COMS 4771 Lecture 21

1. Exponential families

Exponential families

Exponential families

We saw that solutions to the maximum entropy optimization problem are distributions of the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle - G(\boldsymbol{\eta}) \right\} \cdot \pi(x) \quad \forall x \in \mathcal{X}. \qquad (\star)$$

- π: base distribution over domain $\mathcal{X}$ (okay if unnormalized).
- $\boldsymbol{T} \colon \mathcal{X} \to \mathbb{R}^k$: vector-valued feature function.
- $G(\boldsymbol{\eta})$: $\log$ of normalizer that makes $P_{\boldsymbol{\eta}}$ a probability distribution over $\mathcal{X}$.

Exponential families

We saw that solutions to the maximum entropy optimization problem are distributions of the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - G(\boldsymbol{\eta})\Big\} \cdot \pi(x) \quad \forall x \in \mathcal{X}. \tag{$\star$}$$

- π: base distribution over domain $\mathcal{X}$ (okay if unnormalized).
- $\boldsymbol{T} \colon \mathcal{X} \to \mathbb{R}^k$: vector-valued feature function.
- $G(\boldsymbol{\eta})$: $\log$ of normalizer that makes $P_{\boldsymbol{\eta}}$ a probability distribution over $\mathcal{X}$.

Each choice of π and $\boldsymbol{T}$ leads to a **family of probability distributions**

$$\Big\{P_{\boldsymbol{\eta}} \text{ as in } (\star) : \boldsymbol{\eta} \in \mathbb{R}^k\Big\}.^{\ddagger}$$

We call such a family an **exponential family**.

Exponential families

We saw that solutions to the maximum entropy optimization problem are distributions of the form

?

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - G(\boldsymbol{\eta})\Big\} \cdot \pi(x) \quad \forall x \in \mathcal{X}. \qquad (\star)$$

- π: base distribution over domain $\mathcal{X}$ (okay if unnormalized).
- $\boldsymbol{T}\colon \mathcal{X} \to \mathbb{R}^k$: vector-valued feature function.
- $G(\boldsymbol{\eta})$: $\log$ of normalizer that makes $P_{\boldsymbol{\eta}}$ a probability distribution over $\mathcal{X}$.

Each choice of π and $\boldsymbol{T}$ leads to a **family of probability distributions**

$$\Big\{P_{\boldsymbol{\eta}} \text{ as in } (\star) : \boldsymbol{\eta} \in \mathbb{R}^k\Big\}.^{\ddagger}$$

We call such a family an **exponential family**.

‡ **This is not entirely accurate, due to an important technicality.**

Important technicality

The log of the normalizer (which is also called the **log partition function**),

$$G(\boldsymbol{\eta}) \;=\; \ln Z(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\pi(x)\right),$$

could be infinite for some values of $\boldsymbol{\eta} \in \mathbb{R}^k$.

Important technicality

The log of the normalizer (which is also called the **log partition function**),

$$G(\boldsymbol{\eta}) \;=\; \ln Z(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x\in\mathcal{X}} \exp\Big\{\langle\boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\pi(x)\right),$$

could be infinite for some values of $\boldsymbol{\eta} \in \mathbb{R}^k$.

Example: $\mathcal{X} = \mathbb{N}$, $T_1(x) = \ln(x)$, $\pi(x) = 1/x^2$, so

$$G(\eta) \;=\; \sum_{x\in\mathbb{N}} \exp\Big\{\eta\ln(x)\Big\} \cdot \frac{1}{x^2}$$

Important technicality

The log of the normalizer (which is also called the **log partition function**),

$$G(\boldsymbol{\eta}) \;=\; \ln Z(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\pi(x)\right),$$

could be infinite for some values of $\boldsymbol{\eta} \in \mathbb{R}^k$.

Example: $\mathcal{X} = \mathbb{N}$, $T_1(x) = \ln(x)$, $\pi(x) = 1/x^2$, so

$$G(\eta) \;=\; \sum_{x\in\mathbb{N}} \exp\Big\{\eta \ln(x)\Big\} \cdot \frac{1}{x^2} \;=\; \sum_{x\in\mathbb{N}} \frac{1}{x^{2-\eta}}$$

which finite if and only if $\eta < 1$.

IMPORTANT TECHNICALITY

The $\boxed{\text{log of the normalizer}}$ (which is also called the **log partition function**),

$$G(\boldsymbol{\eta}) = \boxed{\ln Z(\boldsymbol{\eta})} = \ln\left(\sum_{x\in\mathcal{X}} \exp\Big\{\langle\boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\pi(x)\right),$$

could be infinite for some values of $\boldsymbol{\eta} \in \mathbb{R}^k$.

Example: $\mathcal{X} = \mathbb{N}$, $T_1(x) = \ln(x)$, $\pi(x) = 1/x^2$, so

$$G(\eta) = \sum_{x\in\mathbb{N}} \exp\Big\{\eta \ln(x)\Big\} \cdot \frac{1}{x^2} = \sum_{x\in\mathbb{N}} \frac{1}{x^{2-\eta}}$$

which finite if and only if $\eta < 1$.

Parameter values $\eta \geq 1$ cause $G(\eta)$ to be infinite, and hence do not yield valid probability distributions.

Revised definition
The **exponential family** corresponding to $\boldsymbol{T}$ and π is

$$\left\{P_{\boldsymbol{\eta}} \text{ as in } (\star) : \boldsymbol{\eta} \in \mathcal{N}\right\}$$

where

$$\boxed{\mathcal{N} \;=\; \left\{\boldsymbol{\eta} \in \mathbb{R}^k : G(\boldsymbol{\eta}) < \infty\right\}}$$

is the **natural parameter space** for this exponential family.

$$\begin{aligned} P_{\boldsymbol{\eta}}(x) &= \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - G(\boldsymbol{\eta})\right\} \cdot \pi(x) \quad \forall x \in \mathcal{X}, \\ G(\boldsymbol{\eta}) &= \ln\left(\sum_{x \in \mathcal{X}} \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\right\}\pi(x)\right). \end{aligned} \qquad (\star)$$

EXAMPLE #1

$\mathcal{X} = \{0, 1\}$, $T_1(x) = \mathbb{1}\{x = 1\} = x$, $\pi(x) = 1$.

Example #1

equal meaning

$\mathcal{X} = \{0, 1\}$, $T_1(x) = \boxed{\mathbb{1}\{x = 1\} = x,}$ $\pi(x) = 1$.

What is the natural parameter space $\mathcal{N} = \big\{\eta \in \mathbb{R} : G(\eta) < \infty\big\}$?

Example #1

$\mathcal{X} = \{0,1\}$, $T_1(x) = \mathbb{1}\{x=1\} = x$, $\pi(x) = 1$.

What is the natural parameter space $\mathcal{N} = \big\{\eta \in \mathbb{R} : G(\eta) < \infty\big\}$?

$$G(\eta) \;=\; \ln\left(\sum_{x\in\{0,1\}} \exp\{\eta x\}\right) \;=\; \ln\big(1+\exp(\eta)\big) \;<\; \infty \quad \text{for all } \eta \in \mathbb{R},$$

so $\mathcal{N} = \mathbb{R}$.

Example #1

$\mathcal{X} = \{0, 1\}$, $T_1(x) = \mathbb{1}\{x = 1\} = x$, $\pi(x) = 1$.

What is the natural parameter space $\mathcal{N} = \big\{\eta \in \mathbb{R} : G(\eta) < \infty\big\}$?

$$G(\eta) \;=\; \ln\left(\sum_{x\in\{0,1\}} \exp\{\eta x\}\right) \;=\; \ln\big(1 + \exp(\eta)\big) \;<\; \infty \quad \text{for all } \eta \in \mathbb{R},$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Big\{\eta x - \ln\big(1 + \exp(\eta)\big)\Big\}$$

Example #1

$\mathcal{X} = \{0, 1\}$, $T_1(x) = \mathbb{1}\{x = 1\} = x$, $\pi(x) = 1$.

What is the natural parameter space $\mathcal{N} = \big\{\eta \in \mathbb{R} : G(\eta) < \infty\big\}$?

$$G(\eta) \;=\; \ln\left(\sum_{x \in \{0,1\}} \exp\{\eta x\}\right) \;=\; \ln\big(1 + \exp(\eta)\big) \;<\; \infty \quad \text{for all } \eta \in \mathbb{R},$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Big\{\eta x - \ln\big(1 + \exp(\eta)\big)\Big\} \;=\; \frac{\exp(\eta x)}{1 + \exp(\eta)}$$

EXAMPLE #1

$\mathcal{X} = \{0, 1\}$, $T_1(x) = \mathbb{1}\{x = 1\} = x$, $\pi(x) = 1$.

What is the natural parameter space $\mathcal{N} = \big\{\eta \in \mathbb{R} : G(\eta) < \infty\big\}$?

$$G(\eta) \;=\; \ln\left(\sum_{x\in\{0,1\}} \exp\{\eta x\}\right) \;=\; \ln\big(1 + \exp(\eta)\big) \;<\; \infty \quad \text{for all } \eta \in \mathbb{R},$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Big\{\eta x - \ln\big(1 + \exp(\eta)\big)\Big\} \;=\; \frac{\exp(\eta x)}{1 + \exp(\eta)} \;=\; \begin{cases} \dfrac{1}{1 + e^\eta} & \text{if } x = 0; \\ \dfrac{e^\eta}{1 + e^\eta} & \text{if } x = 1. \end{cases}$$

\pi (x) is a constant function.

equal meaning

$$\mathcal{X} = \{0, 1\}, \ \ T_1(x) = \boxed{\mathbb{1}\{x = 1\}} = x, \ \ \pi(x) = 1.$$

What is the natural parameter space $\mathcal{N} = \{\eta \in \mathbb{R} : G(\eta) < \infty\}$?

$$G(\eta) \ = \ \ln\left(\sum_{x \in \{0,1\}} \exp\{\eta x\}\right) \ = \ \ln\big(1 + \exp(\eta)\big) \ < \ \infty \quad \text{for all } \eta \in \mathbb{R},$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \ = \ \exp\Big\{\eta x - \ln\big(1 + \exp(\eta)\big)\Big\} \ = \ \frac{\exp(\eta x)}{1 + \exp(\eta)} \ = \ \begin{cases} \dfrac{1}{1 + e^\eta} & \text{if } x = 0; \\ \dfrac{e^\eta}{1 + e^\eta} & \text{if } x = 1. \end{cases}$$

Bernoulli distributions $\mathrm{Bern}\left(\frac{e^\eta}{1+e^\eta}\right)$

Example #1.5

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ for $i \in \{1, \ldots, d-1\}$, $\pi(x) = 1$.

Example #1.5

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ for $i \in \{1, \ldots, d-1\}$, $\pi(x) = 1$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x=0}^{d-1} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\right) \;=\; \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)$$

so $\mathcal{N} = \mathbb{R}^{d-1}$.

Example #1.5

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ for $i \in \{1, \ldots, d-1\}$, $\pi(x) = 1$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x=0}^{d-1} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\right) \;=\; \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)$$

so $\mathcal{N} = \mathbb{R}^{d-1}$.

Probability distributions have the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)\right\}$$

Example #1.5

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ for $i \in \{1, \ldots, d-1\}$, $\pi(x) = 1$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x=0}^{d-1} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\right) \;=\; \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)$$

so $\mathcal{N} = \mathbb{R}^{d-1}$.

Probability distributions have the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)\right\} \;=\; \begin{cases} \dfrac{1}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = 0; \\ \dfrac{e^{\eta_i}}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = i \neq 0. \end{cases}$$

Example #1.5

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ for $i \in \{1, \ldots, d-1\}$, $\pi(x) = 1$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x=0}^{d-1} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\}\right) \;=\; \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)$$

so $\mathcal{N} = \mathbb{R}^{d-1}$.

Probability distributions have the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)\right\} \;=\; \begin{cases} \dfrac{1}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = 0; \\ \dfrac{e^{\eta_i}}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = i \neq 0. \end{cases}$$

Categorical distributions (generalizes Bernoulli)
For $X \sim P_{\boldsymbol{\eta}}$, $\Pr[X = i] \propto e^{\eta_i}$ for $i \neq 0$.

Example #1.5

no need to include 0, redundant! can get the information from other Ti

$\mathcal{X} = \{0, 1, \ldots, d-1\}$, $T_i(x) = \mathbb{1}\{x = i\}$ $\boxed{\text{for } i \in \{1, \ldots, d-1\},}$ $\pi(x) = 1$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\sum_{x=0}^{d-1} \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\right\}\right) \;=\; \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)$$

so $\mathcal{N} = \mathbb{R}^{d-1}$.

Probability distributions have the form

$$P_{\boldsymbol{\eta}}(x) \;=\; \exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - \ln\left(1 + \sum_{i=1}^{d-1} e^{\eta_i}\right)\right\} \;=\; \begin{cases} \dfrac{1}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = 0; \\[2ex] \dfrac{e^{\eta_i}}{1 + \sum_{i=1}^{d-1} e^{\eta_i}} & \text{if } x = i \neq 0. \end{cases}$$

Categorical distributions (generalizes Bernoulli)
For $X \sim P_{\boldsymbol{\eta}}$, $\Pr[X = i] \propto e^{\eta_i}$ for $i \neq 0$.

Note: keep track of $d-1$ features—almost one for every $x \in \mathcal{X}$.

Example #2

$\mathcal{X} =$ non-negative integers, $T_1(x) = x$, $\pi(x) = 1/x!$.

EXAMPLE #2

$\mathcal{X}$ = non-negative integers, $T_1(x) = x$, $\pi(x) = 1/x!$.

Using the fact that $\ln\Big(\sum_{j=0}^{\infty} z^j/j!\Big) = \ln\exp(z) = z$,

$$G(\eta) \;=\; \ln\left(\sum_{x=0}^{\infty} \exp\Big\{\eta x\Big\} \cdot \frac{1}{x!}\right) \;=\; \exp(\eta),$$

so $\mathcal{N} = \mathbb{R}$.

Example #2

$\mathcal{X} =$ non-negative integers, $T_1(x) = x$, $\pi(x) = 1/x!$.

Using the fact that $\ln\Big(\sum_{j=0}^{\infty} z^j/j!\Big) = \ln\exp(z) = z$,

$$G(\eta) \;=\; \ln\left(\sum_{x=0}^{\infty} \exp\Big\{\eta x\Big\} \cdot \frac{1}{x!}\right) \;=\; \exp(\eta),$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Big\{\eta x - \exp(\eta)\Big\} \cdot \frac{1}{x!}$$

Example #2

$\mathcal{X} =$ non-negative integers, $T_1(x) = x$, $\pi(x) = 1/x!$.

Using the fact that $\ln\Bigl(\sum_{j=0}^{\infty} z^j/j!\Bigr) = \ln\exp(z) = z$,

$$G(\eta) \;=\; \ln\left(\sum_{x=0}^{\infty} \exp\Bigl\{\eta x\Bigr\} \cdot \frac{1}{x!}\right) \;=\; \exp(\eta),$$

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Bigl\{\eta x - \exp(\eta)\Bigr\} \cdot \frac{1}{x!} \;=\; \frac{(e^\eta)^x \cdot \exp(-e^\eta)}{x!}.$$

Example #2

$\mathcal{X}$ = non-negative integers, $T_1(x) = x$, $\pi(x) = 1/x!$.

Using the fact that $\ln\left(\sum_{j=0}^{\infty} z^j/j!\right) = \ln\exp(z) = z$,

$$G(\eta) \;=\; \ln\left(\sum_{x=0}^{\infty} \exp\Big\{\eta x\Big\} \cdot \frac{1}{x!}\right) \;=\; \exp(\eta),$$

always converage

so $\mathcal{N} = \mathbb{R}$.

Probability distributions have the form

$$P_\eta(x) \;=\; \exp\Big\{\eta x - \exp(\eta)\Big\} \cdot \frac{1}{x!} \;=\; \frac{(e^\eta)^x \cdot \exp(-e^\eta)}{x!}.$$

Poisson distributions $\mathrm{Pois}(e^\eta)$

Continuous domains

If $\mathcal{X}$ is a continuous domain (e.g., $\mathbb{R}^d$), we can also obtain exponential families comprised of probability densities having the form

$$\begin{aligned} p_{\boldsymbol{\eta}}(x) &= \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle - G(\boldsymbol{\eta}) \right\} \cdot \pi(x) \quad \forall x \in \mathcal{X}, \\ G(\boldsymbol{\eta}) &= \ln\left(\int_{\mathcal{X}} \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \right\} \cdot \pi(x) \, \mathrm{d}x \right). \end{aligned} \tag{$*$}$$

where π is a "base measure" (often a density) on $\mathcal{X}$.

Continuous domains

If $\mathcal{X}$ is a continuous domain (e.g., $\mathbb{R}^d$), we can also obtain exponential families comprised of probability densities having the form

$$\begin{aligned} p_{\boldsymbol{\eta}}(x) &= \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle - G(\boldsymbol{\eta}) \right\} \cdot \pi(x) \quad \forall x \in \mathcal{X}, \\ G(\boldsymbol{\eta}) &= \ln\left(\int_{\mathcal{X}} \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \right\} \cdot \pi(x) \, \mathrm{d}x \right). \end{aligned} \tag{$*$}$$

where π is a "base measure" (often a density) on $\mathcal{X}$.

The exponential family corresponding to $\boldsymbol{T}$ and π is

$$\left\{ P_{\boldsymbol{\eta}} \text{ as in } (*) : \boldsymbol{\eta} \in \mathcal{N} \right\} \text{ where } \mathcal{N} = \left\{ \boldsymbol{\eta} \in \mathbb{R}^k : G(\boldsymbol{\eta}) < \infty \right\}.$$

Continuous domains

If $\mathcal{X}$ is a continuous domain (e.g., $\mathbb{R}^d$), we can also obtain exponential families comprised of probability densities having the form

$$\begin{aligned} p_{\boldsymbol{\eta}}(x) &= \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle - G(\boldsymbol{\eta}) \right\} \cdot \pi(x) \quad \forall x \in \mathcal{X}, \\ G(\boldsymbol{\eta}) &= \ln\left(\int_{\mathcal{X}} \exp\left\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \right\} \cdot \pi(x) \, \mathrm{d}x \right). \end{aligned} \qquad (*)$$

where π is a "base measure" (often a density) on $\mathcal{X}$.

The exponential family corresponding to $\boldsymbol{T}$ and π is

$$\left\{ P_{\boldsymbol{\eta}} \text{ as in } (*) : \boldsymbol{\eta} \in \mathcal{N} \right\} \text{ where } \mathcal{N} = \left\{ \boldsymbol{\eta} \in \mathbb{R}^k : G(\boldsymbol{\eta}) < \infty \right\}.$$

(A more general treatment requires a little bit of measure theory, which we'll forego.)

Example #3

$$\mathcal{X} = \mathbb{R},\ \ T_1(x) = x,\ \ \pi(x) = \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Example #3

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $\pi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$$G(\eta) \;=\; \ln\!\left(\int_{\mathbb{R}} \exp\{\eta x\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x \right) \;=\; \frac{\eta^2}{2},$$

so $\mathcal{N} = \mathbb{R}$.

Example #3

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $\pi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$$G(\eta) \;=\; \ln\left(\int_{\mathbb{R}} \exp\{\eta x\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x \right) \;=\; \frac{\eta^2}{2},$$

so $\mathcal{N} = \mathbb{R}$.

$$p_\eta(x) \;=\; \exp\Big\{ \eta x - \eta^2/2 \Big\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Example #3

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $\pi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$$G(\eta) \;=\; \ln\left(\int_{\mathbb{R}} \exp\{\eta x\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x\right) \;=\; \frac{\eta^2}{2},$$

so $\mathcal{N} = \mathbb{R}$.

$$p_\eta(x) \;=\; \exp\Big\{\eta x - \eta^2/2\Big\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \;=\; \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\eta)^2}{2}\right\}.$$

EXAMPLE #3

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $\pi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$$G(\eta) \;=\; \ln\left(\int_{\mathbb{R}} \exp\{\eta x\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x\right) \;=\; \frac{\eta^2}{2},$$

so $\mathcal{N} = \mathbb{R}$.

$$p_\eta(x) \;=\; \exp\left\{\eta x - \eta^2/2\right\} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \;=\; \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\eta)^2}{2}\right\}.$$

Gaussian distributions with unit variance $\mathrm{N}(\eta, 1)$

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

$$G(\eta) \quad = \quad \ln\left(\int_{\mathbb{R}} \exp\left\{\eta_1 x + \eta_2 x^2\right\} \cdot \frac{1}{\sqrt{2\pi}}\, \mathrm{d}x\right)$$

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

$$\begin{aligned} G(\eta) &= \ln\left(\int_{\mathbb{R}} \exp\left\{\eta_1 x + \eta_2 x^2\right\} \cdot \frac{1}{\sqrt{2\pi}}\,\mathrm{d}x\right) \\ &= \ln\left(\int_{\mathbb{R}} \exp\left\{-\frac{1}{2} \cdot \frac{x^2 + \eta_1 x/\eta_2}{-1/(2\eta_2)}\right\} \cdot \frac{1}{\sqrt{2\pi}}\,\mathrm{d}x\right) \\ \text{(completing the square \ldots)} &= -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}. \end{aligned}$$

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

$$
\begin{aligned}
G(\eta) &= \ln\left(\int_{\mathbb{R}} \exp\left\{\eta_1 x + \eta_2 x^2\right\} \cdot \frac{1}{\sqrt{2\pi}}\,\mathrm{d}x\right) \\
&= \ln\left(\int_{\mathbb{R}} \exp\left\{-\frac{1}{2} \cdot \frac{x^2 + \eta_1 x/\eta_2}{-1/(2\eta_2)}\right\} \cdot \frac{1}{\sqrt{2\pi}}\,\mathrm{d}x\right) \\
\text{(completing the square \ldots)} &= -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}.
\end{aligned}
$$

The "completing the square" step assumes that $-1/(2\eta_2) > 0$, or else the integral would diverge. So $\mathcal{N} = \mathbb{R} \times \mathbb{R}_{--}$ (i.e., η_2 must be negative).

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

$$\begin{aligned} G(\eta) &= \ln\left(\int_{\mathbb{R}} \exp\left\{ \eta_1 x + \eta_2 x^2 \right\} \cdot \frac{1}{\sqrt{2\pi}} \, \mathrm{d}x \right) \\ &= \ln\left(\int_{\mathbb{R}} \exp\left\{ -\frac{1}{2} \cdot \frac{x^2 + \eta_1 x/\eta_2}{-1/(2\eta_2)} \right\} \cdot \frac{1}{\sqrt{2\pi}} \, \mathrm{d}x \right) \\ \text{(completing the square \ldots)} &= -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}. \end{aligned}$$

The "completing the square" step assumes that $-1/(2\eta_2) > 0$, or else the integral would diverge. So $\mathcal{N} = \mathbb{R} \times \mathbb{R}_{--}$ (i.e., η_2 must be negative).

$$p_\eta(x) \;=\; \exp\left\{ \eta_1 x + \eta_2 x^2 + \ln\sqrt{-2\eta_2} + \frac{\eta_1^2}{4\eta_2} \right\} \cdot \frac{1}{\sqrt{2\pi}}$$

Example #4

$\mathcal{X} = \mathbb{R}$, $T_1(x) = x$, $T_2(x) = x^2$, $\pi(x) = \frac{1}{\sqrt{2\pi}}$ (this is actually okay).

$$\begin{aligned} G(\eta) &= \ln\left(\int_{\mathbb{R}} \exp\left\{\eta_1 x + \eta_2 x^2\right\} \cdot \frac{1}{\sqrt{2\pi}}\, \mathrm{d}x\right) \\ &= \ln\left(\int_{\mathbb{R}} \exp\left\{-\frac{1}{2} \cdot \frac{x^2 + \eta_1 x/\eta_2}{-1/(2\eta_2)}\right\} \cdot \frac{1}{\sqrt{2\pi}}\, \mathrm{d}x\right) \\ \text{(completing the square ...)} &= -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}. \end{aligned}$$

The "completing the square" step assumes that $-1/(2\eta_2) > 0$, or else the integral would diverge. So $\mathcal{N} = \mathbb{R} \times \mathbb{R}_{--}$ (i.e., η_2 must be negative).

$$p_\eta(x) = \exp\left\{\eta_1 x + \eta_2 x^2 + \ln\sqrt{-2\eta_2} + \frac{\eta_1^2}{4\eta_2}\right\} \cdot \frac{1}{\sqrt{2\pi}} = \cdots$$

Gaussian distributions $\mathrm{N}\left(-\frac{\eta_1}{2\eta_2}, -\frac{1}{2\eta_2}\right)$

Parameterization

Natural parameters

In most of the examples so far, the "natural parameters" $\boldsymbol{\eta}$ are not the parameters we typically use for the distribution families.

Natural parameters

In most of the examples so far, the "natural parameters" $\boldsymbol{\eta}$ are not the parameters we typically use for the distribution families.

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^{\eta}}{1+e^{\eta}}$.

$$X \sim \mathrm{Bern}(p) \quad \Longrightarrow \quad \mathbb{E}[X] = p.$$

Natural parameters

In most of the examples so far, the "natural parameters" $\boldsymbol{\eta}$ are not the parameters we typically use for the distribution families.

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$X \sim \mathrm{Bern}(p) \quad \Longrightarrow \quad \mathbb{E}[X] = p.$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$X \sim \mathrm{Pois}(\lambda) \quad \Longrightarrow \quad \mathbb{E}[X] = \lambda.$$

Natural parameters

In most of the examples so far, the "natural parameters" $\boldsymbol{\eta}$ are not the parameters we typically use for the distribution families.

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$X \sim \mathrm{Bern}(p) \quad \Longrightarrow \quad \mathbb{E}[X] = p.$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$X \sim \mathrm{Pois}(\lambda) \quad \Longrightarrow \quad \mathbb{E}[X] = \lambda.$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad \mathbb{E}[X] = \mu, \quad \mathbb{E}[X^2] = \sigma^2 + \mu^2.$$

Natural parameters

In most of the examples so far, the "natural parameters" $\boldsymbol{\eta}$ are not the parameters we typically use for the distribution families.

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^{\eta}}{1+e^{\eta}}$.

$$X \sim \mathrm{Bern}(p) \quad \Longrightarrow \quad \mathbb{E}[X] = p.$$

usually use the form of E(X)

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^{\eta}$.

$$X \sim \mathrm{Pois}(\lambda) \quad \Longrightarrow \quad \mathbb{E}[X] = \lambda.$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad \mathbb{E}[X] = \mu, \quad \mathbb{E}[X^2] = \sigma^2 + \mu^2.$$

Is there a more general relationship between these expectations and the natural parameters?

Expectations

Let $\{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Fact: For any $\boldsymbol{\eta} \in \mathcal{N}$ (except on the boundary), all derivatives of the log partition function G exist at $\boldsymbol{\eta}$, and

$$X \sim P_{\boldsymbol{\eta}} \quad \Longrightarrow \quad \nabla G(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{T}(X)].$$

Let $\{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Fact: For any $\boldsymbol{\eta} \in \mathcal{N}$ (except on the boundary), all derivatives of the log partition function G exist at $\boldsymbol{\eta}$, and

$$X \sim P_{\boldsymbol{\eta}} \quad \Longrightarrow \quad \nabla G(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{T}(X)].$$

Proof in discrete case:
Recall that $G(\boldsymbol{\eta}) = \ln\Big(\sum_{x\in\mathcal{X}} \exp\big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\big\} \cdot \pi(x)\Big)$, so its gradient is

$$\nabla G(\boldsymbol{\eta}) \;=\; \frac{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x) \cdot \boldsymbol{T}(x)}{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x)}$$

Expectations

Let $\{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Fact: For any $\boldsymbol{\eta} \in \mathcal{N}$ (except on the boundary), all derivatives of the log partition function G exist at $\boldsymbol{\eta}$, and

$$X \sim P_{\boldsymbol{\eta}} \quad \Longrightarrow \quad \nabla G(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{T}(X)].$$

Proof in discrete case:
Recall that $G(\boldsymbol{\eta}) = \ln\Big(\sum_{x\in\mathcal{X}} \exp\big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\big\} \cdot \pi(x)\Big)$, so its gradient is

$$\nabla G(\boldsymbol{\eta}) \;=\; \frac{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x) \cdot \boldsymbol{T}(x)}{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x)} \;=\; \sum_{x\in\mathcal{X}} P_{\boldsymbol{\eta}}(x)\cdot\boldsymbol{T}(x) \;=\; \mathbb{E}[\boldsymbol{T}(X)].$$

$\square$

Expectations

Let $\{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Fact: For any $\boldsymbol{\eta} \in \mathcal{N}$ (except on the boundary), all derivatives of the log partition function G exist at $\boldsymbol{\eta}$, and

$$X \sim P_{\boldsymbol{\eta}} \quad \Longrightarrow \quad \nabla G(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{T}(X)].$$

\lambda is the target for gradient

X is also a random variable

Proof in discrete case:
Recall that $G(\boldsymbol{\eta}) = \ln\Big(\sum_{x\in\mathcal{X}} \exp\big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\big\} \cdot \pi(x)\Big)$, so its gradient is

$$\nabla G(\boldsymbol{\eta}) \;=\; \frac{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x) \cdot \boldsymbol{T}(x)}{\displaystyle\sum_{x\in\mathcal{X}} \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\Big\} \cdot \pi(x)} \;=\; \sum_{x\in\mathcal{X}} P_{\boldsymbol{\eta}}(x)\cdot\boldsymbol{T}(x) \;=\; \mathbb{E}[\boldsymbol{T}(X)].$$

the gradient gives the

$\square$

In fact, $\nabla^2 G(\boldsymbol{\eta}) = \mathrm{cov}(\boldsymbol{T}(X))$, and higher-order derivatives correspond to higher-order moments of $\boldsymbol{T}(X)$.

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big)$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\bigl(1 + \exp(\eta)\bigr), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta)$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\bigl(1 + \exp(\eta)\bigr), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1 + e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$G(\eta_1, \eta_2) = -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$G(\eta_1, \eta_2) = -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}, \quad \frac{\partial G(\eta_1, \eta_2)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}[X]$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$G(\eta_1, \eta_2) = -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}, \quad \frac{\partial G(\eta_1, \eta_2)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}[X],$$

$$\frac{\partial G(\eta_1, \eta_2)}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \sigma^2 + \mu^2 = \mathbb{E}[X^2].$$

Expectations: examples

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$G(\eta_1, \eta_2) = -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}, \quad \frac{\partial G(\eta_1, \eta_2)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}[X],$$

$$\frac{\partial G(\eta_1, \eta_2)}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \sigma^2 + \mu^2 = \mathbb{E}[X^2].$$

In each of these cases, the distributions are also parameterized by these expectations. Is this true for all exponential families?

we don't care about
\lambda, we want to get p

Example #1: Bernoulli distribution $\mathrm{Bern}(p)$ for $p = \frac{e^\eta}{1+e^\eta}$.

$$G(\eta) = \ln\big(1 + \exp(\eta)\big), \quad \frac{\partial G(\eta)}{\partial \eta} = \frac{e^\eta}{1+e^\eta} = p = \mathbb{E}[X].$$

Example #2: Poisson distribution $\mathrm{Pois}(\lambda)$ for $\lambda = e^\eta$.

$$G(\eta) = \exp(\eta), \quad \frac{\partial G(\eta)}{\partial \eta} = e^\eta = \lambda = \mathbb{E}[X].$$

Example #4: Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ for $\mu = -\frac{\eta_1}{2\eta_2}$ and $\sigma^2 = -\frac{1}{2\eta_2}$.

$$G(\eta_1, \eta_2) = -\ln\sqrt{-2\eta_2} - \frac{\eta_1^2}{4\eta_2}, \quad \frac{\partial G(\eta_1, \eta_2)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}[X],$$

$$\frac{\partial G(\eta_1, \eta_2)}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \sigma^2 + \mu^2 = \mathbb{E}[X^2].$$

In each of these cases, the distributions are also parameterized by these expectations. Is this true for all exponential families? Yes!

Expectation parameters

$g := \nabla G$ **is an invertible map between natural parameters and expectations**:

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\eta}) \quad \Longleftrightarrow \quad \boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}).$$

Expectation parameters

$\boldsymbol{g} := \nabla G$ **is an invertible map between natural parameters and expectations**:

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\eta}) \quad \iff \quad \boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}).$$

Fact: G is a strictly convex function.[‡] (Proof by generalized Cauchy-Schwarz ineq.)

[‡]Strict convexity, as opposed to just convexity, actually requires that the exponential family be *minimal*: feature functions should be linearly independent.

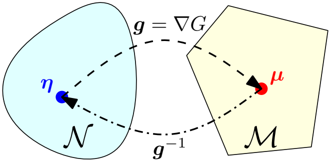
Expectation parameters

$\boldsymbol{g} := \nabla G$ **is an invertible map between natural parameters and expectations**:

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\eta}) \quad \Longleftrightarrow \quad \boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}).$$

Fact: G is a strictly convex function.[‡] (Proof by generalized Cauchy-Schwarz ineq.)

Corollary: $\boldsymbol{g}$ is a 1-to-1 map.



$\mathcal{M} := \{g(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathcal{N}\}$ = **expectation parameter space**.

[‡]Strict convexity, as opposed to just convexity, actually requires that the exponential family be *minimal*: feature functions should be linearly independent.

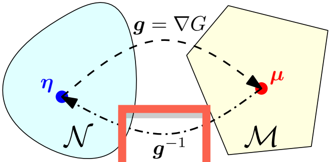
Expectation parameters

$\boldsymbol{g} := \nabla G$ **is an invertible map between natural parameters and expectations**:

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\eta}) \quad \Longleftrightarrow \quad \boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}).$$

Fact: G is a strictly convex function.‡ (Proof by generalized Cauchy-Schwarz ineq.)

Corollary: $\boldsymbol{g}$ is a 1-to-1 map.



$\mathcal{M} := \{g(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathcal{N}\}$ = **expectation parameter space**.

Therefore, exponential families can equivalently be parameterized by the **expectation (mean) parameters** $\boldsymbol{g}(\boldsymbol{\eta}) = \mathbb{E}_{X \sim P_{\boldsymbol{\eta}}}[\boldsymbol{T}(X)]$.

‡Strict convexity, as opposed to just convexity, actually requires that the exponential family be *minimal*: feature functions should be linearly independent.

Special case: Gaussians with unit covariance

Example:
$\mathcal{X} = \mathbb{R}^d$, $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x}$, $\pi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}\|_2^2/2}$.

$$G(\boldsymbol{\eta}) \; = \; \ln\left(\int_{\mathbb{R}^d} \exp\{\langle \boldsymbol{\eta}, \boldsymbol{x} \rangle\} \cdot \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}\|_2^2/2} \, \mathrm{d}\boldsymbol{x} \right) \; = \; \frac{1}{2} \|\boldsymbol{\eta}\|_2^2 \, .$$

Special case: Gaussians with unit covariance

Example:
$\mathcal{X} = \mathbb{R}^d$, $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x}$, $\pi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}\|_2^2/2}$.

$$G(\boldsymbol{\eta}) \;=\; \ln\left(\int_{\mathbb{R}^d} \exp\big\{ \langle \boldsymbol{\eta}, \boldsymbol{x} \rangle \big\} \cdot \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}\|_2^2/2} \, \mathrm{d}\boldsymbol{x} \right) \;=\; \frac{1}{2} \|\boldsymbol{\eta}\|_2^2 \,.$$

In this case,

$$\nabla G(\boldsymbol{\eta}) = \boldsymbol{\eta}.$$

Natural parameters and expectation parameters coincide ($\mathcal{N} = \mathcal{M}$).

Parameter estimation

Let $\mathcal{P} = \{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Parameter estimation

Let $\mathcal{P} = \{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Estimation: Given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, how do we pick $P_{\boldsymbol{\eta}} \in \mathcal{P}$?

Let $\mathcal{P} = \{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Estimation: Given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, how do we pick $P_{\boldsymbol{\eta}} \in \mathcal{P}$?

We'll look at two approaches to estimation:

1. Maximum entropy principle.
2. Maximum likelihood principle.

Parameter estimation

Let $\mathcal{P} = \{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{N}\}$ be the exponential family corresponding to $\boldsymbol{T}$ and π.

Estimation: Given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, how do we pick $P_{\boldsymbol{\eta}} \in \mathcal{P}$?

We'll look at two approaches to estimation:

1. Maximum entropy principle.
2. Maximum likelihood principle.

In both cases, we'll use the fact that

$$\boxed{\mathbb{E}_{X \sim P_{\boldsymbol{\eta}}}[\boldsymbol{T}(X)]} = \boldsymbol{g}(\boldsymbol{\eta}) \quad \forall \boldsymbol{\eta} \in \mathcal{N}$$

where $\boldsymbol{g} \colon \mathcal{N} \to \mathbb{R}^k$ is gradient of log partition function G.

Maximum entropy principle

Maximum entropy principle (for discrete $\mathcal{X}$): given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick the distribution that solves the optimization problem

$$\min_{P \in \Delta(\mathcal{X})} \quad \mathrm{RE}(P \| \pi)$$

$$\text{s.t.} \quad \mathbb{E}_{X \sim P}[\boldsymbol{T}(X)] \;=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)] \left(= \; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{T}(x_i) \right).$$

Maximum entropy principle

Maximum entropy principle (for discrete $\mathcal{X}$): given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick the distribution that solves the optimization problem

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[\boldsymbol{T}(X)] \;=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)] \left(= \; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{T}(x_i) \right). \end{aligned}$$

The solution $P_\star$ has the form $P_\star = P_{\boldsymbol{\eta}_\star} \in \mathcal{P}$, and therefore

$$\boldsymbol{g}(\boldsymbol{\eta}_\star) \;=\; \mathbb{E}_{X \sim P_{\boldsymbol{\eta}_\star}}[\boldsymbol{T}(X)] \;=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)]$$

and

$$\boldsymbol{\eta}_\star \;=\; \boldsymbol{g}^{-1}\left(\widehat{\mathbb{E}}[\boldsymbol{T}(X)]\right).$$

Maximum entropy principle

Maximum entropy principle (for discrete $\mathcal{X}$): given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick the distribution that solves the optimization problem

$$\min_{P \in \Delta(\mathcal{X})} \quad \mathrm{RE}(P \| \pi)$$

$$\text{s.t.} \quad \mathbb{E}_{X \sim P}[\boldsymbol{T}(X)] \; = \; \widehat{\mathbb{E}}[\boldsymbol{T}(X)] \; \left(= \; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{T}(x_i) \right).$$

The solution $P_\star$ has the form $P_\star = P_{\boldsymbol{\eta}_\star} \in \mathcal{P}$, and therefore

it's interesting!

$$\boldsymbol{g}(\boldsymbol{\eta}_\star) \; = \; \mathbb{E}_{X \sim P_{\boldsymbol{\eta}_\star}}[\boldsymbol{T}(X)] \; = \; \widehat{\mathbb{E}}[\boldsymbol{T}(X)]$$

and

$$\boldsymbol{\eta}_\star \; = \; \boldsymbol{g}^{-1}\Big(\widehat{\mathbb{E}}[\boldsymbol{T}(X)]\Big).$$

Pick the distribution $P_{\boldsymbol{\eta}} \in \mathcal{P}$ whose **expectation parameters $\boldsymbol{g}(\boldsymbol{\eta}) = \boldsymbol{\mu}$ are equal to the corresponding empirical expectations**.

Maximum likelihood principle

Maximum likelihood principle: given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick $\boldsymbol{\eta} \in \mathcal{N}$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{\eta}) := \sum_{i=1}^{n} \ln P_{\boldsymbol{\eta}}(x_i)$.

Maximum likelihood principle

Maximum likelihood principle: given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick $\boldsymbol{\eta} \in \mathcal{N}$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{\eta}) := \sum_{i=1}^{n} \ln P_{\boldsymbol{\eta}}(x_i)$.

$$\mathcal{L}(\boldsymbol{\eta}) \;=\; \sum_{i=1}^{n} \Big(\langle \boldsymbol{\eta}, \boldsymbol{T}(x_i) \rangle - G(\boldsymbol{\eta}) + \ln \pi(x_i) \Big)$$

which is a *concave* function of $\boldsymbol{\eta}$.

Maximum likelihood principle

Maximum likelihood principle: given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick $\boldsymbol{\eta} \in \mathcal{N}$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{\eta}) := \sum_{i=1}^{n} \ln P_{\boldsymbol{\eta}}(x_i)$.

$$\mathcal{L}(\boldsymbol{\eta}) \;=\; \sum_{i=1}^{n} \Big(\langle \boldsymbol{\eta}, \boldsymbol{T}(x_i) \rangle - G(\boldsymbol{\eta}) + \ln \pi(x_i) \Big)$$

which is a *concave* function of $\boldsymbol{\eta}$.

Pick $\boldsymbol{\eta}_{\mathsf{ML}} \in \mathcal{N}$ so that gradient of $\mathcal{L}$ at $\boldsymbol{\eta}_{\mathsf{ML}}$ is zero:

$$\nabla \mathcal{L}(\boldsymbol{\eta}_{\mathsf{ML}}) \;=\; \sum_{i=1}^{n} \Big(\boldsymbol{T}(x_i) - \nabla G(\boldsymbol{\eta}_{\mathsf{ML}}) \Big) \;=\; \mathbf{0}$$

Maximum likelihood principle

Maximum likelihood principle: given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick $\boldsymbol{\eta} \in \mathcal{N}$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{\eta}) := \sum_{i=1}^{n} \ln P_{\boldsymbol{\eta}}(x_i)$.

$$\mathcal{L}(\boldsymbol{\eta}) \;=\; \sum_{i=1}^{n} \Big(\langle \boldsymbol{\eta}, \boldsymbol{T}(x_i) \rangle - G(\boldsymbol{\eta}) + \ln \pi(x_i) \Big)$$

which is a *concave* function of $\boldsymbol{\eta}$.

Pick $\boldsymbol{\eta}_{\mathsf{ML}} \in \mathcal{N}$ so that gradient of $\mathcal{L}$ at $\boldsymbol{\eta}_{\mathsf{ML}}$ is zero:

$$\nabla \mathcal{L}(\boldsymbol{\eta}_{\mathsf{ML}}) \;=\; \sum_{i=1}^{n} \Big(\boldsymbol{T}(x_i) - \nabla G(\boldsymbol{\eta}_{\mathsf{ML}}) \Big) \;=\; \mathbf{0},$$

i.e.,

$$\nabla G(\boldsymbol{\eta}_{\mathsf{ML}}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{T}(x_i) \;=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)].$$

Maximum likelihood principle

Maximum likelihood principle: given data $x_1, x_2, \ldots, x_n \in \mathcal{X}$, pick $\boldsymbol{\eta} \in \mathcal{N}$ that maximizes the log-likelihood function $\mathcal{L}(\boldsymbol{\eta}) := \sum_{i=1}^{n} \ln P_{\boldsymbol{\eta}}(x_i)$.

$$\mathcal{L}(\boldsymbol{\eta}) \;=\; \sum_{i=1}^{n} \Big(\langle \boldsymbol{\eta}, \boldsymbol{T}(x_i) \rangle - G(\boldsymbol{\eta}) + \ln \pi(x_i) \Big)$$

which is a *concave* function of $\boldsymbol{\eta}$.

use maximum likelihood method.

Pick $\boldsymbol{\eta}_{\mathsf{ML}} \in \mathcal{N}$ so that gradient of $\mathcal{L}$ at $\boldsymbol{\eta}_{\mathsf{ML}}$ is zero:

$$\nabla \mathcal{L}(\boldsymbol{\eta}_{\mathsf{ML}}) \;=\; \sum_{i=1}^{n} \Big(\boldsymbol{T}(x_i) - \nabla G(\boldsymbol{\eta}_{\mathsf{ML}}) \Big) \;=\; \mathbf{0},$$

i.e.,

$$\nabla G(\boldsymbol{\eta}_{\mathsf{ML}}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{T}(x_i) \;=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)].$$

For exponential families, the maximum likelihood principle is the **same as the maximum entropy principle!**

Empirical expectations

Maximum entropy / maximum likelihood for exponential families is reduced to **computing empirical expectations of feature functions**

$$\boldsymbol{b} \;:=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)] \;=\; \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{T}(x_j),$$

and then applying the inverse of ∇G to obtain the natural parameter estimate.

Empirical expectations

Maximum entropy / maximum likelihood for exponential families is reduced to **computing empirical expectations of feature functions**

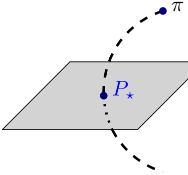
$$\boxed{\boldsymbol{b} \;:=\; \widehat{\mathbb{E}}[\boldsymbol{T}(X)] \;=\; \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{T}(x_j),}$$

reduce to compute feature function

and then applying the inverse of ∇G to obtain the natural parameter estimate.

Unfortunately, except for some simple exponential families (like the well-known distribution families), computing the inverse of ∇G can be difficult.

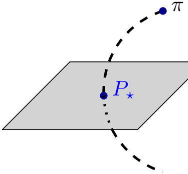
Csiszar's iterative projection algorithm



Let $H_i := \{P : \mathbb{E}_{X\sim P}[T_i(X)] = b_i\}$ for each $i \in [k]$, so

$$H \;:=\; \left\{P : \mathbb{E}_{X\sim P}[\boldsymbol{T}(X)] = \boldsymbol{b}\right\} \;=\; \bigcap_{i=1}^{k} H_i.$$

Csiszar's iterative projection algorithm



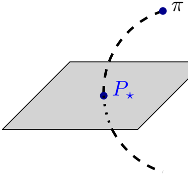
Let $H_i := \{P : \mathbb{E}_{X\sim P}[T_i(X)] = b_i\}$ for each $i \in [k]$, so

$$H := \{P : \mathbb{E}_{X\sim P}[\boldsymbol{T}(X)] = \boldsymbol{b}\} = \bigcap_{i=1}^{k} H_i.$$

Goal: compute **entropy projection** $P_\star$ of π onto H.

Iterative projection algorithm

Csiszar's iterative projection algorithm



Let $H_i := \{P : \mathbb{E}_{X\sim P}[T_i(X)] = b_i\}$ for each $i \in [k]$, so

$$H \;:=\; \{P : \mathbb{E}_{X\sim P}[\boldsymbol{T}(X)] = \boldsymbol{b}\} \;=\; \bigcap_{i=1}^{k} H_i .$$

Goal: compute **entropy projection** $P_\star$ of π onto H.

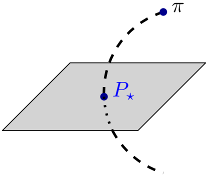
Start with $P^{(0)} := \pi$.
For $t = 1, 2, \ldots$:

$$P^{(t)} := \operatorname*{arg\,min}_{P\in\Delta(\mathcal{X})} \mathrm{RE}(P \| P^{(t-1)}) \;\text{ subject to }\; P \in H_{i_t} .$$

Here, $(i_1, i_2, \ldots)$ is a sequence like $(1, 2, \ldots, k, 1, 2, \ldots, k, \ldots)$.

Iterative projection algorithm

Csiszar's iterative projection algorithm



Let $H_i := \{P : \mathbb{E}_{X\sim P}[T_i(X)] = b_i\}$ for each $i \in [k]$, so

$$H \;:=\; \{P : \mathbb{E}_{X\sim P}[\boldsymbol{T}(X)] = \boldsymbol{b}\} \;=\; \bigcap_{i=1}^{k} H_i .$$

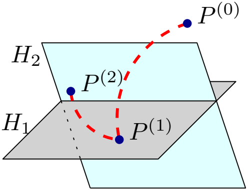
Goal: compute **entropy projection** $P_\star$ of π onto H.

Start with $P^{(0)} := \pi$.
For $t = 1, 2, \dots$:

$$P^{(t)} := \operatorname*{arg\,min}_{P \in \Delta(\mathcal{X})} \mathrm{RE}(P \| P^{(t-1)}) \quad \text{subject to} \quad P \in H_{i_t} .$$

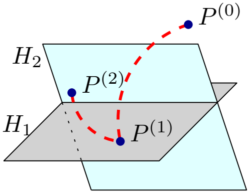
Here, $(i_1, i_2, \dots)$ is a sequence like $(1, 2, \dots, k, 1, 2, \dots, k, \dots)$.

This converges to the entropy projection $P_\star$ of π onto H.



Start with $P^{(0)} := \pi$.
For $t = 1, 2, \ldots$:

$$P^{(t)} := \arg\min_{P \in \Delta(\mathcal{X})} \mathrm{RE}(P \| P^{(t-1)}) \quad \text{s.t.} \quad P \in H_{i_t}.$$



Start with $P^{(0)} := \pi$.
For $t = 1, 2, \ldots$:

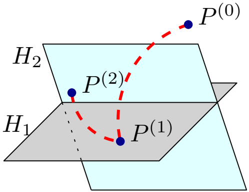
$$P^{(t)} := \arg\min_{P \in \Delta(\mathcal{X})} \mathrm{RE}(P \| P^{(t-1)}) \;\; \text{s.t.} \;\; P \in H_{i_t}.$$

Entropy projection of Q onto H_i: Find η_i such that

$$G_i(\eta) := \ln\left(\sum_{x \in \mathcal{X}} \exp\{\eta T_i(x)\} \cdot Q(x)\right)$$

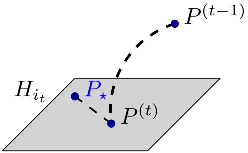
has derivative at η_i equal to b_i. **Can be solved using a line search.**

Iterative projection algorithm



Start with $P^{(0)} := \pi$.
For $t = 1, 2, \ldots$:

$$P^{(t)} := \arg\min_{P \in \Delta(\mathcal{X})} \mathrm{RE}(P \| P^{(t-1)}) \text{ s.t. } P \in H_{i_t}.$$



Convergence: since $P^{(t)}$ is the entropy projection of $P^{(t-1)}$ onto H_{i_t}, and $P_\star \in H_{i_t}$,

$$\mathrm{RE}(P_\star \| P^{(t)}) = \mathrm{RE}(P_\star \| P^{(t-1)}) - \mathrm{RE}(P^{(t)} \| P^{(t-1)}).$$

no negative

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:
  - environmental features of geographic location, e.g.,

$$T_i(x) = \text{temperature at } x \text{ in Celsius;}$$

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:
  - environmental features of geographic location, e.g.,

$$T_i(x) = \text{temperature at } x \text{ in Celsius;}$$

  - products of environmental features;

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:
  - environmental features of geographic location, e.g.,

$$T_i(x) = \text{temperature at } x \text{ in Celsius;}$$

  - products of environmental features;
  - indicator of whether environmental variable exceeds threshold, e.g.,

$$T_i(x) = \mathbb{1}\{\text{temperature at } x \text{ exceeds } 20 \text{ degrees}\}.$$

EXAMPLE: SPECIES DISTRIBUTION MODELING

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:
  - environmental features of geographic location, e.g.,

$$T_i(x) = \text{temperature at } x \text{ in Celsius;}$$

  - products of environmental features;
  - indicator of whether environmental variable exceeds threshold, e.g.,

$$T_i(x) = \mathbb{1}\{\text{temperature at } x \text{ exceeds } 20 \text{ degrees}\}.$$

- Base distribution π = uniform.

Example: species distribution modeling

Phillips, Dudík, and Schapire (2004): model the geographic distribution of particular animal species.

- $\mathcal{X}$ = set of geographic locations.
- Sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$, corresponding to locations where a particular animal species was observed.
- Feature functions $T_1, T_2, \ldots,$ of the form:
  - environmental features of geographic location, e.g.,

$$T_i(x) = \text{temperature at } x \text{ in Celsius;}$$

  - products of environmental features;
  - indicator of whether environmental variable exceeds threshold, e.g.,

$$T_i(x) = \mathbb{1}\{\text{temperature at } x \text{ exceeds } 20 \text{ degrees}\}.$$

- Base distribution $\pi =$ uniform.

Lots of feature functions, but n (number of sightings of an animal) is small.

Example: species distribution modeling

More details:

- Used a variant of maximum entropy where only require

$$\left|\mathbb{E}_{X\sim P}[T_i(X)] - \widehat{\mathbb{E}}[T_i(X)]\right| \;\leq\; \beta$$

for some *bound parameter* $\beta > 0$ (determined using some other method based on the sample size).

Example: species distribution modeling

More details:

- Used a variant of maximum entropy where only require

$$\left|\mathbb{E}_{X\sim P}[T_i(X)] - \widehat{\mathbb{E}}[T_i(X)]\right| \;\leq\; \beta$$

  for some *bound parameter* $\beta > 0$ (determined using some other method based on the sample size).

- Corresponds to *penalized maximum likelihood*: pick $\boldsymbol{\eta} \in \mathcal{N}$ to maximize

$$\mathcal{L}(\boldsymbol{\eta}) - \beta\|\boldsymbol{\eta}\|_1$$

  (similar to Lasso).

  Better able to cope with many feature functions despite small sample size.

Example: species distribution modeling

More details:

- Used a variant of maximum entropy where only require

$$\left|\mathbb{E}_{X\sim P}[T_i(X)] - \widehat{\mathbb{E}}[T_i(X)]\right| \;\leq\; \beta$$

  for some *bound parameter* $\beta > 0$ (determined using some other method based on the sample size).

- Corresponds to *penalized maximum likelihood*: pick $\boldsymbol{\eta} \in \mathcal{N}$ to maximize

$$\mathcal{L}(\boldsymbol{\eta}) - \beta \|\boldsymbol{\eta}\|_1$$

  (similar to Lasso).

  Better able to cope with many feature functions despite small sample size.

- Used resulting $P_{\boldsymbol{\eta}}$ to rank locations by *habitability* for animal species.

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.
- Many commonly used distribution families are exponential families.

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.
- Many commonly used distribution families are exponential families.
- Log partition function (and its gradient) connects the *natural parameters* and the *expectation parameters*.

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.
- Many commonly used distribution families are exponential families.
- Log partition function (and its gradient) connects the *natural parameters* and the *expectation parameters*.
- Maximum entropy and maximum likelihood estimation are the same for exponential families:

  Pick the distribution whose expectation parameters are equal to the corresponding empirical expectations.

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.
- Many commonly used distribution families are exponential families.
- Log partition function (and its gradient) connects the *natural parameters* and the *expectation parameters*.
- Maximum entropy and maximum likelihood estimation are the same for exponential families:

  Pick the distribution whose expectation parameters are equal to the corresponding empirical expectations.
- Many variants of maximum entropy (see, e.g., Phillips, Dudík, and Schapire, 2004).

Recap

- **Exponential families**: naturally arise as solution to maximum entropy optimization problem.
- Many commonly used distribution families are exponential families.
- Log partition function (and its gradient) connects the *natural parameters* and the *expectation parameters*.
- Maximum entropy and maximum likelihood estimation are the same for exponential families:

  Pick the distribution whose expectation parameters are equal to the corresponding empirical expectations.
- Many variants of maximum entropy (see, e.g., Phillips, Dudík, and Schapire, 2004).
- Exponential families also important in / related to other subjects:
  - Generalized linear models (e.g., linear regression, logistic regression).
  - Bayesian inference (due to convenience of *conjugate priors*).
  - Convex optimization (via *Bregman divergences*).
  - ...