TOPIC 4

Probabilistic modeling, inference and sampling

1 Background: conditioning; parametric families

We review a fundamental probability concept: conditional probability. We also introduce a common family of distributions, the exponential family.

Throughout this topic, all formal proofs are done under the assumption of a discrete distribution with finite support to avoid unnecessary technicalities and focus rather on the concepts. But everything we discuss can be adapted to continuous or hybrid distributions.

1.1 Conditioning

We begin with the probabilistic notion of conditioning, which plays a key role in probabilistic modeling and reasoning.

1.1.1 Conditional probability
We start with events. Throughout, we work on a fixed probability space \((\Omega, \mathcal{F}, \mathbb{P})\), which we assume is finite.

**Definition (Conditional Probability):** Let \(A\) and \(B\) be two events with \(\mathbb{P}[B] > 0\). The conditional probability of \(A\) given \(B\) is defined as

\[
\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.
\]

\(\triangleright\)

The intuitive interpretation goes something like this: knowing that event \(B\) has occurred, the updated probability of observing \(A\) is the probability of its restriction to \(B\) properly normalized to reflect that outcomes outside \(B\) have updated probability 0. This is illustrated next.


Conditional probabilities behave like unconditional probabilities. For instance:

**Exercise:** Let \(A_1, \ldots, A_m\) be disjoint events and let \(B\) be such that \(\mathbb{P}[B] > 0\). Show that

\[
\mathbb{P}[A_1 \cup \cdots \cup A_m | B] = \mathbb{P}[A_1 | B] + \cdots + \mathbb{P}[A_m | B].
\]

\(\triangleright\)

Independence can be characterized in terms of conditional probability. In words, \(A\) and \(B\) are independent if conditioning on one of them does not change the probability of observing the other.

**Lemma (Independence and Conditional Probability):** Let \(A\) and \(B\) be two events of positive probability. Then \(A\) and \(B\) are independent, which we will denote as \(A \perp B\), if and only if \(\mathbb{P}[A | B] = \mathbb{P}[A]\) and \(\mathbb{P}[B | A] = \mathbb{P}[B]\).
Proof: If $A$ and $B$ are independent, then $c$ which implies
\[ \mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]. \]
In the other direction,
\[ \mathbb{P}[A] = \mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \]
implies $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$ after rearranging. \(\square\)

The conditional probability is often used in three fundamental ways, which we recall next. Proofs can be found in most probability textbooks.

- **Product (or chain) rule:** For any collection of events $A_1, \ldots, A_r$,
  \[ \mathbb{P} \left[ \cap_{i=1}^{r} A_i \right] = \prod_{i=1}^{r} \mathbb{P} \left[ A_i \mid \cap_{j=1}^{i-1} A_j \right]. \]

- **Law of total probability:** For any event $B$ and any partition (A.1, \ldots, A_r) $A_1, \ldots, A_r$ of $\Omega$,
  \[ \mathbb{P}[B] = \sum_{i=1}^{r} \mathbb{P}[B|A_i]\mathbb{P}[A_i]. \]

- **Bayes' rule:** For any events $A$ and $B$ with positive probability,
  \[ \mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \]

### 1.1.2 Conditioning on a random variable

Conditional probabilities extend naturally to random variables. If $X$ is a discrete random variable, we let $p_X$ be its probability mass function and $S_X$ be its support, that is, the set of values where it has positive probability. Then we can for instance condition on the event \{ $X = x$ \} for any $x \in S_X$. 
We define the conditional probability mass function.

**Definition (Conditional Probability Mass Function):** Let $X$ and $Y$ be discrete random variables with joint probability mass function $p_{X,Y}$ and marginals $p_X$ and $p_Y$. The conditional probability mass function of $X$ given $Y$ is defined as

$$p_{X|Y}(x|y) := P[X = x|Y = y] = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

which is defined for all $x \in S_X$ and $y \in S_Y$. \(\triangleleft\)

The conditional expectation can then be defined in a natural way as the expectation over the conditional probability mass function.

**Definition (Conditional Expectation):** Let $X$ and $Y$ be discrete random variables where $X$ takes real values. The conditional expectation of $X$ given $Y = y$ is given by

$$\mathbb{E}[X|Y = y] = \sum_{x \in S_X} x \cdot p_{X|Y}(x|y).$$

\(\triangleleft\)

Thinking of $\mathbb{E}[X|Y = y]$ as a function of $y$ leads to a fundamental characterization of the conditional expectation.

**Theorem (Conditional Expectation and Least Squares):** Let $X$ and $Y$ be discrete random variables where $X$ takes real values. Then the conditional expectation $h(y) = \mathbb{E}[X|Y = y]$ minimizes the least squares criterion

$$\min_h \mathbb{E} \left[ (X - h(Y))^2 \right]$$

where the minimum is over all real-valued functions of $y$.

**Proof:** Think of $h(y)$ as a vector $h = (h_y)_{y \in S_Y}$ indexed by $S_Y$, where $h(y) = h_y \in \mathbb{R}$. Then

$$\mathcal{L}(h) = \mathbb{E} \left[ (X - h(Y))^2 \right] = \sum_{x \in S_X} \sum_{y \in S_Y} (x - h_y)^2 p_{X,Y}(x, y).$$
Take partial derivatives with respect to $h$:

$$
\frac{\partial}{\partial h_y} \mathcal{L} = -2 \sum_{x \in S_X} (x - h_y)p_{X,Y}(x, y) = -2 \sum_{x \in S_X} xp_{X,Y}(x, y) + 2h_y p_Y(y)
$$

where we used that the marginal of $Y$ satisfies $p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x, y)$. The Hessian is the diagonal matrix with positive entries $2p_Y(y)$, so it is positive definite and the problem is convex.

After rearranging, the stationary points satisfy

$$
h_y = \sum_{x \in S_X} x \frac{p_{X,Y}(x, y)}{p_Y(y)} = \sum_{x \in S_X} xp_{X|Y}(x|y) = \mathbb{E}[X|Y = y]
$$

as claimed. □

### 1.2 Parametric families

Parametric families of probability distributions serve as basic building blocks for more complex models. By this, we mean a collection $\{p_\theta : \theta \in \Theta\}$, where $p_\theta$ is a probability distribution over a set $S_\theta$ and $\theta$ is a vector-valued parameter.

**Example (Bernoulli):** The random variable $X$ is Bernoulli with parameter $q \in [0, 1]$, denoted $X \sim \text{Ber}(q)$, if it takes values in $S_X = \{0, 1\}$ and $\mathbb{P}[X = 1] = q$. Varying $p$ produces the family of Bernoulli distributions. □

Here we focus on the exponential family, which includes many common distributions (including the Bernoulli distribution).

#### 1.2.1 Exponential family
One particularly useful class of probability distributions in data science is the exponential family (https://en.wikipedia.org/wiki/Exponential_family#Vector_parameter), which includes many well-known cases.

**Definition (Exponential Family - Discrete Case):** A parametric collection of probability distributions \( \{p_\theta : \theta \in \Theta \} \) over a discrete space \( S \) is an exponential family if it can be written in the form
\[
p_\theta(x) = \frac{1}{Z(\theta)} h(x) \exp(\theta^T \phi(x))
\]
where \( \theta \in \mathbb{R}^m \) are the canonical parameters, \( \phi : S \to \mathbb{R}^m \) are the sufficient statistics and \( Z(\theta) \) is the partition function. It is often convenient to introduce the function \( A(\theta) = \log Z(\theta) \) and re-write
\[
p_\theta(x) = h(x) \exp(\theta^T \phi(x) - A(\theta)).
\]

\[\triangle\]

**Example (Bernoulli, continued):** For \( x \in \{0, 1\} \), the Bernoulli \( \text{Ber}(q) \) distribution for \( 0 < q < 1 \) can be written as
\[
q^x(1 - q)^{1-x} = (1 - q) \left( \frac{q}{1 - q} \right)^x \\
= (1 - q) \exp\left[ x \log \left( \frac{q}{1 - q} \right) \right] \\
= \frac{1}{Z(\theta)} h(x) \exp(\theta \phi(x))
\]
where we define \( h(x) = 1, \phi(x) = x, \theta = \log \left( \frac{q}{1 - q} \right) \) and, since \( Z(\theta) \) serves as the normalization constant in \( p_\theta \),
\[
Z(\theta) = \sum_{x \in \{0, 1\}} h(x) \exp(\theta \phi(x)) = 1 + e^\theta.
\]

\[\triangle\]

The following is an important generalization.

**Example (Categorical and Multinomial):** A categorical variable \( X \) takes \( K \geq 2 \) possible values. A standard choice is to use one-hot encoding \( S = \{ e_i : i = 1, \ldots, K \} \) where \( e_i \) is the \( i \)-th canonical basis in \( \mathbb{R}^K \). The distribution is specified by setting the probabilities \( \pi = (\pi_1, \ldots, \pi_K) \)
\[
\pi_i = \Pr[X = e_i].
\]
We denote this \( X \sim \text{Cat}(\pi) \) and we assume \( \pi_i > 0 \) for all \( i \).
To see that this is an exponential family, write the probability mass function at \( \mathbf{x} = (x_1, \ldots, x_K) \) as

\[
\prod_{i=1}^{K} \pi_i^{x_i} = \exp \left( \sum_{i=1}^{K} x_i \log \pi_i \right).
\]

So we can take \( h(\mathbf{x}) = 1, \theta = (\log \pi_i)_{i=1}^{K}, \phi(\mathbf{x}) = (x_i)_{i=1}^{K} \) and \( Z(\theta) = 1 \).

The multinomial distribution ([https://en.wikipedia.org/wiki/Multinomial_distribution](https://en.wikipedia.org/wiki/Multinomial_distribution)) arises as a sum of independent categorical variables. Let \( n \geq 1 \) be the number of trials (or samples) and let \( Y_1, \ldots, Y_n \) be IID \( \text{Cat}(\pi) \). Define \( X = \sum_{i=1}^{n} Y_i \). The probability mass function of \( X \) at

\[
\mathbf{x} = (x_1, \ldots, x_K) \in \left\{ \mathbf{x} \in \{0, 1, \ldots, n\}^K : \sum_{i=1}^{K} x_i = n \right\} =: S
\]

is

\[
\frac{n!}{x_1! \cdots x_K!} \prod_{i=1}^{K} \pi_i^{x_i} = \frac{n!}{x_1! \cdots x_K!} \exp \left( \sum_{i=1}^{K} x_i \log \pi_i \right)
\]

and we can take \( h(\mathbf{x}) = \frac{n!}{x_1! \cdots x_K!}, \theta = (\log \pi_i)_{i=1}^{K}, \phi(\mathbf{x}) = (x_i)_{i=1}^{K} \) and \( Z(\theta) = 1 \). This is an exponential family if we think of \( n \) as fixed.

We use the notation \( X \sim \text{Mult}(n, \pi) \).

**Exercise (Moments of Multinomial):** Let \( X \sim \text{Mult}(n, \pi) \) with \( n \geq 1 \) and \( \pi \in \Delta_K \). Establish the following formulas:

(a) \( \mathbb{E}[X] = n\pi \)

(b) \( \text{Var}[X] = n[\text{Diag}(\pi) - \pi \pi^T] \)

While we have focused so far on discrete distributions, one can adapt the definitions above by replacing mass functions with density functions. We give two important examples.

**Exercise:** Recall that the [trace (https://en.wikipedia.org/wiki/Trace_(linear_algebra))] of a square matrix \( A \), denoted \( \text{tr}(A) \), is the sum of its diagonal entries.

(a) Show that, for any \( A \in \mathbb{R}^{n \times m} \) and \( B \in \mathbb{R}^{m \times n} \), it holds that \( \text{tr}(AB) = \text{tr}(BA) \).

(b) Use (a) to show that more generally \( \text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA) \).
**Example (Multivariate Gaussian):** A multivariate Gaussian vector \( \mathbf{X} \) on \( \mathbb{R}^d \) with mean \( \mathbf{\mu} \in \mathbb{R}^d \) and positive definite covariance matrix \( \mathbf{\Sigma} \in \mathbb{R}^{d \times d} \) has probability density function

\[
f_{\mathbf{\mu},\mathbf{\Sigma}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{\mu})\right)
\]

where \(|A|\) is the [determinant](https://en.wikipedia.org/wiki/Determinant) of \( A \). We use the notation \( \mathbf{X} \sim \mathcal{N}(\mathbf{\mu}, \mathbf{\Sigma}) \).

![Multivariate Normal Distribution](https://commons.wikimedia.org/wiki/File:Multivariate_Gaussian.png)

Rewriting the density as

\[
f_{\mathbf{\mu},\mathbf{\Sigma}}(\mathbf{x}) = \frac{e^{- (1/2) \mathbf{\mu}^T \mathbf{\Sigma}^{-1} \mathbf{\mu}}}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(- \mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{\mu} - \frac{1}{2} \text{tr} \left( \mathbf{x} \mathbf{x}^T \mathbf{\Sigma}^{-1} \right) \right)
\]

where we used the symmetric of \( \mathbf{\Sigma}^{-1} \) in the first term of the exponential and trace identity of the previous exercise in the second term. The expression in parentheses is linear in the entries of \( \mathbf{x} \) and \( \mathbf{x} \mathbf{x}^T \), which can then be taken as sufficient statistics (formally, using [vectorization](https://en.wikipedia.org/wiki/Vectorization_(mathematics))). This shows that the multivariate normal is an exponential family. \( \triangleleft \)

The [Dirichlet distribution](https://en.wikipedia.org/wiki/Dirichlet_distribution), which we describe next, is a natural probability distribution over probability distributions. In particular, it is used in [Bayesian data analysis](https://en.wikipedia.org/wiki/Bayesian_statistics) as a prior on the parameters of categorical and multinomial distribution, largely because of a property known as [conjugacy](https://en.wikipedia.org/wiki/Conjugate_prior).
Example (Dirichlet): The Dirichlet distribution is a distribution over the $(K - 1)$-simplex

\[ S = \Delta_K := \left\{ \mathbf{x} = (x_1, \ldots, x_K) : x \geq 0, \sum_{i=1}^{K} x_i = 1 \right\}. \]

Its parameters are $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}$ and the density is

\[ f_\alpha (\mathbf{x}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad \mathbf{x} \in \Delta_K \]

where the normalizing constant $B(\alpha)$ is the multivariate Beta function (https://en.wikipedia.org/wiki/Beta_function#Multivariate_beta_function).

Rewriting the density as

\[ \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i \exp \left( \sum_{i=1}^{K} \alpha_i \log x_i \right) \]

shows that this is an exponential family with the canonical parameters $\alpha$ and sufficient statistics $(\log x_i)_{i=1}^{K}$.

See here (https://en.wikipedia.org/wiki/Exponential_family#Table_of_distributions) for many more examples. Observe, in particular, that the same distribution can have several representations as an exponential family.

NUMERICAL CORNER In Julia, the package Distributions.jl (https://juliastats.org/Distributions.jl/stable/) provides a way to sample from a variety of standard distributions. We first initialize the pseudorandom number generator (https://en.wikipedia.org/wiki/Pseudorandom_number_generator) with a random seed (https://en.wikipedia.org/wiki/Random_seed). In particular it allows the results to be reproducible: using the same seed produces the same results again.

```julia
In [1]: # Julia version: 1.5.1
using Random, Distributions

In [2]: Random.seed!(123); # Setting the seed

In [3]: p = 0.1
d = Bernoulli(p)

Out[3]: Bernoulli{Float64}(p=0.1)
```
We then use `rand` to generate samples.

```
In [4]: N = 5  # number of samples
    : rand(d,N)
```

```
Out[4]: 5-element Array{Bool,1}:
         0
         0
         0
         0
         0
```

Here are a few other examples.

```
In [5]: p = [0.1, 0.2, 0.7]
    : n = 100
    : d = Multinomial(n,p)
    : rand(d, N)
```

```
Out[5]: 3×5 Array{Int64,2}:
        15  4  10  13   4
       15  27  19  31  18
      70  69  71  56  78
```

```
In [6]: mu = [0.1, -0.3]
    : sig = [2. 0.; 0. 3.]
    : d = MvNormal(mu, sig)
    : rand(d, N)
```

```
Out[6]: 2×5 Array{Float64,2}:
         3.78819  0.549552  -1.85312   0.672791  1.9125
        2.16195  1.92556   2.63177  -1.92266   0.959448
```

### 1.2.2 Parameter estimation

When modeling data via a parametric family of distributions, the parameters must be determined from the data itself. In a typical setting, we assume that the data comprises $n$ independent samples $X_1, \ldots, X_n$ from a parametric family $p_\theta$ with unknown $\theta \in \Theta$. Many methods exist for estimating $\theta$, depending on the context. Here we focus on perhaps the best-known approach, maximum likelihood estimation ([https://en.wikipedia.org/wiki/Maximum_likelihood_estimation](https://en.wikipedia.org/wiki/Maximum_likelihood_estimation)). It has many good theoretical properties ([https://en.wikipedia.org/wiki/Maximum_likelihood_estimation#Properties](https://en.wikipedia.org/wiki/Maximum_likelihood_estimation#Properties)) which we will not describe here, as well as drawbacks ([https://stats.stackexchange.com/questions/261056/why-does-maximum-likelihood-estimation-have-issues-with-over-fitting](https://stats.stackexchange.com/questions/261056/why-does-maximum-likelihood-estimation-have-issues-with-over-fitting)).
The idea behind maximum likelihood estimation is simple and intuitive: we choose the parameter that maximizes the probability of observing the data.

**Definition (Maximum likelihood estimator):** Assume that \( X_1, \ldots, X_n \) are \( n \) independent samples from a parametric family \( p_{\theta^*} \) with unknown \( \theta^* \in \Theta \). The maximum likelihood estimator of \( \theta \) is defined as

\[
\hat{\theta}_{\text{MLE}} \in \arg \max \left\{ \prod_{i=1}^{n} p_{\theta}(X_i) : \theta \in \Theta \right\}.
\]

It is often advantageous to work with the minus log-likelihood

\[
L_n(\theta; \{X_i\}_{i=1}^{n}) = - \sum_{i=1}^{n} \log p_{\theta}(X_i).
\]

\( \triangleleft \)

For exponential families, maximum likelihood estimation takes a particularly natural form. We provide details in the discrete case.

**Theorem (Maximum Likelihood Estimator for Exponential Families):** Assume that \( p_{\theta} \) takes the exponential family form

\[
p_{\theta}(x) = h(x) \exp \left( \theta^T \phi(x) - A(\theta) \right)
\]

and that the support \( S \) is finite. Let \( X_1, \ldots, X_n \) be \( n \) independent samples from a parametric family \( p_{\theta^*} \) with unknown \( \theta^* \in \Theta \). Then \( L_n(\theta; \{X_i\}_{i=1}^{n}) \), as a function of \( \theta \), is convex and the maximum likelihood estimator of \( \theta \) - if it exists - solves the system of moment-matching equations

\[
\mathbb{E}[\phi(X)] = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i),
\]

where \( X \sim p_{\hat{\theta}_{\text{MLE}}} \).

We will need some auxiliary formulas and results.

**Definition (Covariance matrix):** Let \( Z \) be a random vector. The covariance matrix of \( Z \) is defined as

\[
K_{Z,Z} = \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T].
\]

\( \triangleleft \)

**Exercise:** Show that

\[
K_{Z,Z} = \mathbb{E}[ZZ^T] - \mathbb{E}[Z]\mathbb{E}[Z]^T
\]

\( \triangleleft \)
**Lemma:** Let $\mathbf{Z}$ be a random vector taking values in $\mathbb{R}^m$ whose components have finite variances. Then the covariance matrix of $\mathbf{Z}$, $\mathbf{K}_{\mathbf{Z},\mathbf{Z}}$, is positive semidefinite.

**Proof idea:** We show that the quadratic form $\langle \mathbf{y}, \mathbf{K}_{\mathbf{Z},\mathbf{Z}} \mathbf{y} \rangle$ can be written as a variance, and so is always non-negative.

**Proof:** For any $\mathbf{y} \in \mathbb{R}^m$, by the exercise above,

\[
\langle \mathbf{y}, \mathbf{K}_{\mathbf{Z},\mathbf{Z}} \mathbf{y} \rangle = \mathbf{y}^T \left( \mathbb{E}[\mathbf{ZZ}^T] - \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{Z}]^T \right) \mathbf{y} \\
= \mathbb{E}[\mathbf{y}^T \mathbf{Z}^T \mathbf{y}] - \mathbb{E}[\mathbf{y}^T \mathbf{Z}]\mathbb{E}[\mathbf{Z}^T \mathbf{y}] \\
= \mathbb{E}[(\mathbf{y}^T \mathbf{Z})^2] - (\mathbb{E}[\mathbf{y}^T \mathbf{Z}])^2 \\
= \text{Var}[\mathbf{y}^T \mathbf{Z}] \\
\geq 0.
\]

Since this inequality holds for any $\mathbf{y}$, it follows by definition that $\mathbf{K}_{\mathbf{Z},\mathbf{Z}} \succeq 0$. □

The function $A$ has properties worth highlighting.

**Lemma (Derivatives of $A$):** Assume that $p_\theta$ takes the exponential family form

\[ p_\theta(\mathbf{x}) = h(\mathbf{x}) \exp(\theta^T \phi(\mathbf{x}) - A(\theta)) \]

and that the support $\mathcal{S}$ is finite. Then

\[ \nabla A(\theta) = \mathbb{E}[\phi(\mathbf{X})] \quad \text{and} \quad \mathbf{H}_A(\theta) = \mathbf{K}_{\phi(\mathbf{X}),\phi(\mathbf{X})}, \]

where $\mathbf{X} \sim p_\theta$.

**Proof idea:** Follows from a direct calculation.

**Proof:** We observe first that

\[ A(\theta) = \log Z(\theta) = \log \left( \sum_{\mathbf{x} \in \mathcal{S}} h(\mathbf{x}) \exp(\theta^T \phi(\mathbf{x})) \right), \]

where we used the fact that, by definition, $Z(\theta)$ is the normalization constant of $p_\theta$. In particular, as the logarithm of a finite, weighted sum of exponentials, the function $A(\theta)$ is continuously differentiable. Hence so is $p_\theta(\mathbf{x})$ as a function of $\theta$. 

From the formula above and the basic rules of calculus,

\[
\frac{\partial}{\partial \theta_j} A(\theta) = \frac{\partial}{\partial \theta_j} \log \left( \sum_{x \in S} h(x) \exp(\theta^T \phi(x)) \right) = \frac{\sum_{x \in S} h(x) \phi_j(x) \exp(\theta^T \phi(x))}{\sum_{x \in S} h(x) \exp(\theta^T \phi(x))} = \sum_{x \in S} \phi_j(x) \frac{1}{Z(\theta)} h(x) \exp(\theta^T \phi(x)) = \sum_{x \in S} \phi_j(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) = \mathbb{E} [\phi_j(X)],
\]

where \( X \sim p_\theta \).

Differentiating again, this time with respect to \( \theta_i \), we obtain

\[
\frac{\partial^2}{\partial \theta_i \partial \theta_j} A(\theta) = \frac{\partial}{\partial \theta_i} \left\{ \sum_{x \in S} \phi_j(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) \right\} = \sum_{x \in S} \phi_j(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) \left\{ \phi_i(x) - \frac{\partial}{\partial \theta_i} A(\theta) \right\} = \sum_{x \in S} \phi_i(x) \phi_j(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) - \left( \sum_{x \in S} \phi_i(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) \right) \times \left( \sum_{x \in S} \phi_j(x) h(x) \exp(\theta^T \phi(x) - A(\theta)) \right) = \mathbb{E} [\phi_i(X) \phi_j(X)] - \mathbb{E} [\phi_i(X)] \mathbb{E} [\phi_j(X)],
\]

where again \( X \sim p_\theta \). That concludes the proof. \( \square \)

We are now ready to prove the main theorem.
Proof (Maximum Likelihood Estimator for Exponential Families): We begin by computing the stationary points of the log-likelihood, for which we need the gradient with respect to \( \theta \in \mathbb{R}^m \). We will also need the second-order derivatives to establish convexity. We have

\[
\frac{\partial}{\partial \theta_j} \{- \log p(x)\} = \frac{\partial}{\partial \theta_j} \{- \log h(x) - \theta^T \phi(x) + A(\theta)\} \\
= -\phi_j(x) + \frac{\partial}{\partial \theta_j} A(\theta).
\]

and

\[
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \{- \log p(x)\} = \frac{\partial}{\partial \theta_i} \left\{-\phi_j(x) + \frac{\partial}{\partial \theta_j} A(\theta)\right\} \\
= \frac{\partial^2}{\partial \theta_i \partial \theta_j} A(\theta).
\]

We use the expressions for the derivatives of \( A \) obtained above.

Plugging into the formula for the minus log-likelihood (as a function of \( \theta \)), we get for the gradient with respect to \( \theta \)

\[
\nabla L_n(\theta; \{X_i\}_{i=1}^n) = \sum_{i=1}^n \{-\phi(X_i) + \nabla A(\theta)\} \\
= \sum_{i=1}^n \{-\phi(X_i) + \mathbb{E}[\phi(X)]\}.
\]

For the Hessian with respect to \( \theta \), we get

\[
H_{L_n}(\theta; \{X_i\}_{i=1}^n) = \sum_{i=1}^n H_A(\theta) = n H_{\phi(X), \phi(X)}.
\]

Since \( H_{\phi(X), \phi(X)} \) is positive semidefinite, so is \( H_{L_n}(\theta; \{X_i\}_{i=1}^n) \).
Hence, a stationary point \( \hat{\theta}_{\text{MLE}} \) must satisfy
\[
0 = \nabla L_n(\theta; \{X_i\}_{i=1}^n) = \sum_{i=1}^{n} \{-\phi(X_i) + \mathbb{E}[\phi(X)]\},
\]
where \( X \sim p_{\hat{\theta}_{\text{MLE}}} \) or, after re-arranging,
\[
\mathbb{E}[\phi(X)] = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i).
\]
Because \( L_n \) is convex, a stationary point - if it exists - is necessarily a global minimum by the First-Order Sufficient Condition for Convex Functions. \( \square \)

Example (Bernoulli, continued): For \( x \in \{0, 1\} \), recall that the \( \text{Ber}(q) \) distribution can be written as
\[
p_{\theta}(x) = \frac{1}{Z(\theta)} h(x) \exp(\theta \phi(x))
\]
where we define \( h(x) = 1, \phi(x) = x, \theta = \log \left( \frac{q}{1-q} \right) \) and \( Z(\theta) = \sigma(\theta) \). Let \( X_1, \ldots, X_n \) be independent samples from \( p_{\theta^*} \).

For \( X \sim p_{\hat{\theta}_{\text{MLE}}} \), the moment-matching equations reduce to
\[
\hat{q}_{\text{MLE}} := \mathbb{E}[X] = \mathbb{E}[\phi(X)] = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i) = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]
To compute the left-hand side in terms of \( \hat{\theta}_{\text{MLE}} \) we use the relationship \( \theta = \log \left( \frac{q}{1-q} \right) \), that is,
\[
\hat{\theta}_{\text{MLE}} = \log \left( \frac{\frac{1}{n} \sum_{i=1}^{n} X_i}{1 - \frac{1}{n} \sum_{i=1}^{n} X_i} \right).
\]
Hence, \( \hat{\theta}_{\text{MLE}} \) is well-defined when \( \frac{1}{n} \sum_{i=1}^{n} X_i \neq 0, 1 \).
Define $q^*$ as the solution to

$$\theta^* = \log\left(\frac{q^*}{1 - q^*}\right)$$

that is,

$$q^* = \frac{e^{\theta^*}}{1 + e^{\theta^*}} = \frac{1}{1 + e^{-\theta^*}} = \sigma(\theta^*),$$

where $\sigma$ is the sigmoid function.

By the law of large numbers, as $n \to +\infty$, we get the convergence

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to q^*,\]$$

with probability one.

Because the function $\log\left(\frac{q}{1-q}\right)$ is continuous for $q \in (0, 1)$, we have furthermore

$$\hat{\theta}_{\text{MLE}} = \log\left(\frac{\frac{1}{n} \sum_{i=1}^{n} X_i}{1 - \frac{1}{n} \sum_{i=1}^{n} X_i}\right) \to \log\left(\frac{q^*}{1 - q^*}\right) = \theta^*.\]$$

In words, the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ is guaranteed to converge to the true parameter $\theta^*$ when the number of samples grows. This fundamental property is known as statistical consistency (https://en.wikipedia.org/wiki/Consistent_estimator).

Statistical consistency holds more generally for the maximum likelihood estimator under exponential families (and parametric families more broadly), provided certain technical conditions hold. We will not provide further details here.