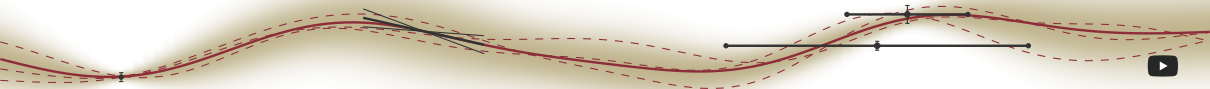
Probabilistic Machine Learning
Lecture 15
Exponential Families

Philipp Hennig

15 June 2020

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

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

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

Practical probabilistic inference is chiefly a *computational* task.

The Toolbox

Framework:

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

Modelling:

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

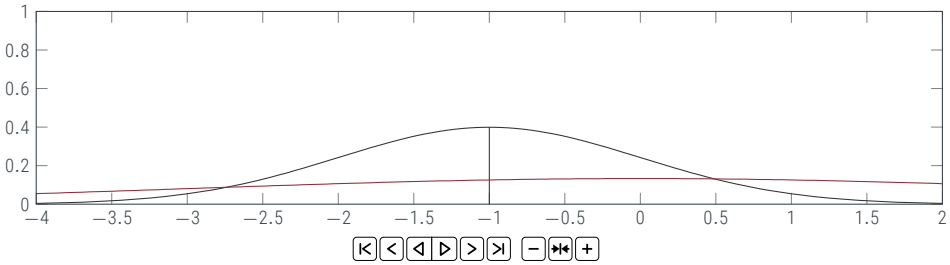
Computation:

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

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

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

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



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

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

$$p(x \mid f) = \prod_{i=1}^{n} f^x \cdot (1-f)^{1-x} \qquad x \in \{0;1\}$$

$$= f^{n_1} \cdot (1-f)^{n_0} \qquad n_0 := n - n_1$$

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

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

Pierre Simon, marquis de Laplace, 1749–1827

image: Deutsches Museum München

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} f_{x_i} \qquad x \in \{0; \dots, K\}$$

$$= \prod_{k=1}^{K} f_k^{n_k} \qquad n_k := |\{x_i \mid x_i = k\}|$$

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

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

Peter Gustav Lejeune Dirichlet
(1805–1859)

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

$$p(\sigma) = ?$$

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

$$p(\sigma) = ?$$

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

image: ETH Zürich

$$
\begin{aligned}
p(\boldsymbol{x} \mid \sigma) &= \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2) \\
p(\sigma) &= ? \\
\log p(\boldsymbol{x} \mid \sigma) &= -\frac{1}{2} \log \sigma^2 - \frac{1}{2}(x-\mu)^2 \cdot \frac{1}{\sigma^2} - \frac{1}{2} \log 2\pi \\
\log p(\sigma \mid \alpha, \beta) &= (\alpha + 1) \log \sigma^{-2} - \beta \cdot \frac{1}{\sigma^2} - Z(\alpha, \beta) \\
p(\sigma \mid \alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^{-2})^{\alpha+1} \mathrm{e}^{-\beta\sigma^{-2}} =: \mathcal{G}(\sigma^{-2}; \alpha, \beta) \\
p(\sigma \mid \alpha, \beta, \boldsymbol{x}) &= \mathcal{G}\left(\sigma^{-2}; \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_i (x_i - \mu)^2\right)
\end{aligned}
$$

Daniel Bernoulli (1700–1782)

Analytic Hierarchical Bayesian Inference

Inferring Mean *and* Co-Variance of a Gaussian

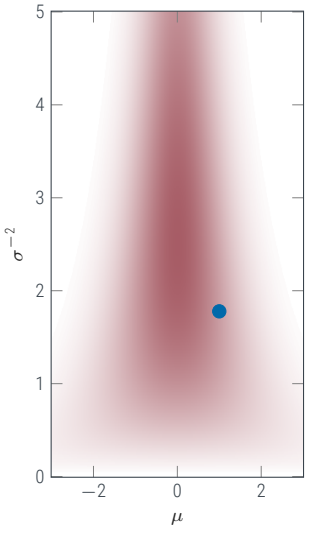
$$p(\boldsymbol{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

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

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

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

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



Analytic Hierarchical Bayesian Inference

Inferring Mean *and* Co-Variance of a Gaussian

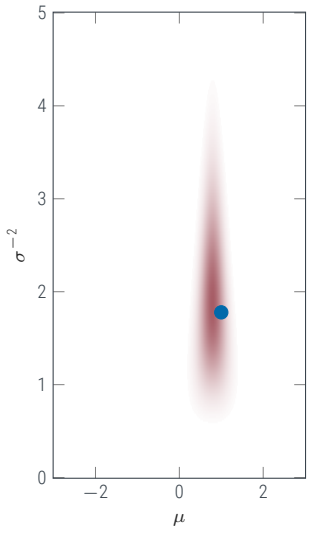
$$p(\boldsymbol{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

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

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

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

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



Definition (Conjugate Prior)

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

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

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

$$p_w(x) = h(x) \exp\left[\phi(x)^\intercal w - \log Z(w)\right] = \frac{h(x)}{Z(w)} e^{\phi(x)^\intercal w} = p(x \mid w)$$

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

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

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

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	$\phi(x) = [1, x]^\intercal$	$\mathbb{X} = \mathbb{R}$	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	$\phi(X) = [X, \log \lvert X \rvert]$	$\mathbb{X} = \{X \in \mathbb{R}^{N \times N} \mid v^\intercal X v \geq 0 \forall v \in \mathbb{R}^N\}$	covariances
Gauss	$\phi(X) = [X, XX^\intercal]$	$\mathbb{X} = \mathbb{R}^N$	functions
Boltzmann	$\phi(X) = [X, \text{triag}(XX^\intercal)]$	$\mathbb{X} = \{0; 1\}^N$	thermodynamics

Exponential Families have Conjugate Priors

but the prior's normalization constant can be tricky

- Consider the exponential family $p_w(x \mid w) = h(x) \exp\left[\phi(x)^\intercal w - \log Z(w)\right]$
- its conjugate prior is the exponential family $\quad F(\alpha, \nu) = \int \exp(\alpha^\intercal w - \nu \log Z(w))\, dw$

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

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

- and the predictive is

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

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

Exponential Families Allow Efficient Maximum Likelihood Estimation

why they're called sufficient statistics

- Consider the exponential family

$$p_w(x \mid w) = \exp\left[\phi(x)^\intercal w - \log Z(w)\right]$$

- for iid data:

$$p_w(x_1, x_2, \ldots, x_n \mid w) = \prod_i^n p_w(x_i \mid w) = \exp\left(\sum_i^n \phi^\intercal(x_i) w - n \log Z(w)\right)$$

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

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

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

Other great properties

exponential families make many things easy

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

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

- 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:

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

- 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 $p_w(x) = e^{\phi(x)^\intercal w}/Z(w)$ to learn distributions,
just like we used linear forms $f(x) = \phi(x)^\intercal w$ to learn functions?

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

Recap: Regression on Functions

The ℓ_2 loss

- Recall previous lectures: **regression** on **real functions**:
  Given $(y_i, x_i)_{i=1,\dots,n}$, and assume $p(y_i \mid f(x_i)) = \mathcal{N}(y_i; f(x_i), \sigma^2)$ and $f(x) = \phi(x)^\intercal w$. Notice conjugate Gaussian prior $p(w) = \mathcal{N}(w; \mu, \Sigma)$, get Gaussian posterior $p(w \mid \boldsymbol{y}) = \mathcal{N}(\dots)$
- statistical analysis: interpret negative log posterior as empirical risk $\mathcal{L}_2(w) \propto -\log p(w \mid \boldsymbol{y})$. MAP estimate at

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

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

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

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

Interlude: KL divergence

The most mis-spelled names in statistics

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

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

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

Some properties:

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

Solomon Kullback
(1907–1994)

Richard Leibler
(1914–2003)

Maximum Likelihood Regression on Distributions!

Fitting distributions with exponential families

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

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

- to find $\hat{w}$, consider

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} D_{\mathrm{KL}}(p(x) \| \hat{p}(x \mid w)) = \arg\min_{w \in \mathbb{R}^d} \int [\log p(x) - \log \hat{p}(x \mid w)]\, dp(x)$$

$$= \arg\min_{w \in \mathbb{R}^d} \underbrace{\int \log p(x)\, dp(x)}_{-\mathbb{H}(p)} + \mathbb{E}_p(\phi(x))^\intercal w - \log Z(w) = \arg\min_{w \in \mathbb{R}^d} \mathcal{L}_{\log}(w)$$

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

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

MAP Regression on Distributions!

Fitting distributions with exponential families

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

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

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

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} D_{\mathrm{KL}}(p(x) \| \hat{p}(x, w)) = \arg\min_{w \in \mathbb{R}^d} \int [\log p(x) - \log \hat{p}(x \mid w)]\, dp(x) + \alpha^\intercal w - \nu \log Z(w)$$

$$= \arg\min_{w \in \mathbb{R}^d} \underbrace{\int \log p(x)\, dp(x)}_{-\mathbb{H}(p)} + \mathbb{E}_p(\phi(x))^\intercal w - \log Z(w) + \alpha^\intercal w - \nu \log Z(w) = \arg\min_{w \in \mathbb{R}^d} \tilde{\mathcal{L}}_{\log}(w)$$

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

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

Full Bayesian Regression on Distributions!

Fitting distributions with exponential families

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

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

- compute the posterior on w, using the conjugate prior

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

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

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

Learning probability distributions with exponential families

- Given data $x_1, \dots, x_N$ drawn iid. from unknown $p(x)$, consider *approximating* $p(x) \approx p_w(x \mid w)$ 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 $D_{\mathrm{KL}}(p||p_w)$ within the exponential family.

Wouldn't you want to join this club?

Build your own exponential family!






Building our own Exponential Family

just for fun

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

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

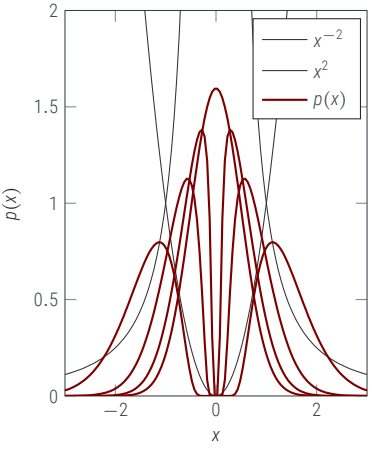
2. solve integral (the hard bit)

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

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. :(

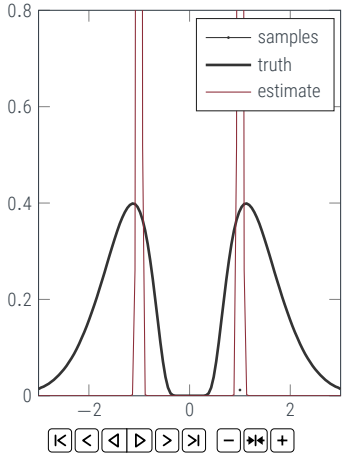


- We need

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

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

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



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!