Stat 238, Fall 2025
Project Minis

Alexander Strang
Due: by **5:00 pm** Friday, April 4th, 2025

Mini Project 7: Exponential Family Proofs

Policies

- This is an individual project.
- All submissions must be properly type-set and sourced. This means providing formal citations and a complete bibliography when sources are used.
- Project minis should be uploaded to Gradescope under the appropriate mini-assignment.

Prompts

This is an analysis. You will be asked to prove the two theorems stated at the end of lecture 10.

In Lecture 10 we stated two remarkable facts about exponential families. First, if we observe a set of moments of an unknown distribution, then the exponential family with features set equal to those moments is, in a sense, the least informative distribution with the known moments. This justifies using exponential families to build minimally informed priors. Second, given an exponential family, the choice of parameters that maximize the likelihood for a set of observed sufficient statistics are equivalent to the parameters which equate the expected value of the feature functions and the observed sufficient statistics.

(a) (8 points) *Exponential Families and Maximum Entropy Priors (Discrete Unknowns):*

Suppose that X is a discrete unknown taking on finitely many possible values in $\mathcal{X}$ (that is, $\mathcal{X} = \{x_1, x_2, \dots, x_{|\mathcal{X}|}\}$ contains $|\mathcal{X}| < \infty$ possible states). Then, all p_X may be represented as nonnegative vectors in $\mathbb{R}^{|\mathcal{X}|}$ with entries that add to one.

Next, we need a way to measure how informative a distribution is. Let, $H[X] = H(p_X)$ denote the Shannon entropy:

$$H[X] = H(p_X) = -\mathbb{E}_{X\sim p_X}[\log_d(p_X)]. \tag{1}$$

The Shannon entropy measures the expected number of d-ary questions (e.g. $d = 2$ corresponds to yes/no questions) needed to resolve $X \sim p_X$ under an optimal search procedure. As such, large entropy means that p_X leaves X uncertain. The larger the entropy, the more information is needed to resolve X, so the less informative p_X.

Suppose that p_X is unknown, but some of the moments of p_X are known. In particular, let $\vec{\phi}(\cdot)$ be a vector-valued function of x: $\vec{\phi}(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_m(x)] \in \mathbb{R}^m$. Then, suppose that $\mathbb{E}_{X\sim p_X}[\vec{\phi}(X)] = \vec{\alpha}$.

Theorem 1: Given X taking values in $\mathcal{X}$, with moments $\mathbb{E}_X[\vec{\phi}(X)] = \vec{\alpha}$, the distribution $p_*(\vec{\alpha})$ that maximizes $H(p)$ over all distributions supported on $\mathcal{X}$ with moments $\mathbb{E}_{Y\sim p}[\vec{\phi}(Y)] = \vec{\alpha}$, is an exponential family with feature functions $\vec{\phi}(\cdot)$ and base measure $f(x) = \mathbf{1}(x \in \mathcal{X})$.

1. Write out the theorem as the solution to a constrained optimization problem on $p \in \mathbb{R}^{|\mathcal{X}|}$ with objective $H(p) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x))$. In particular, write out constraint functions that require each entry of p is nonnegative, the entries sum to one, and the moments of $\vec{\phi}(X)$ match $\vec{\alpha}$. You should have $|\mathcal{X}|$ linear inequality constraints and $m+1$ linear equality constraints. Argue that these constraints define a convex polytope.
2. The Shannon entropy is a convex, continuously differentiable function of the input distribution. The optimizer of a convex, continuously differentiable object on a convex domain must satisfy the Karush-Kuhn-Tucker (KKT) conditions. These generalize the Lagrange conditions used given equality constraints. Write out the KKT conditions p_* must satisfy.
3. Compute each of the gradients needed to satisfy the KKT conditions.
4. Rearrange the KKT conditions to prove the theorem statement.

(b) (4 points) *Exponential Families and Maximum Entropy Priors (Continuous Unknowns):*

Attempt to generalize your proof for discrete $\mathcal{X}$ to continuous $\mathcal{X}$ (e.g. $\mathcal{X} \subseteq \mathbb{R}^d$ with a non-empty interior). This can be done in much the same method, but replacing vector-calculus with functional calculus. The KKT conditions still apply, and specify the optimal distribution. Now you will have one non-negativity constraint for each $x \in \mathcal{X}$, and $m+1$ linear equality constraints. To compute functional gradients let $\nabla U(f) = g$ if $\partial_h U(f) = \lim_{\epsilon\to 0}(U(f+\epsilon h) - U(f))/\epsilon = \langle g, h\rangle = \int_{x\in\mathcal{X}} g(x)h(x)dx$. That is, the functional gradient of a functional of U at f is the function $g : \mathcal{X} \to \mathbb{R}$ such that the inner product between g and h returns the partial derivative of U at f along perturbations in the direction h. This extends the usual definition used in finite-dimensional spaces.[1]

(c) (4 points) *Justifying Standard Priors:*

Using this result, create a table showing the support $\mathcal{X}$, and set of known moments, for which the following priors are minimally informative:

1. A beta distribution
2. An exponential distribution
3. A normal distribution
4. A gamma distribution
5. A Student's t distribution

[1] This is a tricky problem to set up. If you are unfamiliar with functional calculus, but want to try it, come talk to us or ask on Ed. The basic proof machinery is identical to the machinery used in finite-dimensions, but requires functional calculus.

(d) (8 points) *Maximizing Exponential Families and Method of Moments:*

Consider an exponential family of the form:

$$p(x;\vec{\phi}) = \frac{f(x)}{Z(\vec{\phi})}\exp(\vec{\phi}\cdot\vec{u}(x)). \tag{2}$$

with known features and base measure but unknown parameters $\vec{\phi}$. Suppose that $Z(\vec{\phi})$ is finite for all $\vec{\phi} \in \Phi$ where Φ is a convex set with a non-empty interior. Suppose, in addition, that the feature functions are linearly independent.

Suppose we observe $X^{(n)} = \{X_1, X_2, \dots X_n\}$ drawn i.i.d. from the exponential family. Let $\hat{\phi}_{\text{MLE}}$. Let:

$$\vec{s}(X^{(n)}) = \frac{1}{n}\sum_{j=1}^{n}\vec{u}(X^{(n)}) \tag{3}$$

be the sufficient statistics. Then:

Theorem 2: The maximum likelihood estimator for $\vec{\phi}$, $\hat{\phi}_{\text{MLE}}(x^{(n)}) = \text{argmax}_{\vec{\phi}\in\Phi}\{\text{Lik}(\vec{\phi}|X^{(n)} = x^{(n)})\}$, is the unique solution to the moment matching problem:

$$\text{Find } \vec{\phi} \text{ such that } \mathbb{E}_{X\sim p(\cdot;\vec{\phi})}[\vec{u}(X)] = \vec{s}(X^{(n)}). \tag{4}$$

Prove Theorem 2 by following the sequence of steps outlined in the lecture notes for lecture 10. You should proceed by:

1. Showing that the log partition function $\log(Z(\vec{\phi})$ has gradient equal to the expected value of the feature functions and Hessian equal to their covariance when X is sampled from $p(\cdot;\vec{\phi})$.
2. Use this conclusion to prove that the log partition function is strictly convex in ϕ given linearly independent feature functions (argue by contradiction).
3. Argue that, any strictly convex, twice-differentiable function admits an invertible mapping between its arguments and its gradient.
4. Show that setting the gradient of the likelihood to zero is equivalent to matching the gradient of the log-partition to the sufficient statistics.