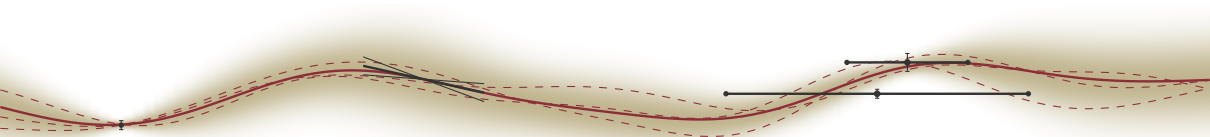
Probabilistic Inference and Learning
Lecture 11
Exponential Families

Philipp Hennig

21 November 2018

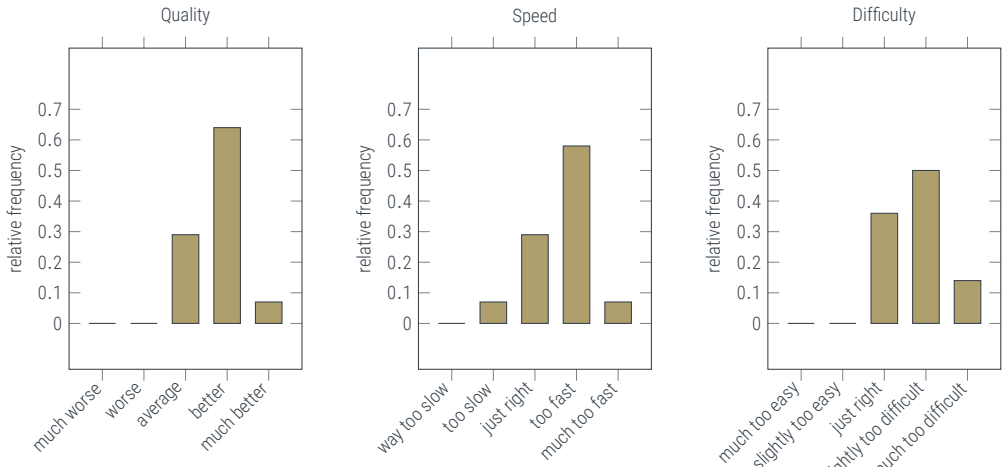
Eberhard Karls
Universität
Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you did not like:

- the exercises are too hard

Things you did not understand:

- steady-state
- what kind of object is an SDE?
- too many new concepts in one lecture

Things you enjoyed:

- **room for discussion**
- recap
- the exercises are of a good difficulty level

Overview of Lectures so far:

0. Introduction to Reasoning under Uncertainty
1. Probabilistic Reasoning
2. Probabilities over Continuous Variables
3. Gaussian Probability Distributions
4. Gaussian Parametric Regression
5. More on Parametric Regression
6. Gaussian Processes
7. More on Kernels & GPs
8. A practical GP example
9. Markov Chains, Time Series, Filtering
10. Classification
11. Empirical Example of Classification
12. Bayesianism and Frequentism
13. Stochastic Differential Equations
14. Exponential Families

Today:

- A Toolbox of Probability Distributions
- A clean framework for efficient probabilistic inference, with some caveats

Why is this hard?

The computational challenge in Bayesian Inference

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

Why is this hard?

The computational challenge in Bayesian Inference

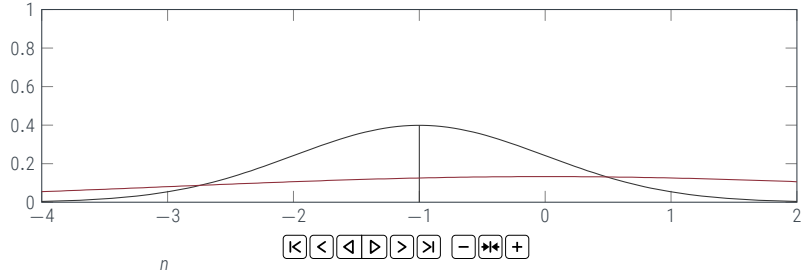
$$p(x \mid y) = \frac{p(y \mid x)p(x)}{\int p(y \mid x)p(x)\, dx}$$

- 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
- even if we just want $\arg\max_x p(x \mid y)$, the optimization may be hard

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…



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

$$p(\mu \mid \mu_0, \Sigma_0) = \mathcal{N}(\mu; \mu_0, \Sigma_0)$$

$$p(\mu \mid x) = \frac{p(x \mid \mu) p(\mu \mid \mu_0, \Sigma_0)}{p(x)}$$

$$= \mathcal{N}(\mu; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_i x_i), (\Sigma_0^{-1} + n\Sigma^{-1})^{-1})$$

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

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

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

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

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)

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

$$p(\sigma) = ?$$

Analytic Hierarchical Bayesian Inference

Inferring the (Co-) Variance of a Gaussian

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

$$p(\sigma) = ?$$

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

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

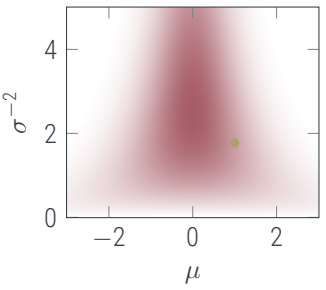
$$p(\boldsymbol{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

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

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

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

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



Analytic Hierarchical Bayesian Inference

Inferring Mean *and* Co-Variance of a Gaussian

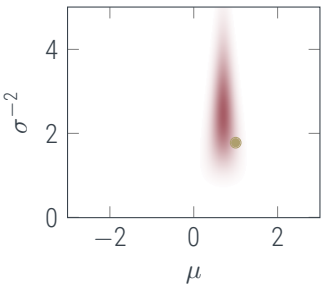
$$p(\boldsymbol{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

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

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

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

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



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

$$p_w(x) = \exp\left[\phi(x)^\intercal w - \log Z(w)\right] = \frac{1}{Z(w)} e^{\phi(x)^\intercal w} = p(x \mid w)$$

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**.

Definition (Exponential Family, formal definition, only if you care for it)

Consider a σ-finite measure μ on the Borel sets of $\mathbb{R}^d$. Consider the convex hull of the support set of μ, and let $\mathbb{X}$ be the interior of this convex set, assumed to be nonempty and open in $\mathbb{R}^d$. For $w \in \mathbb{R}^d$, define $Z(w) = \ln \int e^{w \cdot x} \, d\mu(x)$ and let $W = \{w \mid Z(w) < \infty\}$. Hölder's inequality shows that Θ is a convex set, called the **natural parameter** space. Assume further that W is nonempty and open in $\mathbb{R}^d$. The **exponential family** $\{P_w\}$ of probability measures **through** μ is determined by

$$dP_w(x) = e^{x \cdot w - Z(w)} \, d\mu(x).$$

Nb: Relative to our more practically-minded definition on the previous slide, this definition absorbs the sufficient statistics ϕ into the definition of X.

A Family Meeting

incomplete list of exponential families

Bernoulli	$\phi(x) = [x]$	$\mathbb{X} = \{0; 1\}$
Poisson	$\phi(x) = [x]$	$\mathbb{X} = \mathbb{R}_+$
Laplace	$\phi(x) = [1, x]^\intercal$	$\mathbb{X} = \mathbb{R}$
χ^2	$\phi(x) = [x, -\log x]$	$\mathbb{X} = \mathbb{R}$
Dirichlet	$\phi(x) = [\log x]$	$\mathbb{X} = \mathbb{R}_+$
Euler (Γ)	$\phi(x) = [x, \log x]$	$\mathbb{X} = \mathbb{R}_+$
Wishart	$\phi(X) = [X, \log \lvert X \rvert]$	$\mathbb{X} = \{X \in \mathbb{R}^{N \times N} \mid v^\intercal X v \geq 0 \forall v \in \mathbb{R}^N\}$
Gauss	$\phi(X) = [X, XX^\intercal]$	$\mathbb{X} = \mathbb{R}^N$
Boltzmann	$\phi(X) = [X, \text{triag}(XX^\intercal)]$	$\mathbb{X} = \{0; 1\}^N$

Exponential Families have Conjugate Priors

but the prior's normalization constant can be tricky

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

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

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

- and the predictive is

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

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

Example: The Gaussian distribution

There it is, again!

$$
\begin{aligned}
p_w(x \mid w) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}(x; \mu, \sigma^2) \\
&= \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right) \\
&= \exp\left(\begin{bmatrix} x & -1/2\, x^2 \end{bmatrix} \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2}\right)\right) \\
&= \exp\left(\begin{bmatrix} \phi_1(x) & \phi_2(x) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \underbrace{\left(\frac{w_1^2}{2w_2} - \frac{1}{2}\log w_2 + \log\sqrt{2\pi}\right)}_{\log Z(w)}\right)
\end{aligned}
$$

- The natural parameters are **precision** σ^{-2} and **precision-adjusted mean** $\mu\sigma^{-2}$
- the sufficient statistics are the sample *mean* and $(-1/2)$ *variance*
- The conjugate prior is the *Normal-Gamma*, the predictive marginal is the *Student-t* distribution

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.

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 lectures 4–8: **regression** on **real functions**:
  Given $(y_i, x_i)_{i=1,\dots,n}$, and assuming $y_i = f(x_i)$, assume $f(x) = \phi(x)^\intercal w$ and find

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

- can interpret $\exp -\mathcal{L}_2(w)$ as the (unnormalized) **posterior** on w from the Gaussian prior $\exp(-\|w\|_\Sigma^2)$.
- assume $x_i \sim p(x)$, then the Loss approximates an expected log posterior

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

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)

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

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

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

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

- If the conjugate prior has tractable normalizer, we can do full Bayesian inference on w!

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

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

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

- If the conjugate prior is intractable, we can still do MAP:

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

$$\text{equates to setting} \quad w = \arg\max_w \frac{p_w(x \mid w) p_\alpha(w \mid \alpha, \nu)}{\int p_w(x \mid w) p_\alpha(w \mid \alpha, \nu)\, dw}$$

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)

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

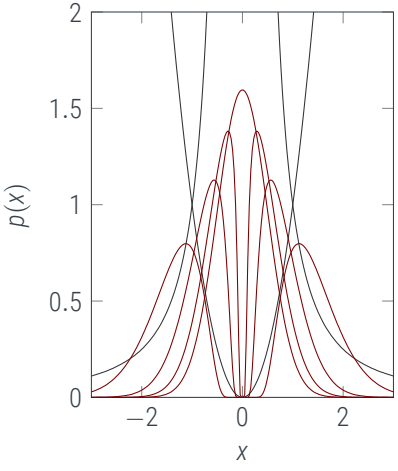
2. solve integral (the hard bit)

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

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. :(

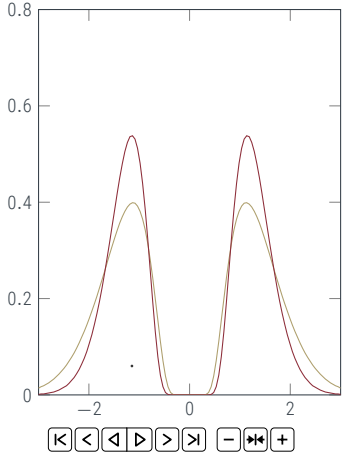


- We need

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

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

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



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!

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!

If you happen to know (a good approximation to)

$$Z = \int_{-\infty}^{\infty} \exp\left[\sum_i w_i \exp\left(-\frac{(x - c_i)^2}{2\lambda^2}\right)\right] \mathrm{d}x$$

I may have a job for you …