$\LaTeX$ command declarations here.

EECS 545: Machine Learning

Lecture 12: Information Theory and Exponential Families

- Instructor: **Jacob Abernethy**
- Date: March 7, 2016

Lecture Exposition Credit: Benjamin Bray & Saket Dewangan

Outline

- Information Theory
  - Information, Entropy, Maximum Entropy Distributions
  - Entropy and Encoding, Cross Entropy, Relative Entropy
  - Mutual Information & Collocations
- Exponential Family
  - Sufficient Statistic
  - General Form of Exponential Family
  - Likelihood and MLE

Reading List

- Required:
  - **[PRML]**, §1.6: Information Theory
  - **[PRML]**, §2.4: The Exponential Family
- Optional:
  - **[MLAPP]**, §2.8: Information Theory
  - **[MLAPP]**, §9.2: Exponential Families

Other References

- Information Theory:
  - **[Shannon 1951]** Shannon, Claude E.. *The Mathematical Theory of Communication* (http://worrydream.com/refs/Shannon%20-%20A%20Mathematical%20Theory%20of%20Communication.pdf). 1951.
  - **[Pierce 1980]** Pierce, John R.. *An Introduction to Information Theory: Symbols, Signals, and Noise* (http://www.amazon.com/An-Introduction-Information-Theory-Mathematics/dp/0486240614). 1980.
  - **[Stone 2015]** Stone, James V.. *Information Theory: A Tutorial Introduction* (http://jim-stone.staff.shef.ac.uk/BookInfoTheory/InfoTheoryBookMain.html). 2015.
- Exponential Families:
  - **[MLAPP]** Murphy, Kevin. *Machine Learning: A Probabilistic Perspective* (https://mitpress.mit.edu/books/machine-learning-0). 2012.
  - **[Hero 2008]** Hero, Alfred O.. *Statistical Methods for Signal Processing* (http://web.eecs.umich.edu/~hero/Preprints/main_564_08_new.pdf). 2008.
  - **[Blei 2011]** Blei, David. *Notes on Exponential Families* (https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/exponential-families.pdf). 2011.
  - **[Wainwright & Jordan 2008]** Wainwright, Martin J. and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference* (https://www.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf). 2008.

> This lecture, we will not cover any classifier or regressor. Instead, some basics of information theory and exponential family will be introduced. These will provide some important background for **Probabilistic Graphical Models**, which is a big topic that we will cover for several following lectures. For information theory, some definitions like information, entropy, cross entropy, relative entropy, etc. are to be introduced. We could see how entropy is related to compression theory. As for applications, we will show how information theory can help us select features and find most frequent collocations in a novel. As for exponential family, we study it because it has some nice properties and will be frequently used in following lectures. Starting with definition of sufficient statistics, we will go through the general form, likelihood function and maximum likelihood estimator of exponential family

Information Theory

> Uses material from **[MLAPP]** §2.8, **[Pierce 1980]**, **[Stone 2015]**, and **[Shannon 1951]**.

Information Theory

- Information theory is concerned with
  - **Compression:** Representing data in a compact fashion
  - **Error Correction:** Transmitting and storing data in a way that is robust to errors
- In machine learning, information-theoretic quantities are useful for
  - manipulating probability distributions
  - interpreting statistical learning algorithms

What is Information?

- Can we measure the amount of **information** we gain from an observation?
  - Information is measured in *bits* (don't confuse with *binary digits*, $0110001\dots$)
  - Intuitively, observing a fair coin flip should give 1 bit of information
  - Observing two fair coins should give 2 bits, and so on...

Information: Definition

- The **information content** of an event E with probability p defined as

$$I(E) = I(p) = -\log_2 p = \log_2 \frac{1}{p} \geq 0$$

  - Information theory is about *probabilities* and *distributions*
  - The "meaning" of events doesn't matter.
  - Using bases other than 2 yields different units (Hartleys, nats, ...)

Information Example: Fair Coin—$P(\text{Head}) = 0.5$

- **One Coin:** If we observe one head, then

$$I(\text{Head}) = -\log_2 P(\text{Head}) = 1 \text{ bit}$$

- **Two Coins:** If we observe two heads in a row,

$$\begin{aligned} I(\text{Head}, \text{Head}) &= -\log_2 P(\text{Head}, \text{Head}) \\ &= -\log_2 P(\text{Head})P(\text{Head}) \\ &= -\log_2 P(\text{Head}) - \log_2 P(\text{Head}) = 2 \text{ bits} \end{aligned}$$

Information Example: Unfair Coin

- Suppose the coin has two heads, so $P(\text{Head}) = 1$. Then,

$$I(\text{Head}) = -\log_2 1 = 0$$

  - We will gain no information!
- On the contrary, if we observe tail

$$I(\text{Tail}) = -\log_2 0 = +\infty$$

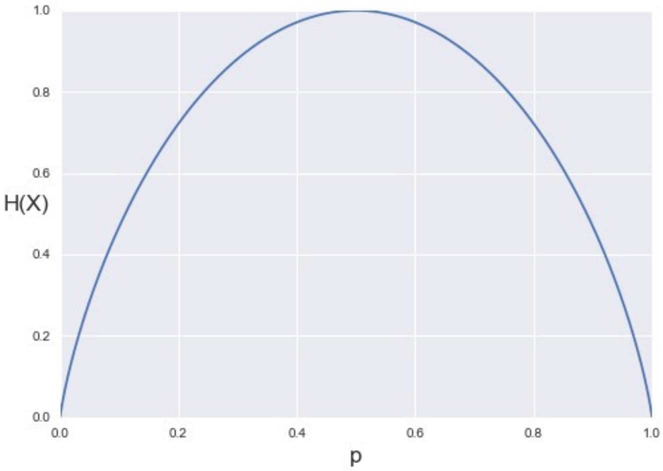
  - We will gain *infinite* information because we observe an impossible thing!
- Information is a measure of how **surprised** we are by an outcome.

Entropy: Definition

- The **entropy** of a discrete random variable X with distribution p is
$$H[X] = E[I(p(X))] = -\sum_{x \in X} p(x) \log p(x)$$
    - Entropy is the expected information received when we sample from X.
    - Entropy measures how *surprised* we are on average
    - When X is continuous random variable, summation is replaced with integral

Entropy: Coin Flip

- If X is binary, entropy is
$$H[X] = -p \log p + (1-p) \log(1-p)$$



- Entropy is highest when X is close to uniform.
    - Large entropy $\iff$ high uncertainty, more information from each new observation
    - Small entropy $\iff$ more knowledge about possible outcomes
- The farther from uniform X is, the smaller the entropy.

Maximum Entropy Principle

- Suppose we sample data from an unknown distribution p, and
    - we collect statistics (mean, variance, etc.) from the data
    - we want an *objective* or unbiased estimate of p The **Maximum Entropy Principle** states that:

> We should choose p to have maximum entropy $H[p]$ among all distributions satisfying our constraints.

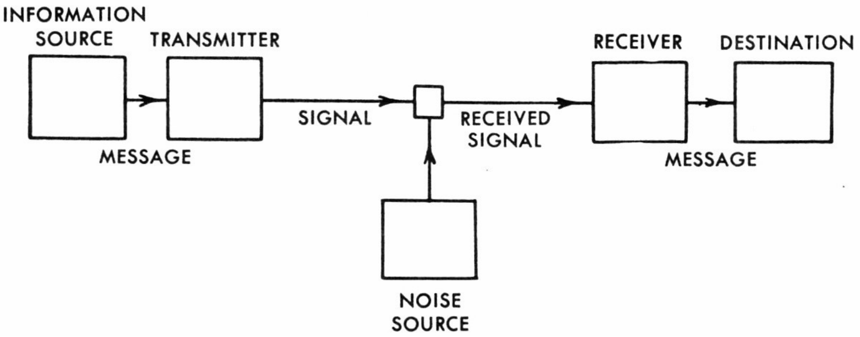
- Some examples of maximum entropy distributions:

Constraints	**Maximum Entropy Distribution**
Min a, Max b	Uniform $U[a,b]$
Mean μ, Support $(0, +\infty)$	Exponential $Exp(\mu)$
Mean μ, Variance σ^2	Gaussian $\mathcal{N}(\mu, \sigma^2)$

- Later, **Exponential Family Distributions** will generalize this concept.

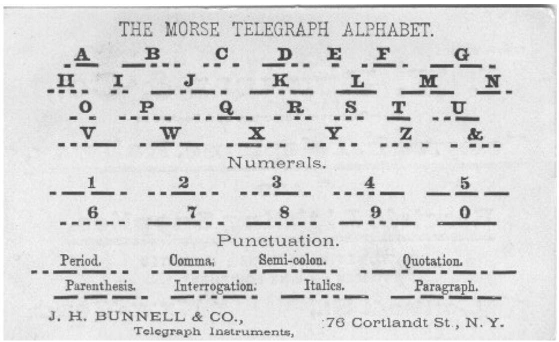
Entropy and Encoding: Communication Channel

- Now let's see how entropy is related to encoding theory
- **Communication channel** can be characterized as:
  - **[Source]** generates messages.
  - **[Encoder]** converts the message to a **signal** for transmission.
  - **[Channel]** is the path along which signals are transmitted, possibly under the influence of **noise**.
  - **[Decoder[** attempts to reconstruct the original message from the transmitted signal.
  - **[Destination]** is the intended recipient.



Entropy and Encoding: Encoding

- Suppose we draw messages from a distribution p.
  - Certain messages may be more likely than others.
  - For example, the letter **e** is most frequent in English
- An **efficient** encoding minimizes the average code length,
  - assign *short* codewords to common messages
  - and *longer* codewords to rare messages
- Example: **Morse Code**



Entropy and Encoding: Source Coding Theorem

- Claude Shannon proved that for discrete noiseless channels:

> It is impossible to encode messages drawn from a distribution p with fewer than $H[p]$ bits, on average.

- Here, *bits* refers to *binary digits*, i.e. encoding messages in binary.

> $H[p]$ measures the optimal code length, in bits, for messages drawn from p

Cross Entropy & Relative Entropy

- Consider different distributions p and q
  - What if we use a code optimal for q to encode messages from p?
- For example, suppose our encoding scheme is optimal for German text.
  - What if we send English messages instead?
  - Certainly, there will be some waste due to different letter frequencies, umlauts, ...

Cross Entropy & Relative Entropy

- **Cross entropy** measures the average number of bits needed to encode messages drawn from p when we use a code optimal for q:
$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) = E_p[\log q(x)]$$
- Intuitively, $H(p,q) \geq H(p)$.
- **Relative entropy** is the difference $H(p,q) - H(p)$.
- Relative entropy, aka **Kullback-Leibler divergence**, of q from p is
$$\begin{aligned} D_{KL}(p \| q) &= H(p,q) - H(p) \\ &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

> Measures the number of *extra* bits needed to encode messages from p if we use a code optimal for q.

Mutual Information: Definition

- **Mutual information** between discrete variables X and Y is
$$\begin{aligned} I(X;Y) &= \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D_{KL}(p(x,y) \| p(x)p(y)) \end{aligned}$$
  - If X and Y are independent, $p(x,y) = p(x)p(y)$ and $I(X;Y) = 0$
  - So, $I(X;Y)$ measures how *dependent* X and Y are!
  - Related to correlation $\rho(X,Y)$

Mutual Information: Example of Feature Selection

- Mutual information can also be used for **feature selection**.
  - In classification, features that *depend* most on the class label C are useful
  - So, choose features X_k such that $I(X_k;C)$ is large
  - This helps to avoid *overfitting* by ignoring irrelevant features!

> See **[MLAPP]** §3.5.4 for more information

Pointwise Mutual Information

- A **collocation** is a sequence of words that co-occur more often than expected by chance.
  - fixed expression familiar to native speakers (hard to translate)
  - meaning of the whole is more than the sum of its parts
  - See these slides (https://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/philip-pmi.pdf) for more details
- Substituting a synonym sounds unnatural:
  - "fast food" vs. "quick food"
  - "Great Britain" vs. "Good Britain"
  - "warm greetings" vs "hot greetings"
- How can we find collocations in a corpus of text?

Pointwise Mutual Information

- The **pointwise mutual information (PMI)** between words x and y is
$$\mathrm{pmi}(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$$
  - $p(x)p(y)$ is how frequently we **expect** x and y to co-occur, if x and y are independent.
  - $p(x,y)$ measures how frequently x and y **actually** occur together
- **Idea:** Rank word pairs by $\mathrm{pmi}(x,y)$ to find collocations!
  - $\mathrm{pmi}(x,y)$ is large if x and y co-occur more frequently together than expected
- **Example:** Let's try it on the novel *Crime and Punishment*!
  - Pre-computed unigram and bigram counts are found in the `collocations/data` folder

Pointwise Mutual Information: Example

- Here are the most frequent bigrams--these aren't collocations!

Bigram	in the	of the	he was	he had	to the	on the	i am	at the	it was	that he
Count	778	598	505	498	488	479	460	459	413	335

- Sorting bigrams by PMI, we first get names...

Collocation	andrey semyonovitch	nikodim fomitch	hay market	dmitri prokofitch	honoured sir	sofya semyonovna	marfa petrovna	police station	rodion romanovitch
PMI	-3.18	-3.18	-3.48	-3.87	-4.27	-4.33	-4.37	-4.48	-4.57

- ...then more interesting collocations! This is much more useful than sorting by frequency alone.

Collocation	thank god	police office	great deal	ten minutes	good heavens	thousand roubles	katerina ivanovnas	old womans
PMI	-5.20	-5.23	-5.28	-5.40	-5.51	-5.54	-5.57	-5.57

Exponential Families

> Uses material from **[MLAPP]** §9.2 and **[Hero 2008]** §3.5, §4.4.2

Exponential Family: Introduction

- We have seen many distributions.
  - Bernoulli
  - Gaussian
  - Exponential
  - Gamma
- Many of these belong to a more general class called the **exponential family**.
- Why do we care?
  - only family of distributions with finite-dimensional **sufficient statistics**
  - only family of distributions for which **conjugate priors** exist
  - makes the least set of assumptions subject to some user-chosen constraints (**Maximum Entropy**)
  - core of generalized linear models and **variational inference**

Sufficient Statistics: Definition

- **Recall:** A **statistic** $T(\mathcal{D})$ is a function of the observed data $\mathcal{D}$.
  - Mean, $T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^{n} x_k$
  - Variance, maximum, mode, etc.
- Suppose we have some distribution with parameters θ. Then,

> A statistic $T(\mathcal{D})$ is **sufficient** for θ if no other statistic calculated from the same sample provides any additional information about θ.

- Mathematically,

$$P(\theta \,|\, \mathcal{D}, T(\mathcal{D})) = P(\theta \,|\, T(\mathcal{D}))$$

  Given statistic $T(\mathcal{D})$, θ is independent of data $\mathcal{D}$

Sufficient Statistics: Example

- Suppose $X \sim \text{Bernoulli}(\theta)$, i.e. $P(X=1) = \theta, P(X=0) = 1-\theta$ and we observe $\mathcal{D} = \{x_1, \ldots, x_N\} \in \{0,1\}^N$
- Then statistic $T(\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} x_n$, i.e. number of occurrence, is *sufficient* for θ

- **Proof for sufficiency**
  - Let $\tau = T(\mathcal{D})$, we have
$$\begin{aligned} P(\mathcal{D}\,|\,\theta) &= P(\mathcal{D},\tau\,|\,\theta) \\ &= \theta^{\tau}(1-\theta)^{N-\tau} \end{aligned} \qquad P(\tau\,|\,\theta) = \binom{N}{\tau}\theta^{\tau}(1-\theta)^{N-\tau} \qquad P(\mathcal{D}\,|\,\tau) = 1\Big/\binom{N}{\tau}$$
    Therefore,
$$P(D,\tau\,|\,\theta) = P(\tau\,|\,\theta)P(D\,|\,\tau)$$
  - For $P(\theta\,|\,\mathcal{D},\tau)$, we have
$$\begin{aligned} P(\theta\,|\,\mathcal{D},\tau) &= \frac{P(\mathcal{D},\tau\,|\,\theta)P(\theta)}{P(\mathcal{D},\tau)} = \frac{P(\tau\,|\,\theta)P(D\,|\,\tau)P(\theta)}{P(\mathcal{D},\tau)} \\ &= \frac{P(\tau\,|\,\theta)P(D\,|\,\tau)P(\theta)}{P(\mathcal{D}\,|\,\tau)P(\tau)} = \frac{P(\tau\,|\,\theta)P(\theta)}{P(\tau)} \\ &= P(\theta\,|\,\tau) \qquad \mathbf{Q.E.D.} \end{aligned}$$

Exponential Family: Definition

- $p(x\,|\,\theta)$ has **exponential family form** if:
$$\begin{aligned} p(x\,|\,\theta) &= \frac{1}{Z(\theta)}h(x)\exp\left[\eta(\theta)^T\phi(x)\right] \\ &= h(x)\exp\left[\eta(\theta)^T\phi(x) - A(\theta)\right] \end{aligned}$$
  of which $p(x\,|\,\theta)$ means *distribution of x parameterized by θ*
  - $Z(\theta) = \int h(x)\exp\left[\eta(\theta)^T\phi(x)\right]\mathrm{d}x$ is the **partition function** for normalization
  - $A(\theta) = \log Z(\theta)$ is the **log partition function**
  - $\phi(x) \in \mathbb{R}^d$ is a vector of **sufficient statistics**
  - $\eta(\theta)$ maps θ to a set of **natural parameters**
  - $h(x)$ is a scaling constant, usually $h(x) = 1$

Exponential Family: Example—Bernoulli

- The Bernoulli distribution can be written as
$$\begin{aligned} \mathrm{Bern}(x\,|\,\theta) &= \theta^x(1-\theta)^{1-x} \\ &= \exp[x\log\theta + (1-x)\log(1-\theta)] \\ &= \exp\left[\eta(\theta)^T\phi(x)\right] \end{aligned}$$
  where $\eta(\theta) = (\log\theta, \log(1-\theta))$ and $\phi(x) = (x, 1-x)$
  - There is a linear dependence between features $\phi(x)$
  - This representation is **overcomplete**
  - η is not uniquely determined
- Instead, we can find a **minimal** parameterization:
$$\mathrm{Ber}(x\,|\,\theta) = (1-\theta)\exp\left[x\log\frac{\theta}{1-\theta}\right]$$
- This gives **natural parameters** $\eta(\theta) = \log\frac{\theta}{1-\theta}$.
  - Now, η is unique

Exponential Family: Example—Gaussian

- The Gaussian distribution can be written as
$$\begin{aligned} \mathcal{N}(x\,|\,\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\exp\left\{\begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2}\end{bmatrix}\begin{bmatrix} x^2 \\ x\end{bmatrix}\right\} \end{aligned}$$
  of which
$$\frac{1}{Z(\theta)} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \qquad \eta(\theta) = \begin{bmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2}\end{bmatrix} \qquad \phi(x) = \begin{bmatrix} x^2 \\ x\end{bmatrix}$$

Exponential Family: Example—Others

- Exponential Family Distributions:
  - Multivariate normal
  - Exponential
  - Dirichlet
- Non-examples:
  - Student t-distribution can't be written in exponential form
  - Uniform distribution support depends on the parameters θ

Exponential Family: Notation Change

- Recall our exponential family has the form

$$p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp\left[\eta(\theta)^T \phi(x)\right] = h(x) \exp\left[\eta(\theta)^T \phi(x) - A(\theta)\right]$$

  of which natural parameter is $\eta(\theta)$.
- Now we change the notation a little bit
  - let θ denote **natural parameter**, i.e. replace $\eta(\theta)$ with θ, so that we could manipulate natural parameter directly. So we have a new form of exponential family

$$p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp\left[\theta^T \phi(x)\right] = h(x) \exp\left[\theta^T \phi(x) - A(\theta)\right]$$

  - Note that this new function $Z(\theta)$ and $A(\theta)$ is different from old $Z(\theta)$ and $A(\theta)$ because we have changed the notation of θ
- After this notation change, we have log-partition function:

$$A(\theta) = \log Z(\theta) = \log \int h(x) \exp\left[\theta^T \phi(x)\right] \mathrm{d}x$$

Exponential Family: Log-partition Function

- Recall our log-partition function is

$$A(\theta) = \log \int h(x) \exp\left[\theta^T \phi(x)\right] \mathrm{d}x$$

- Derivatives of **log-partition function** $A(\theta)$ yield **cumulants** of sufficient statistics (Proof in the note)
  - $\nabla_\theta A(\theta) = E\left[\phi(x)\right]$
  - $\nabla_\theta^2 A(\theta) = Cov[\phi(x)]$
- Since covariance $Cov[\phi(x)]$ is positive definite,i.e. $Cov[\phi(x)] \succ 0$, we have
  - $\nabla_\theta^2 A(\theta)$ is positive definite
  - and $A(\theta)$ is *strictly convex*!
- Later, we will see this could guarantee a unique global maximum of the likelihood $P(\mathcal{D} \mid \theta)$

Remark

- Proof of Convexity: **First Derivative**

$$\begin{aligned}
\frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log \int h(x) \exp\left[\theta^T \phi(x)\right] \mathrm{d}x\right] \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} \int h(x) \exp\left[\theta^T \phi(x)\right] \mathrm{d}x}{\int h(x) \exp[\theta^T \phi(x)] \mathrm{d}x} \\
&= \frac{\int \phi(x) h(x) \exp\left[\theta^T \phi(x)\right] \mathrm{d}x}{\exp[A(\theta)]} \\
&= \int \phi(x) \underbrace{h(x) \exp\left[\theta^T \phi(x) - A(\theta)\right]}_{p(x)} \mathrm{d}x \\
&= \int \phi(x) p(x) dx = E[\phi(x)]
\end{aligned}$$

- Proof of Convexity: **Second Derivative**
- Recall we just derived

$$\frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta} = \int \phi(x)h(x)\exp\big[\theta^T\phi(x) - A(\theta)\big]\mathrm{d}x$$

So the second derivative is

$$\begin{aligned}\frac{\mathrm{d}^2A}{\mathrm{d}\theta^2} &= \int \phi(x)h(x)\exp\big[\theta^T\phi(x) - A(\theta)\big]\left[\phi(x) - \frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta}\right]\mathrm{d}x \\ &= \int \phi(x)p(x)\left[\phi(x) - \frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta}\right]\mathrm{d}x \\ &= \int \phi^2(x)p(x)\mathrm{d}x - \frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta}\int \phi(x)p(x)\mathrm{d}x \\ &= E[\phi^2(x)] - E[\phi(x)]^2 \qquad (\because \mathrm{d}A(\theta)/\mathrm{d}\theta = E[\phi(x)]) \\ &= Var[\phi(x)]\end{aligned}$$

- For multi-variate case, we have

$$\frac{\partial^2 A}{\partial\theta_i\partial\theta_j} = E[\phi_i(x)\phi_j(x)] - E[\phi_i(x)]E[\phi_j(x)]$$

and hence,

$$\nabla^2 A(\theta) = Cov[\phi(x)]$$

Since covariance is positive definite, we have $A(\theta)$ strictly convex as required.

Exponential Family: Likelihood

- For single data x_n, its likelihood is

$$p(x_n \,|\, \theta) = h(x_n)\exp\big[\theta^T\phi(x_n) - A(\theta)\big]$$

- For data $\mathcal{D} = \{x_1, \ldots, x_N\}$, the likelihood is

$$p(\mathcal{D}\,|\,\theta) = \left[\prod\nolimits_{n=1}^N h(x_n)\right]\exp\left[\theta^T\sum\nolimits_{n=1}^N \phi(x_n) - NA(\theta)\right]$$

- The sufficient statistics are now $\phi(\mathcal{D}) = \sum_{n=1}^N \phi(x_n)$.
  - **Bernoulli:** $\phi(\mathcal{D}) = \sum_{n=1}^N x_n$
  - **Normal:** $\phi(\mathcal{D}) = [\sum_n x_n, \sum_n x_n^2]$
- The log-likelihood is (we have omitted terms independent of θ)

$$\log p(\mathcal{D}\,|\,\theta) = \theta^T\phi(\mathcal{D}) - NA(\theta)$$

- Since $-A(\theta)$ is *strictly concave* and $\theta^T\phi(\mathcal{D})$ *linear* w.r.t θ,
  - the log-likelihood is *strictly concave*
  - there is a *unique* global maximum for likelihood!
  - we have *unique* **maximum likelihood estimate (MLE)** for θ!

Exponential Family: MLE

- At the MLE $\hat{\theta}_{MLE}$, we have

$$\nabla_\theta \log p(\mathcal{D}\,|\,\theta) = 0$$

- For the derivative of log-likelihood, we have

$$\nabla_\theta \log p(\mathcal{D}\,|\,\theta) = \nabla_\theta\left[\theta^T\phi(\mathcal{D}) - NA(\theta)\right] \overset{\nabla_\theta A(\theta) = E[\phi(x)]}{=} \phi(\mathcal{D}) - NE[\phi(X)]$$

- In conclusion, at the MLE $\hat{\theta}_{MLE}$ we have

$$E[\phi(x)] = \frac{\phi(\mathcal{D})}{N} = \frac{1}{N}\sum\nolimits_{n=1}^N \phi(x_n)$$

  - Expected value (parameterized by θ) of sufficient statistics equals empirical average of them when $\theta = \hat{\theta}_{MLE}$
  - This is called **moment matching**
  - We could obtain MLE in this way

Exponential Family: MLE—Bernoulli

- Recall we just showed for MLE $\hat{\theta}_{MLE}$, we have
$$E[\phi(X)] = \frac{1}{N}\sum_{n=1}^{N}\phi(x_n)$$
- For $\text{Bernoulli}(\theta)$, we know
$$E[\phi(X)] = E[x] = \theta$$
and we have showed
$$\phi(x) = x$$
- So the MLE $\hat{\theta}_{MLE}$ can be obtained by
$$\hat{\theta}_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$$