha + n_1, \beta + n_0) = \frac{1}{B(\alpha + n_1, \beta + n_0)} f^{\alpha+n_1-1} (1 - f)^{\beta+n_0-1}$$



Pierre Simon, marquis de
Laplace, 1749–1827



$$p(\mathbf{x}) = \prod_{i=1}^n f_{x_i} \quad \mathbf{x} \in \{0; \dots, K\}$$
$$= \prod_{k=1}^K f_k^{n_k} \quad n_k := |\{x_i \mid x_i = k\}|$$

$$p(f \mid \boldsymbol{\alpha}) = \mathcal{D}(\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K f_k^{\alpha_k - 1}$$

$$p(f \mid \mathbf{x}) = \mathcal{D}(\boldsymbol{\alpha} + \mathbf{n})$$



Peter Gustav Lejeune Dirichlet
(1805–1859)

$$p(\mathbf{x} \mid \sigma) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2)$$
$$p(\sigma) = ?$$

$$p(\mathbf{x} | \sigma) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2)$$

$$p(\sigma) = ?$$

$$\log p(\mathbf{x} | \sigma) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2} (x - \mu)^2 \cdot \frac{1}{\sigma^2} - \frac{1}{2} \log 2\pi$$



$$p(\mathbf{x} | \sigma) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2)$$

$$p(\sigma) = ?$$

$$\log p(\mathbf{x} | \sigma) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2} (x - \mu)^2 \cdot \frac{1}{\sigma^2} - \frac{1}{2} \log 2\pi$$

$$\log p(\sigma | \alpha, \beta) = (\alpha + 1) \log \sigma^{-2} - \beta \cdot \frac{1}{\sigma^2} - Z(\alpha, \beta)$$

$$p(\sigma | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^{-2})^{\alpha+1} e^{-\beta \sigma^{-2}} =: \mathcal{G}(\sigma^{-2}; \alpha, \beta)$$

$$p(\sigma | \alpha, \beta, \mathbf{x}) = \mathcal{G} \left(\sigma^{-2}; \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_i (x_i - \mu)^2 \right)$$



Daniel Bernoulli (1700–1782)

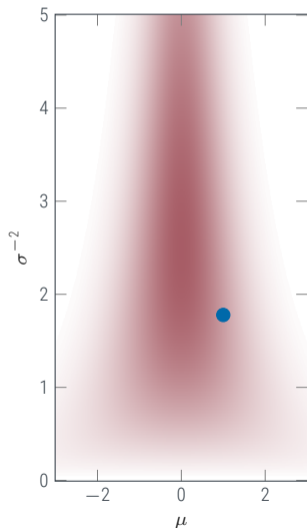
$$p(\mathbf{x} \mid \mu, \sigma) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2)$$

$$p(\mu, \sigma \mid \mu_0, \nu, \alpha, \beta) = \mathcal{N}\left(\mu; \mu_0, \frac{\sigma^2}{\nu}\right) \mathcal{G}(\sigma^{-2}; \alpha, \beta)$$

$$p(\mu, \sigma \mid \mathbf{x}, \mu_0, \nu, \alpha, \beta) = \mathcal{N}\left(\mu; \frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \frac{\sigma^2}{\nu + n}\right) \cdot$$

$$\mathcal{G}\left(\sigma^{-2}; \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{2(n + \nu)} (\bar{x} - \mu_0)^2\right)$$

$$\text{where } \bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$



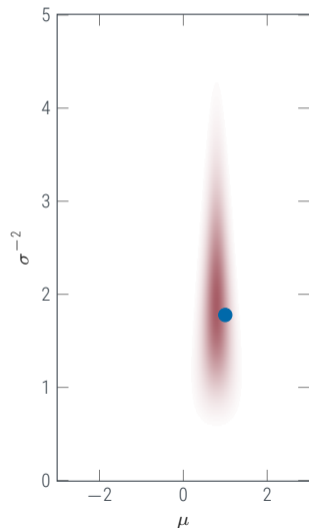
$$p(\mathbf{x} \mid \mu, \sigma) = \prod_{i=1}^n \mathcal{N}(x_i; \mu, \sigma^2)$$

$$p(\mu, \sigma \mid \mu_0, \nu, \alpha, \beta) = \mathcal{N}\left(\mu; \mu_0, \frac{\sigma^2}{\nu}\right) \mathcal{G}(\sigma^{-2}; \alpha, \beta)$$

$$p(\mu, \sigma \mid \mathbf{x}, \mu_0, \nu, \alpha, \beta) = \mathcal{N}\left(\mu; \frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \frac{\sigma^2}{\nu + n}\right) \cdot$$

$$\mathcal{G}\left(\sigma^{-2}; \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{2(n + \nu)} (\bar{x} - \mu_0)^2\right)$$

$$\text{where } \bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$



Definition (Conjugate Prior)

Let D and x be a data-set and a variable to be inferred, respectively, connected by the likelihood $p(D | x) = \ell(D; x)$. A **conjugate prior to ℓ for x** is a probability measure with pdf $p(x) = \pi(x; \theta)$ of functional form π , such that

$$p(x | D) = \frac{\ell(D; x)\pi(x; \theta)}{\int \ell(D; x)\pi(x; \theta) dx} = \pi(x; \theta').$$

That is, such that the posterior arising from ℓ is of the same functional form as the prior, with updated parameters.

E. Pitman. *Sufficient statistics and intrinsic accuracy* (1936). Math. Proc. Cambr. Phil. Soc. 32(4), 1936.
P. Diaconis and D. Ylvisaker, *Conjugate priors for exponential families*. Annals of Statistics 7(2), 1979.

- ▶ **Conjugate priors** allow analytic Bayesian inference
- ▶ How can we construct them in general?



Definition (Exponential Family, simplified form)

Consider a random variable X taking values $x \in \mathbb{X} \subset \mathbb{R}^n$. A probability distribution for X with pdf of the functional form

$$p_w(x) = h(x) \exp[\phi(x)^\top w - \log Z(w)] = \frac{h(x)}{Z(w)} e^{\phi(x)^\top w} = p(x | w)$$

is called an **exponential family** of probability measures. The function $\phi : \mathbb{X} \rightarrow \mathbb{R}^d$ is called the **sufficient statistics**. The parameters $w \in \mathbb{R}^d$ are the **natural parameters** of p_w . The normalization constant $Z(w) : \mathbb{R}^d \rightarrow \mathbb{R}$ is the **partition function**. The function $h(x) : \mathbb{X} \rightarrow \mathbb{R}_+$ is the **base measure**.

$$\begin{aligned} p(k | q) &= \binom{n}{k} q^k \cdot (1 - q)^{n-k} \quad (\text{nb: treating } n \text{ as fixed}) \\ &= \binom{n}{k} \exp(k \log q + (n - k) \log(1 - q)) \\ &= \underbrace{\binom{n}{k}}_{=: h(k)} \exp \left(\underbrace{k}_{\phi(k)} \underbrace{\log \frac{q}{1 - q}}_w + \underbrace{n \log(1 - q)}_{-\log Z(w)} \right) \\ \log Z(w) &= n \log(1 + e^w) \end{aligned}$$

$$\begin{aligned} p(q \mid \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} q^{\alpha-1} (1-q)^{\beta-1} \\ &= \underbrace{1}_{h(q)} \exp \left(\underbrace{\begin{bmatrix} \log q \\ \log(1-q) \end{bmatrix}}_{=: \phi^\tau(q)}^\top \underbrace{\begin{bmatrix} \alpha-1 \\ \beta-1 \end{bmatrix}}_w - \log B(\alpha, \beta) \right) \\ &= \frac{1}{\underbrace{q(1-q)}_{\tilde{h}(q)}} \exp \left(\begin{bmatrix} \log q \\ \log(1-q) \end{bmatrix}^\top \underbrace{\begin{bmatrix} \alpha \\ \beta \end{bmatrix}}_{\tilde{w}} - \log B(\alpha, \beta) \right) \end{aligned}$$

sufficient statistics ϕ , natural parameters w and base measure h are not uniquely defined.

A Family Meeting

incomplete list of exponential families

Name	sufficient stats	domain	use case
Bernoulli	$\phi(x) = [x]$	$\mathbb{X} = \{0; 1\}$	coin toss
Poisson	$\phi(x) = [x]$	$\mathbb{X} = \mathbb{R}_+$	emails per day
Laplace	$\phi(x) = [1, x]^\top$	$\mathbb{X} = \mathbb{R}$	floods
Helmert (χ^2)	$\phi(x) = [x, -\log x]$	$\mathbb{X} = \mathbb{R}$	variances
Dirichlet	$\phi(x) = [\log x]$	$\mathbb{X} = \mathbb{R}_+$	class probabilities
Euler (Γ)	$\phi(x) = [x, \log x]$	$\mathbb{X} = \mathbb{R}_+$	variances
Wishart	$\phi(X) = [X, \log X]$	$\mathbb{X} = \{X \in \mathbb{R}^{N \times N} \mid v^\top X v \geq 0 \forall v \in \mathbb{R}^N\}$	covariances
Gauss	$\phi(X) = [X, XX^\top]$	$\mathbb{X} = \mathbb{R}^N$	functions
Boltzmann	$\phi(X) = [X, \text{triag}(XX^\top)]$	$\mathbb{X} = \{0; 1\}^N$	thermodynamics

Exponential Families have Conjugate Priors

but the prior's normalization constant can be tricky

- ▶ Consider the exponential family $p_w(x | w) = h(x) \exp[\phi(x)^\top w - \log Z(w)]$
- ▶ its conjugate prior is the exponential family $F(\alpha, \nu) = \int \exp(\alpha^\top w - \nu \log Z(w)) dw$

$$p_\alpha(w | \alpha, \nu) = \exp \left[\begin{pmatrix} w \\ -\log Z(w) \end{pmatrix}^\top \begin{pmatrix} \alpha \\ \nu \end{pmatrix} - \log F(\alpha, \nu) \right]$$

$$\text{because } p_\alpha(w | \alpha, \nu) \prod_{i=1}^n p_w(x_i | w) \propto p_\alpha \left(w \mid \alpha + \sum_i \phi(x_i), \nu + n \right)$$

- ▶ and the predictive is

$$\begin{aligned} p(x) &= \int p_w(x | w) p_\alpha(w | \alpha, \nu) dw = h(x) \int e^{(\phi(x) + \alpha)^\top w + (\nu + 1) \log Z(w) - \log F(\alpha, \nu)} dw \\ &= h(x) \frac{F(\phi(x) + \alpha, \nu + 1)}{F(\alpha, \nu)} \end{aligned}$$

Computing $F(\alpha, \nu)$ can be tricky. In general, this is **the** challenge when constructing an EF.



why they're called **sufficient statistics**

- ▶ Consider the exponential family

$$p_w(x | w) = \exp [\phi(x)^\top w - \log Z(w)]$$

- ▶ for iid data:

$$p_w(x_1, x_2, \dots, x_n | w) = \prod_i^n p_w(x_i | w) = \exp \left(\sum_i^n \phi^\top(x_i) w - n \log Z(w) \right)$$

- ▶ to find the **maximum likelihood** estimate for w , set

$$\nabla_w \log p(x | w) = 0 \quad \Rightarrow \quad \nabla_w \log Z(w) = \frac{1}{n} \sum_i \phi(x_i)$$

- ▶ hence, collect **statistics** of ϕ , compute $\nabla_w \log Z(w)$ and solve the above for w .

- ▶ Re-phrased from above: because $\int_{\mathbb{X}} dp_w(x) = 1$, we have

$$\begin{aligned}\nabla_w \int p_w(x | w) dx &= \int \nabla_w p_w(x | w) dx &= \int \phi(x) dp_w(x | w) - \nabla_w \log Z(w) \int dp_w(x | w) \\ &= \nabla_w 1 &= 0 \\ \Rightarrow \mathbb{E}_{p_w}(\phi(x)) &= \nabla_w \log Z(w)\end{aligned}$$

- ▶ hence, if we should need to compute $\mathbb{E}_{p_w}(\phi(x))$, we can do so by *differentiating* $\log Z$ wrt. w instead of *integrating* p over x . (actually, we're efficiently re-using someone else's integral)
- ▶ Note that an exponential family forms a *Abelian semigroup* on w :

$$p_w(x | w_1) \cdot p_w(x | w_2) \propto p(x | w_1 + w_2)$$

- ▶ Thus, combining information about x from independent p_w -sources can be done by floating point addition. In this sense, exponential families map inference to addition.

Exponential Families

- ▶ have conjugate priors
- ▶ allow maximum likelihood inference on their parameters from N observations in $\mathcal{O}(N)$, because doing so requires only the sufficient statistics ϕ .
- ▶ allow computation of the *integrals* $\mathbb{E}_{p_w}(\phi(x)) = \nabla_w \log Z(w)$

All of this hinges on the fact that $\log Z(w)$ is (analytically) known.

Can we use exponential families $p_w(x) = e^{\phi(x)^\top w} / Z(w)$ to learn **distributions**, just like we used linear forms $f(x) = \phi(x)^\top w$ to learn **functions**?

Yes! In fact, we can even do **Bayesian** distribution regression. It is called *conjugate prior inference*.



- ▶ Recall previous lectures: **regression on real functions:**

Given $(y_i, x_i)_{i=1, \dots, n}$, and assume $p(y_i | f(x_i)) = \mathcal{N}(y_i; f(x_i), \sigma^2)$ and $f(x) = \phi(x)^\top w$. Notice conjugate Gaussian prior $p(w) = \mathcal{N}(w; \mu, \Sigma)$, get Gaussian posterior $p(w | \mathbf{y}) = \mathcal{N}(\dots)$

- ▶ statistical analysis: interpret negative log posterior as empirical risk $\mathcal{L}_2(w) \propto -\log p(w | \mathbf{y})$. MAP estimate at

$$\hat{f}(x) = \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^n \|y_i - \phi(x_i)^\top w\|^2 + \frac{\sigma^2}{n} \|w\|_\Sigma^2 =: \arg \min_{w \in \mathbb{R}^d} \mathcal{L}_2(w)$$

- ▶ assume $x_i \sim p(x)$, then the Loss approximates an expected log posterior

$$\hat{f} \approx \arg \min_{w \in \mathbb{R}^d} \int \|f(x) - \phi(x)^\top w\|^2 dp(x) + \frac{\sigma^2}{n} \|w\|_\Sigma^2$$

- ▶ thus, for $n \rightarrow \infty$, find function \hat{f} that minimizes the *expected square risk* to f in $\mathcal{H}_\phi = \{f: \mathbb{X} \rightarrow \mathbb{R} \mid f(x) = \phi(x)^\top w\}$.

Definition (Kullback-Leibler divergence)

Let P and Q be probability distributions over \mathbb{X} with pdf's $p(x)$ and $q(x)$, respectively. The **KL-divergence from Q to P** is defined as

$$D_{\text{KL}}(P\|Q) := \int \log \left(\frac{p(x)}{q(x)} \right) dp(x)$$

(I will often write $D_{\text{KL}}(p\|q)$ instead)

Some properties:

- ▶ $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$
- ▶ $D_{\text{KL}}(P\|Q) \geq 0, \forall P, Q$ (**Gibbs' inequality**), and
- ▶ $D_{\text{KL}}(P\|Q) = 0 \Leftrightarrow p \equiv q$ almost everywhere



Solomon Kullback
(1907–1994)



Richard Leibler
(1914–2003)

- ▶ Given $[x_i]_{i=1, \dots, n}$ with $x_i \sim p(x)$, assume

$$p(x) \approx \hat{p}(x | w) = \exp(\phi(x)^\top w - \log Z(w))$$

- ▶ to find \hat{w} , consider

$$\begin{aligned} \hat{w} &= \arg \min_{w \in \mathbb{R}^d} D_{\text{KL}}(p(x) \| \hat{p}(x | w)) = \arg \min_{w \in \mathbb{R}^d} \int [\log p(x) - \log \hat{p}(x | w)] dp(x) \\ &= \arg \min_{w \in \mathbb{R}^d} \underbrace{\int \log p(x) dp(x)}_{-\mathbb{H}(p)} + \mathbb{E}_p(\phi(x))^\top w - \log Z(w) = \arg \min_{w \in \mathbb{R}^d} \mathcal{L}_{\log}(w) \end{aligned}$$

- ▶ Find minimum at $\nabla_w \mathcal{L}_{\log}(w) = 0$, where

$$\mathbb{E}_p(\phi(x)) \approx \frac{1}{n} \sum_{i=1}^n \phi(x_i) = \nabla_w \log Z(w) = \mathbb{E}_{\hat{p}}(\phi(x))$$

- ▶ Given $[x_i]_{i=1, \dots, n}$ with $x_i \sim p(x)$, assume

$$p(x) \approx \hat{p}(x | w) = \exp(\phi(x)^\top w - \log Z(w))$$

- ▶ to find \hat{w} , consider (to regularize, include the conjugate prior. No need to know its normalizer!)

$$\begin{aligned} \hat{w} &= \arg \min_{w \in \mathbb{R}^d} D_{\text{KL}}(p(x) \parallel \hat{p}(x, w)) = \arg \min_{w \in \mathbb{R}^d} \int [\log p(x) - \log \hat{p}(x | w)] dp(x) + \alpha^\top w - \nu \log Z(w) \\ &= \arg \min_{w \in \mathbb{R}^d} \underbrace{\int \log p(x) dp(x)}_{-\mathbb{H}(p)} + \mathbb{E}_p(\phi(x))^\top w - \log Z(w) + \alpha^\top w - \nu \log Z(w) = \arg \min_{w \in \mathbb{R}^d} \tilde{\mathcal{L}}_{\log}(w) \end{aligned}$$

- ▶ Find minimum at $\nabla_w \tilde{\mathcal{L}}_{\log}(w) = 0$, where

$$\mathbb{E}_p(\phi(x)) \approx \frac{1}{n} \sum_{i=1}^n \phi(x_i) = \frac{n + \nu}{n} \nabla_w \log Z(w) - \frac{1}{n} \alpha$$

Full Bayesian Regression on Distributions!

Fitting distributions with exponential families

- ▶ Given $[x_i]_{i=1, \dots, n}$ with $x_i \sim p(x)$, assume

$$p(x) \approx p_w(x | w) = \exp(\phi(x)^\top w - \log Z(w)) \quad \text{and} \quad p_F(w | \alpha, \nu) = \exp(w^\top \alpha - \nu \log Z(w) - \log F(\alpha, \nu))$$

- ▶ compute the posterior on w , using the conjugate prior

$$p(w | x, \alpha, \nu) = \frac{\prod_{i=1}^n p_w(x_i | w) p_F(w | \alpha, \nu)}{\int p(x | w) p(w | \alpha, \nu) dx} = p_F \left(w | \alpha + \sum_i \phi(x_i), \nu + n \right)$$

- ▶ note that $\nabla \nabla p_F(w | \alpha, \nu)|_{w_* = \arg \max p(w | \alpha, \nu)} = -\nu p(w_* | \alpha, \nu) \nabla_w \nabla_w^\top \log Z(w_*)$
- ▶ In the limit $n \rightarrow \infty$, posterior concentrates at w_* with

$$\nabla_w \log Z(w_*) = \frac{\alpha}{n} + \frac{1}{n} \sum_{i=1}^n \phi(x_i) = \mathbb{E}_p(\phi(x)) \quad \text{thus} \quad p_w(x | w_*) = \arg \min_w D_{\text{KL}}(p(x) || p_w(x | w))$$

Learning probability distributions with exponential families

- ▶ Given data x_1, \dots, x_N drawn iid. from unknown $p(x)$, consider *approximating* $p(x) \approx p_w(x | w)$ with an EF
- ▶ The *maximum likelihood* and *MAP* estimates for w can be computed in $\mathcal{O}(N)$
- ▶ If the conjugate prior to p_w (which is an EF) is tractable, it allows full Bayesian inference
- ▶ asymptotically, the posterior concentrates around the maximum likelihood estimate, which is the minimizer of the KL-divergence $D_{\text{KL}}(p || p_w)$ within the exponential family.



Wouldn't you want to join this club?

Build your own exponential family!



1. choose features (come up with grand motivation: attraction/repulsion)

$$\phi(x) = \begin{bmatrix} -x^2 \\ -x^{-2} \end{bmatrix}$$

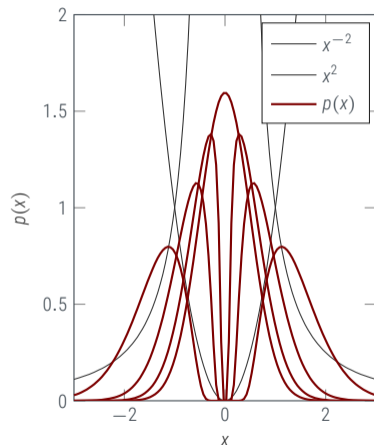
2. solve integral (the hard bit)

$$Z(w) = \int_0^{\infty} \exp(-w_1 x^2 - w_2/x^2) dx = \sqrt{\frac{\pi}{w_1}} e^{-2\sqrt{w_1 w_2}}$$

3. profit! The **bagel-distribution**!

$$\mathcal{H}(x; w) = \sqrt{\frac{w_1}{\pi}} e^{2\sqrt{w_1 w_2}} e^{-w_1 x^2 - w_2/x^2}$$

4. don't know the conjugate prior, though. :(



► We need

$$\log Z(w) = -2(w_1 w_2)^{1/2} - \frac{1}{2} \log w_1 + \frac{1}{2} \log \pi$$

$$-\nabla_w \log Z(w) = \begin{bmatrix} \sqrt{\frac{w_2}{w_1}} + \frac{1}{2w_1} \\ \sqrt{\frac{w_1}{w_2}} \end{bmatrix} \stackrel{!}{=} -\frac{1}{n} \sum_i \begin{bmatrix} x_i^2 \\ x_i^{-2} \end{bmatrix} =: \begin{bmatrix} \bar{\mu} \\ \bar{\omega} \end{bmatrix}$$

$$\Rightarrow \hat{w}_1 = \frac{1}{2(\bar{\mu} - \bar{\omega})} \quad \hat{w}_2 = \frac{\hat{w}_1}{\bar{\omega}^2}$$

Summary:

- ▶ Conjugate Priors allow analytic inference of “nuisance parameters” in probabilistic models
- ▶ Exponential Families
 - ▶ guarantee the existence of conjugate priors, although not always tractable ones
 - ▶ allow analytic MAP inference from only a finite set of *sufficient statistics*

Conjugate prior inference with exponential families is a form of Bayesian **regression on distributions**. Gaussian process inference, in this sense, is inference on the unknown mean of a Gaussian distribution.

- ▶ The hardest part is finding the normalization constant. In fact, finding the normalization constant is *the only* hard part.
- ▶ Exponential families are a way to turn someone else’s integral into an inference algorithm!

