# Exponential families

## CPSC 440/550: Advanced Machine Learning

`cs.ubc.ca/~dsuth/440/23w2`

University of British Columbia, on unceded Musqueam land

2023-24 Winter Term 2 (Jan–Apr 2024)

# Last time: Approximate inference

- Laplace approximation: simple way to find a Gaussian approximation to posterior
  - Fast and easy, but not always accurate
- Rejection sampling: generate exact samples from complicated distributions
  - Tends to reject too many samples in high dimensions
- Importance sampling: re-weights samples from the wrong distribution
  - Tends to have high variance in high dimensions

# Previously: Density Estimation with Categorical/Gaussian Distributions

- We have discussed density estimation with categorical and Gaussian distribution
  - Bernoulli is a special case of categorical (up to notation changes)

- These distributions have a lot of nice properties for learning/inference
  - NLL is convex, and MLE has closed-form (statistics in training data)
  - A conjugate prior exists, so posterior is prior with "updated hyper-parameters"

- But these distributions make restrictive assumptions:
  - Categorical assumes categories are unordered, non-hierarchical, and finite
  - Gaussian assumes symmetry, full support, no outliers, uni-modal

- Many alternatives to categorical/Gaussian exist (examples later)
  - Alternatives that are in the exponential family maintain nice properties

# Exponential Family: Definition

- General form of exponential family likelihood for data $x$ with parameters $\theta$ is

$$p(x \mid \theta) = \frac{h(x) \exp(\eta(\theta)^{\mathsf{T}} s(x))}{Z(\theta)}$$

- The value $s(x)$ is the vector of sufficient statistics
  - $s(x)$ tells us everything that is relevant to $\theta$ about the data point $x$
- The parameter function $\eta$ controls how parameters $\theta$ interact with the statistics
  - We'll focus on $\eta(\theta) = \theta$, which is called the canonical form
- The support function $h$ contains terms that don't depend on $\theta$
  - Also called the base measure
- The normalizing constant $Z$ ensures it sums/integrates to 1 over $x$
  - Also called the partition function

# Bernoulli as Exponential Family

- Is Bernoulli in the exponential family for some parameters $w$?

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \, \mathbb{1}(x \in \{0,1\}) \stackrel{?}{=} \frac{h(x) \exp(\eta(\theta)^\top F(x))}{Z(\theta)}$$

- To get an exponential, take log of exp (cancelling operations),

$$\begin{aligned} p(x \mid \theta) &= \theta^x (1-\theta)^{1-x} \, \mathbb{1}(x \in \{0,1\}) = \exp(\log(\theta^x (1-\theta)^{1-x})) \, \mathbb{1}(x \in \{0,1\}) \\ &= \exp(x \log \theta + (1-x) \log(1-\theta)) \, \mathbb{1}(x \in \{0,1\}) \\ &= (1-\theta) \exp\left(x \log\left(\frac{\theta}{1-\theta}\right)\right) \mathbb{1}(x \in \{0,1\}) \end{aligned}$$

- The sufficient statistic is $s(x) = x$; normalizing constant is $Z(\theta) = 1/(1-\theta)$
- The parameter function is $\eta(\theta) = \log(\theta/(1-\theta))$ (the log odds)
  - Not in canonical form. Canonical form would use log odds directly as the parameter
- The support function is $h(x) = \mathbb{1}(x \in \{0,1\})$ – says if we're "in the support"
- There are also other ways to write Bernoulli as an exponential family

# Gaussian as Exponential Family

- Writing univariate Gaussian as an exponential family:

$$
\begin{aligned}
p(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-(x-\mu)^2/2\sigma^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-x^2/2\sigma^2 + \mu x/\sigma^2 - \mu^2/2\sigma^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\mu^2/2\sigma^2\right)}{\sigma} \exp\left(\begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} x \\ x^2 \end{bmatrix}\right)
\end{aligned}
$$

- The sufficient statistics are $x$ and $x^2$, and parameters are $\mu/\sigma^2$ and $-1/2\sigma^2$
- The normalizing constant is $\sigma \exp(\mu^2/2\sigma^2)$, and support is $1/\sqrt{2\pi}$

- Again, there is more than one way to represent as an exponential family
  - If $\sigma^2$ is fixed, then $x/\sigma^2$ is the sufficient statistic and $\mu$ is canonical

# Learning with Exponential Families

- With $n$ IID examples and canonical parameters $\theta$, the likelihood is

$$
\begin{aligned}
p(\mathbf{X} \mid \theta) &= \prod_{i=1}^{n} h(x^{(i)}) \frac{\exp(\theta^\mathsf{T} s(x^{(i)}))}{Z(\theta)} \\
&= \frac{1}{Z(\theta)^n} \exp\left(\theta^\mathsf{T} \sum_{i=1}^{n} s(x^{(i)})\right) \prod_{j=1}^{n} h(x^i) \\
&= \frac{\exp(\theta^\mathsf{T} s(\mathbf{X}))}{Z(\theta)^n} \prod_{j=1}^{n} h(x^{(i)}),
\end{aligned}
$$

  with sufficient statistics $s(\mathbf{X}) = \sum_{i=1}^{n} s(x^i)$
- $s(\mathbf{X})$ contains everything relevant for learning – can throw away the actual data
  - For Gaussians, only knowledge of data we need is $\sum_{i=1}^{n} x^{(i)}$ and $\sum_{i=1}^{n} (x^{(i)})^2$
  - No point in using SGD: just compute $s$ on each example once
  - Exponential families are the *only* class of distributions with a finite sufficient statistic

# Learning with Exponential Families

- With iid data and canonical $\theta$, NLL is $f(\theta) = -\theta^\mathsf{T} s(\mathbf{X}) + n \log Z(\theta) + \text{const}$
- The gradient divided by $n$ (average NLL) for a feature $j$ is

$$\begin{aligned}
\frac{1}{n}\nabla_{\theta_j} f(\theta) &= -\frac{1}{n} s_j(\mathbf{X}) + \frac{1}{Z(\theta)} \nabla_{\theta_j} Z(\theta) \\
&= -\frac{1}{n} s_j(\mathbf{X}) + \frac{1}{Z(\theta)} \nabla_{\theta_j} \int h(x) \exp\left(\theta^\mathsf{T} s(x)\right) \mathrm{d}x \quad \text{(use } \textstyle\sum \text{ for discrete } x\text{)} \\
&= -\frac{1}{n} s_j(\mathbf{X}) + \int_x h(x) \frac{\exp(\theta^\mathsf{T} s(\mathbf{X}))}{Z(\theta)} s_j(\mathbf{X}) \, \mathrm{d}x \quad \text{(w/ conditions)} \\
&= -\frac{1}{n} s_j(\mathbf{X}) + \int_x p(x \mid \theta) s_j(x) \mathrm{d}x \\
&= -\mathop{\mathbb{E}}_{X \sim \text{data}} [s_j(X)] + \mathop{\mathbb{E}}_{X \sim \text{model } p_\theta} [s_j(X)]
\end{aligned}$$

- The stationary points where $\nabla f(\theta) = 0$ correspond to moment matching:
  - Set parameters $\theta$ so that expected sufficient statistics equal to statistics in data
  - This is the source of the simple/intuitive closed-form MLEs we've seen so far

# Convexity and Entropy in Exponential Families

bonus!

- If you take the second derivative of the NLL you get

$$\nabla^2 f(\theta) = \text{Cov}[s(X)],$$

  the covariance of the sufficient statistics
  - Covariances are positive semi-definite, $\text{Cov}[s(X)] \succeq 0$, so NLL is convex
  - This is why "setting the gradient to zero and solve for $\theta$" gives MLE
- Higher-order derivatives give higher-order moments
  - We call $\log(Z)$ the cumulant function

- Can show MLE maximizes entropy over all distributions that match moments
  - Entropy is a measure of "how random" a distribution is
  - So Gaussian is "most random" distribution that fits means and covariance of data
    - Or you can think of this as Gaussian makes "least assumptions"
  - Details for special case of $h(x) = 1$ in bonus slides

# Conjugate Priors in Exponential Family

- Exponential families in canonical form are guaranteed to have conjugate priors
- For example, we could choose a prior like

$$p(\theta \mid \alpha) \propto \frac{\exp(\theta^{\mathsf{T}}\alpha)}{Z(\theta)^k}$$

  - $\alpha$ is "pseudo-counts" for the sufficient statistics
  - $k$ modifies the stength of the prior ($Z$ above is normalizer for the likelihood)
  - For fixed $k$, itself an exp. family in $\theta$: $s(\theta) = \theta$, parameter $\alpha$, base measure $Z(\theta)^{-k}$
- Then the posterior has the same form,

$$p(\theta \mid \mathbf{X}, \alpha) \propto \frac{\exp(\theta^{\mathsf{T}}(s(\mathbf{X}) + \alpha))}{Z(\theta)^{n+k}}$$

- Prior's normalizing constant (some $\zeta_k(\alpha)$, not $Z(\theta)$) useful for Bayesian inference:
  - e.g. can derive, like before, that $p(\mathbf{X} \mid \alpha) = \zeta_{n+k}(s(x) + \alpha)/\zeta_k(\alpha) \cdot \prod_{i=1}^n h(x^i)$

# Discriminative Models and the Exponential Family

- Going from an exponential family to a discriminative supervised learning:
  - Set canonical parameter to $w^\mathsf{T} x$
  - Gives a convex NLL, where MLE tries to match data/model's conditional statistics
  - Called generalized linear model (GLM) – see Stat 538A, Generalized Linear Models :)

- For example, consider Gaussian with fixed variance for $y$
  - Canonical parameter is $\mu$, and we know setting $\mu = w^\mathsf{T} x$ gives least squares

- If we start with Bernoulli for $y$, we get logistic regression
  - Canonical parmaeter is log-odds
  - Setting $w^\mathsf{T} x = \log(y/(1-y))$ and solving for $y$ gives the sigmoid function
    - Gives a reason (sort of) for using the logistic sigmoid

- You can obtain regression models for other settings using this kind of approach
  - Set canonical parameters to $f_\theta(x)$, the output of a neural network
  - Use a different exponential family to handle a different type of data

# Examples of Exponential Families

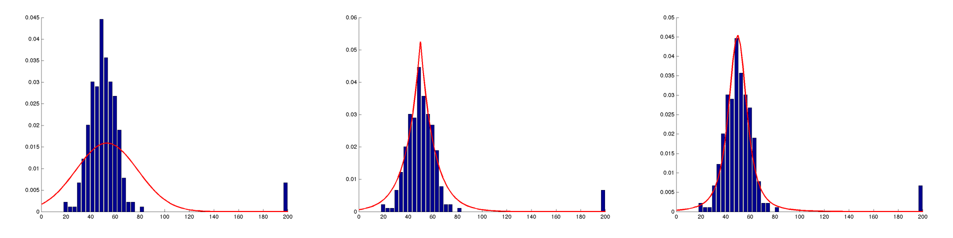
bonus!

- Bernoulli: distribution on $\{0, 1\}$
- Categorical: distribution on $\{1, 2, \ldots, k\}$
- Multivariate Gaussian: distribution on $\mathbb{R}^d$
- Beta: distribution on $[0, 1]$ (including uniform)
- Dirichlet: distribution on discrete probabilities
- Wishart: distribution on positive-definite matrices
- Poisson: distribution on non-negative integers
- Gamma: distribution on positive real numbers
- Many, many others: Wikipedia has a big table
- ...can even have infinite-dimensional statistics via kernel exponential families

# Non-Examples of Exponential Families

bonus!

- Laplace and student $t$ distribution are not exponential families.




  - "Heavy-tailed": have larger probability that data is far from mean
  - More robust to outliers than Gaussian
- Ordinal logistic regression is not in exponential family
  - Can be used for categorical variables where ordering matters
- In these cases, we may not have nice properties:
  - MLE may not be intuitive or closed-form, NLL may not be convex
  - May not have conjugate prior, so need Monte Carlo or variational methods

# Summary

- Exponential families:
  - Have sufficient statistics and canonical parameters
  - Maximimum likelihood becomes moment matching; always have conjugate priors
  - Can build discriminative models by using canonical parameter $s(x) = w^{\mathsf{T}} x$
  - Many things (but not everything!) are exponential families

- Next time: mixing things up

# Convex Conjugate and Entropy

bonus!

- The convex conjugate of a function $A$ is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^\mathsf{T} w - A(w)\}.$$

- E.g., if we consider for logistic regression

$$A(w) = \log(1 + \exp(w)),$$

we have that $A^*(\mu)$ satisfies $w = \log(\mu)/\log(1-\mu)$.

  - When $0 < \mu < 1$ we have

$$\begin{aligned} A^*(\mu) &= \mu \log(\mu) + (1-\mu)\log(1-\mu) \\ &= -H(p_\mu), \end{aligned}$$

    negative entropy of binary distribution with mean $\mu$.
  - If $\mu$ does not satisfy boundary constraint, $\sup$ is $\infty$.

# Convex Conjugate and Entropy

bonus!

- More generally, if $A(w) = \log(Z(w))$ for an exponential family then
$$A^*(\mu) = -H(p_\mu),$$
subject to boundary constraints on $\mu$ and constraint:
$$\mu = \nabla A(w) = \mathbb{E}[s(X)].$$
- Convex set satisfying these is called marginal polytope $\mathcal{M}$.
- If $A$ is convex (and LSC), $A^{**} = A$. So we have
$$A(w) = \sup_{\mu \in \mathcal{U}} \{w^\mathsf{T}\mu - A^*(\mu)\}.$$
and when $A(w) = \log(Z(w))$ we have
$$\log(Z(w)) = \sup_{\mu \in \mathcal{M}} \{w^\mathsf{T}\mu + H(p_\mu)\}.$$
- This can be used to derive variational methods, since we have written computing $\log(Z)$ as a convex optimization problem.

# Maximum Likelihood and Maximum Entropy

bonus!

- The maximum likelihood parameters $w$ in exponential family satisfy:

$$\begin{aligned}
&\min_{w\in\mathbb{R}^d} -w^\mathsf{T}s(D) + \log(Z(w)) \\
&= \min_{w\in\mathbb{R}^d} -w^\mathsf{T}s(D) + \sup_{\mu\in\mathcal{M}}\{w^\mathsf{T}\mu + H(p_\mu)\} && \text{(convex conjugate)} \\
&= \min_{w\in\mathbb{R}^d}\sup_{\mu\in\mathcal{M}}\{-w^\mathsf{T}s(D) + w^\mathsf{T}\mu + H(p_\mu)\} \\
&= \sup_{\mu\in\mathcal{M}}\{\min_{w\in\mathbb{R}^d} -w^\mathsf{T}s(D) + w^\mathsf{T}\mu + H(p_\mu)\} && \text{(convex/concave)}
\end{aligned}$$

which is $-\infty$ unless $s(D) = \mu$ (e.g., maximum likelihood $w$), so we have

$$\begin{aligned}
&\min_{w\in\mathbb{R}^d} -w^\mathsf{T}s(D) + \log(Z(w)) \\
&= \max_{\mu\in\mathcal{M}} H(p_\mu),
\end{aligned}$$

subject to $s(D) = \mu$.

- Maximum likelihood $\Rightarrow$ maximum entropy $+$ moment constraints.