Probabilistic Machine Learning

Lecture 05

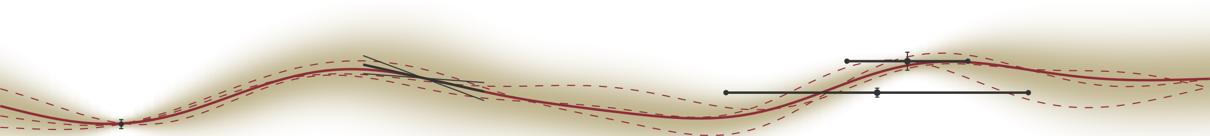
Exponential Families

Philipp Hennig

04 May 2023

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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



Bonus Points

it's ok not to get everything right

- "bonus points" means not everyone should get them
- otherwise, the exam just gets a steeper grading curve, and nobody gains anything
- *It's ok* to not be able to finish the exercise. You do not need all the bonus points to finish the exam!
- please do not post solutions on the forum. Sometimes they even add confusion!
- keep in mind that some things can actually be computed by hand. Auto-diff and optimization are powerful tools, but they do have computational and coding cost.

The skeleton of ML

probabilistic inference

$$p(\boldsymbol{w} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{w})p(\boldsymbol{w})}{\int p(\boldsymbol{x} \mid \boldsymbol{w})p(\boldsymbol{w})\, d\boldsymbol{w}} \overset{\text{“often”}}{=} \frac{\prod_{n=1}^{N} p(x_n \mid \boldsymbol{w})p(\boldsymbol{w})}{\int \prod_{n=1}^{N} p(x_n \mid \boldsymbol{w})p(\boldsymbol{w})\, d\boldsymbol{w}}$$

This is very abstact. It would be nice to reduce this to

- get some **data** $\boldsymbol{x} \in \mathbb{R}^{N \times D}$ (N: “batch dim”, D: “input dim”)
- compute some **statistics** $\phi(\boldsymbol{x}) \in \mathbb{R}^F$ (note this can consume either or both axes of $\boldsymbol{x}$) that capture the **algebraic structure** of $p(\boldsymbol{x} \mid \boldsymbol{w})$
- *somehow* compute $p(\boldsymbol{w} \mid \boldsymbol{x})$, using only $\phi(\boldsymbol{x})$ and nothing else about the data.

Conjugate Priors

Mapping Bayesian inference to tractable code

Definition (Conjugate Prior)

Let $\boldsymbol{x}$ and $\boldsymbol{w}$ be a data-set and a variable to be inferred, respectively, connected by the likelihood $p(\boldsymbol{x} \mid \boldsymbol{w}) = \ell(\boldsymbol{x}; \boldsymbol{w})$. A **conjugate prior to ℓ for $\boldsymbol{w}$** is a probability measure with pdf $p(\boldsymbol{w}) = g(\boldsymbol{w}; \theta)$, such that

$$p(\boldsymbol{w} \mid \boldsymbol{x}) = \frac{\ell(\boldsymbol{x}; \boldsymbol{w}) g(\boldsymbol{w}; \theta)}{\int \ell(\boldsymbol{x}; \boldsymbol{w}) g(\boldsymbol{w}; \theta)\, d\boldsymbol{w}} = g(\boldsymbol{w}; \theta + \phi(\boldsymbol{x})).$$

That is, such that the posterior arising from ℓ is of the same functional form as the prior, with updated parameters arising by **adding** some **sufficient statistics** of the observation $\boldsymbol{x}$ to the prior's 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, by mapping it ot
  - "data processing" $\phi(\boldsymbol{x}) : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^F$
  - "inference" $g(\boldsymbol{w}; \theta + \phi(\boldsymbol{x}))$, simple (floating-point) addition yields the full, normalized posterior!
  - (the implicit assumption in this formulation is that g can be computed in decent time)
- How can we construct them in general?

Exponential Families

Exponentials of a Linear Form

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 $\phi : \mathbb{X} \rightharpoonup \mathbb{R}^d$ is called the **sufficient statistics**. The parameters $w \in \mathbb{R}^d$ are the **natural parameters** of p_w. The normalization constant $Z(w) : \mathbb{R}^d \rightharpoonup \mathbb{R}$ is the **partition function**. The function $h(x) : \mathbb{X} \rightharpoonup \mathbb{R}_+$ is the **base measure**. For notational convenience, it can be useful to re-parametrize the natural parameters w as $w := \eta(\theta)$ in terms of *canonical parameters* θ.

A Family Meeting

Exponential families provide the probabilistic analogue to data types

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	$\phi(X) = [X, \operatorname{triag}(XX^\intercal)]$	$\mathbb{X} = \{0; 1\}^N$	thermodynamics

Note: *Each row* of this table is *one* exponential family. Some authors (e.g. Murphy) call the entire table "the exponential family". We will *not* do this.

CODE

– Thanks to Marvin Pförtner for pair coding –

Exponential Families have Conjugate Priors

but their normalization constant may not be known

- 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 $F(\alpha, \nu) := \int \exp(\alpha^\intercal w - \nu \log Z(w))\, dw$

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

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

Computing $F(\alpha, \nu)$ can be tricky. But **if** we have it, it completely automates Bayesian inference! Finding F is thus **the** challenge when constructing an EF.

Exponential Families have Conjugate Priors

but their normalization constant may not be known

- 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 $F(\alpha, \nu) := \int \exp(\alpha^\intercal w - \nu \log Z(w))\, dw$

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

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

- and the predictive is

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

$$= h(x) \frac{F(\phi(x) + \alpha, \nu + 1)}{F(\alpha, \nu)}$$

Computing $F(\alpha, \nu)$ can be tricky. But **if** we have it, it completely automates Bayesian inference! Finding F is thus **the** challenge when constructing an EF.

Atomic inference machines

Exponential families with tractable conjugate priors make Bayesian ML tractable

- Once we decide to use a particular **exponential family** $(h(\boldsymbol{x}), \phi(\boldsymbol{x}))$ as the **model** for some data $\boldsymbol{x}$, we *automatically get* a *conjugate prior* in the form of another exponential family
- in principle, this provides the means to do analytic Bayesian inference, **if** we can compute the partition function $Z(w)$ of the likelihood and $F(\alpha, \nu)$, that of the conjugate prior.
- This solves the *learning* problem, i.e. how to extract information from the data. It does not guarantee that we will be able to compute moments, or sample from the posterior, but these can be encapsulated to subroutines, now that the data is dealt with.

CODE

Example: The Gaussian distribution

There it is, again!

$$
\begin{aligned}
p_w(x \mid w) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}(x;\mu,\sigma^2) \\
&= \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right) \\
&= \exp\left(\begin{bmatrix} x & -{}^1\!/_2\, x^2 \end{bmatrix} \begin{bmatrix} {}^\mu\!/_{\sigma^2} \\ {}^1\!/_{\sigma^2} \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2}\right)\right) \\
&= \exp\left(\begin{bmatrix} \phi_1(x) & \phi_2(x) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \underbrace{\frac{1}{2}\left(\frac{w_1^2}{w_2} - \log w_2 + \log(2\pi)\right)}_{\log Z(w)}\right)
\end{aligned}
$$

- The natural parameters are **precision** σ^{-2} and **precision-adjusted mean** $\mu\sigma^{-2}$
- the sufficient statistics are the *first two (non-central) sample moments* of the data
- The conjugate prior is the *Normal-Gamma*, the predictive marginal is the *Student-t* distribution

Analytic Hierarchical Bayesian Inference

Inferring Mean *and* Co-Variance of a Gaussian

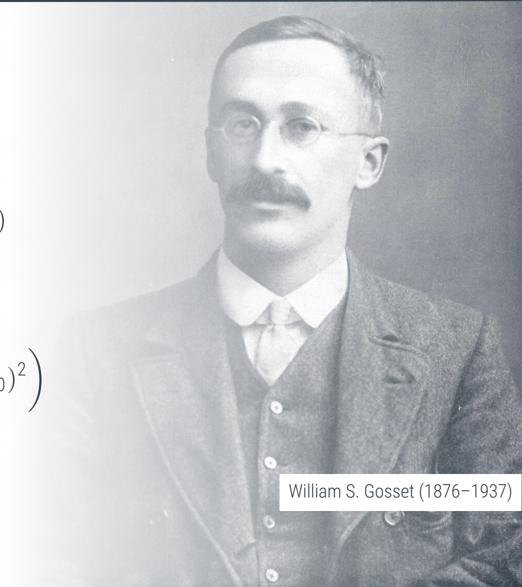
$$p(\boldsymbol{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

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

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

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

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



William S. Gosset (1876–1937)

What if we don’t know the log partition function?

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:

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

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

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

- hence, collect **statistics** of ϕ, compute $\nabla_w \log Z(w)$ and solve the above for w
- this may or may not be possible analytically, but in any case it encapsulates the data!

Example for Maximum Likelihood Estimation

the Gaussian case

Example: Assume we observe samples x_i drawn iid. from

$$p(\boldsymbol{x} \mid \mu, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2) = \prod_{i}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Remember that the Gaussian is an EF with

$$\boldsymbol{w} := \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{1}{\sigma^2} \end{bmatrix} \qquad \phi(x) := \begin{bmatrix} x \\ -\frac{1}{2}x^2 \end{bmatrix} \qquad \log Z(\boldsymbol{w}) := \frac{1}{2}\left(\frac{w_1^2}{w_2} - \log w_2 + \log(2\pi)\right)$$

so we find the maximum likelihood estimate by computing

$$\nabla_{\boldsymbol{w}} \log Z(\boldsymbol{w}) = \begin{bmatrix} \frac{w_1}{w_2} \\ -\frac{1}{2}\left(\frac{w_1^2}{w_2^2} + \frac{1}{w_2}\right) \end{bmatrix} = \begin{bmatrix} \mu \\ -\frac{1}{2}(\mu^2 + \sigma^2) \end{bmatrix} \overset{!}{=} \overline{\phi(x)} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_i \\ -\frac{1}{2}x_i^2 \end{bmatrix}$$

hence setting $\hat{\mu} = \bar{x} =: \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \hat{\mu}^2$

Why stop at maximum likelihood?

- thanks to auto-diff, we can do *approximate* **full** Bayesian inference

Summary: Exponential Families

- reduce Bayesian inference to
  - *modelling*: designing / computing the sufficient statistics $\phi(\boldsymbol{x})$ and partition function $Z(\boldsymbol{w})$
  - computing the posterior: essentially evaluating the log partition function F of the conjugate prior
- if F is not tractable, we can still do *Laplace* approximations:
  - find the *mode* $\hat{w}$ of the posterior, by solving the *root-finding* problem

$$\nabla_w \log p(w \mid \boldsymbol{x}) = \frac{\alpha + \sum_{i=1}^{n} \phi(x_i)}{\nu + n}$$

  - evaluate the *Hessian* $\Psi = \nabla_w \nabla_w^\intercal \log p(w \mid \boldsymbol{x})$ at $\hat{w}$
  - approximate the posterior as $\mathcal{N}(w; \hat{w}, -\Psi^{-1})$

Please cite this course, as

```
@techreport{Tuebingen_ProbML23,
    title =
    {Probabilistic Machine Learning},
    author = {Hennig, Philipp},
    series = {Lecture Notes
        in Machine Learning},
    year = {2023},
    institution = {Tübingen AI Center}}
```

Laplace approximations show that Bayesian inference is not “about the prior” at all, but rather about *capturing the (local) geometry of the likelihood (around the mode)*. Uncertainty is about tracking all possible solutions at once (or at least as many as possible), not one single estimate.