Exponential families

Reading group "Network Theory" at LINCS – April 28, 2021

Céline Comte

Eindhoven University of Technology

References

- M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends® in Machine Learning, 2008. Link towards the book.
  - → Chapters 2 "Background" and 3 "Graphical Models as Exponential Families", plus Appendix A "Background Material".

References

- M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends® in Machine Learning, 2008. Link towards the book.
  - → Chapters 2 "Background" and 3 "Graphical Models as Exponential Families", plus Appendix A "Background Material".
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Link towards the book.
  - → Section 3.3 "The conjugate function".

References

- M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends® in Machine Learning, 2008. Link towards the book.
  → Chapters 2 "Background" and 3 "Graphical Models as Exponential Families", plus Appendix A "Background Material".

- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Link towards the book.
  → Section 3.3 "The conjugate function".

- Wikipedia pages Exponential family, Maximum-entropy probability distribution, Lagrange multiplier, Principle of maximum entropy, Convex conjugate.

Outline

Outline

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
  The functions $\phi_1, \phi_2, \ldots, \phi_n$ are called sufficient statistics.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
  The functions $\phi_1, \phi_2, \ldots, \phi_n$ are called sufficient statistics.
- Vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical or exponential parameters.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
  The functions $\phi_1, \phi_2, \ldots, \phi_n$ are called sufficient statistics.
- Vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical or exponential parameters.

The exponential family associated with ϕ is the collection of probability mass functions

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m,$$

parameterized by the vector θ of canonical parameters.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
  The functions $\phi_1, \phi_2, \ldots, \phi_n$ are called sufficient statistics.
- Vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical or exponential parameters.

The exponential family associated with ϕ is the collection of probability mass functions

$$p_\theta(x) = e^{\langle \eta(\theta), \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m,$$

parameterized by the vector θ of canonical parameters.

Exponential families

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Vector-valued function $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
  The functions $\phi_1, \phi_2, \ldots, \phi_n$ are called sufficient statistics.
- Vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical or exponential parameters.

The exponential family associated with ϕ is the collection of probability mass functions

$$p_\theta(x) = h(x) e^{\langle \eta(\theta), \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m,$$

parameterized by the vector θ of canonical parameters.

Exponential families

Exponential families

The quantity $A(\theta)$ is called the log-partition function or cumulant function

Exponential families

The quantity $A(\theta)$ is called the log-partition function or cumulant function, given by

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right).$$

Exponential families

The quantity $A(\theta)$ is called the log-partition function or cumulant function, given by

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right).$$

The domain Ω of the log-partition function A is the set of canonical parameters θ such that $A(\theta)$ is finite, that is

$$\Omega = \{\theta \in \mathbb{R}^n : A(\theta) < +\infty\}.$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Exponential families

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Exponential families

We make the following technical assumptions:

- Regularity: The domain Ω is open.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Exponential families

We make the following technical assumptions:

- Regularity: The domain Ω is open.
- Minimality: There does not exist a nonzero vector $\theta \in \mathbb{R}^n$ such that

$$\langle \theta, \phi(x) \rangle = \sum_{i=1}^{m} \theta_i \phi_i(x)$$

  is a constant.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Exponential families

We make the following technical assumptions:

- Regularity: The domain Ω is open.
- Minimality: There does not exist a nonzero vector $\theta \in \mathbb{R}^n$ such that

$$\langle \theta, \phi(x) \rangle = \sum_{i=1}^{m} \theta_i \phi_i(x)$$

  is a constant. This implies that there is a unique parameter vector θ associated with each distribution in the exponential family.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Log-partition functions vs. generating functions

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Log-partition functions vs. generating functions

Consider the moment-generating function of the sufficient statistics:

$$M(t) = \mathbb{E}_{p_\theta} \left(e^{\langle t, \phi(X) \rangle} \right), \quad t = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n.$$

TU/e

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Log-partition functions vs. generating functions

Consider the moment-generating function of the sufficient statistics:

$$M(t) = \mathbb{E}_{p_\theta} \left(e^{\langle t, \phi(X) \rangle} \right), \quad t = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n.$$

We have $M(t) = e^{A(\theta+t) - A(\theta)}$ for each $t \in \mathbb{R}^n$ such that $\theta + t \in \Omega$.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Log-partition functions vs. generating functions

Consider the moment-generating function of the sufficient statistics:

$$M(t) = \mathbb{E}_{p_\theta}\left(e^{\langle t, \phi(X) \rangle}\right), \quad t = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n.$$

We have $M(t) = e^{A(\theta+t) - A(\theta)}$ for each $t \in \mathbb{R}^n$ such that $\theta + t \in \Omega$. Indeed,

$$M(t) = \sum_{x \in \mathcal{X}^m} e^{\langle t, \phi(x) \rangle} e^{\langle \theta, \phi(x) \rangle - A(\theta)} = \left(\sum_{x \in \mathcal{X}^m} e^{\langle t+\theta, \phi(x) \rangle} \right) e^{-A(\theta)} = e^{A(t+\theta) - A(\theta)}.$$

Outline

Outline

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution
- Beta distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution
- Beta distribution

Discrete univariate distributions

- Geometric distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution
- Beta distribution

Discrete univariate distributions

- Geometric distribution
- Bernoulli distribution

TU/e

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution
- Beta distribution

Discrete univariate distributions

- Geometric distribution
- Bernoulli distribution
- Binomial distribution (with a fixed number of trials)

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Continuous univariate distributions

- Exponential distribution
- Normal distribution
- Beta distribution

Discrete univariate distributions

- Geometric distribution
- Bernoulli distribution
- Binomial distribution (with a fixed number of trials)
- Poisson distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Probabilistic graphical models

TU/e

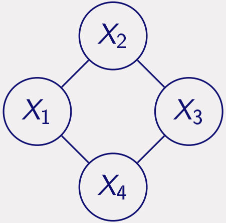
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Probabilistic graphical models

Markov random field

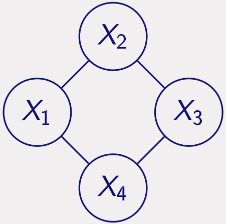


$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

1 – Common distributions

Probabilistic graphical models

Markov random field



Distribution:

$$p(x_1, x_2, x_3, x_4) \propto f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_3, x_4) f_d(x_1, x_4)$$

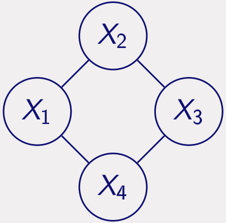
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Probabilistic graphical models

Markov random field



Distribution:

$$p(x_1, x_2, x_3, x_4) \propto f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_3, x_4) f_d(x_1, x_4)$$

$$f_a(x_1, x_2) = e^{(\log f_a(0,0)) 1_{(x_1, x_2) = (0,0)}} \times e^{(\log f_a(0,1)) 1_{(x_1, x_2) = (0,1)}} \times e^{(\log f_a(1,0)) 1_{(x_1, x_2) = (1,0)}} \times e^{(\log f_a(1,1)) 1_{(x_1, x_2) = (1,1)}}$$

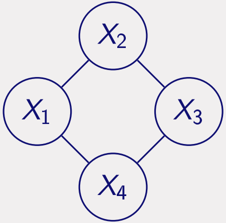
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

1 – Common distributions

Probabilistic graphical models

Markov random field



Distribution:

$$p(x_1, x_2, x_3, x_4) \propto f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_3, x_4) f_d(x_1, x_4)$$

$$f_a(x_1, x_2) = e^{(\log f_a(0,0)) 1_{(x_1, x_2) = (0,0)}} \times e^{(\log f_a(0,1)) 1_{(x_1, x_2) = (0,1)}} \times e^{(\log f_a(1,0)) 1_{(x_1, x_2) = (1,0)}} \times e^{(\log f_a(1,1)) 1_{(x_1, x_2) = (1,1)}}$$

Question: Calculate the normalization constant or marginal distributions.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Limiting distributions of stochastic systems

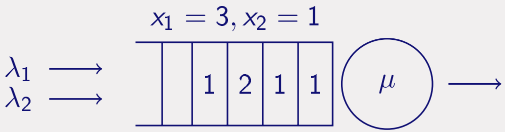
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Limiting distributions of stochastic systems

M/M/1-PS queue with two customer classes



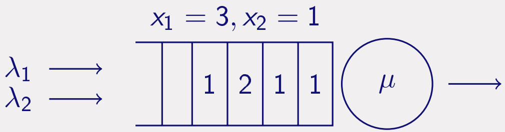
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

1 – Common distributions

Limiting distributions of stochastic systems

M/M/1-PS queue with two customer classes



Stationary distribution:

$$\pi(x) = (1-\rho)\binom{x_1 + x_2}{x_1} {\rho_1}^{x_1} {\rho_2}^{x_2},$$

$$\rho_1 = \frac{\lambda_1}{\mu},\ \rho_2 = \frac{\lambda_2}{\mu},\ \rho = \rho_1 + \rho_2 = \frac{\lambda_1 + \lambda_2}{\mu}.$$

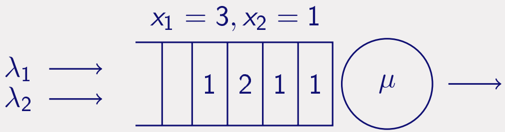
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

1 – Common distributions

Limiting distributions of stochastic systems

M/M/1-PS queue with two customer classes



Stationary distribution:

$$\pi(x) = (1-\rho)\binom{x_1 + x_2}{x_1} {\rho_1}^{x_1} {\rho_2}^{x_2},$$

$$\rho_1 = \frac{\lambda_1}{\mu},\ \rho_2 = \frac{\lambda_2}{\mu},\ \rho = \rho_1 + \rho_2 = \frac{\lambda_1 + \lambda_2}{\mu}.$$

Question: Calculate long-term performance metrics.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Sufficient statistics $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
- Vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$ of mean parameters.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Sufficient statistics $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
- Vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$ of mean parameters.

Moment-matching condition: Find a distribution p on $\mathcal{X}^m$ such that

$$\mathbb{E}_p(\phi(X)) = \mu, \quad \text{that is,} \quad \mathbb{E}_p(\phi_i(X)) = \mu_i, \quad i = 1, 2, \ldots, n.$$

TU/e

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

We introduce:

- Random vector $X = (X_1, X_2, \ldots, X_m)$ taking values in $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$.
- Sufficient statistics $\phi : x \in \mathcal{X}^m \mapsto (\phi_1(x), \ldots, \phi_n(x)) \in \mathbb{R}^n$.
- Vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$ of mean parameters.

Moment-matching condition: Find a distribution p on $\mathcal{X}^m$ such that

$$\mathbb{E}_p(\phi(X)) = \mu, \quad \text{that is,} \quad \mathbb{E}_p(\phi_i(X)) = \mu_i, \quad i = 1, 2, \ldots, n.$$

We let $\mathcal{M}$ denote the set of vectors μ such that such a distribution exists, that is,

$$\mathcal{M} = \{\mu \in \mathbb{R}^n : \exists p \text{ such that } \mathbb{E}_p(\phi(X)) = \mu\}.$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

Principle of maximum entropy: Among all distributions p such that $\mathbb{E}_p(\phi(X)) = \mu$, choose a distribution p that maximizes the Shannon entropy:

$$H(p) = -\sum_{x \in \mathcal{X}^m} (\log p(x)) p(x).$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

2 – Maximum-entropy distribution

Principle of maximum entropy: Among all distributions p such that $\mathbb{E}_p(\phi(X)) = \mu$, choose a distribution p that maximizes the Shannon entropy:

$$H(p) = -\sum_{x \in \mathcal{X}^m} (\log p(x)) p(x).$$

Result: The solution is a member p_θ of the exponential family associated with ϕ, for some vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical parameters:

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m.$$

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle\theta,\phi(x)\rangle}\right)$$

2 – Maximum-entropy distribution

Principle of maximum entropy: Among all distributions p such that $\mathbb{E}_p(\phi(X)) = \mu$, choose a distribution p that maximizes the Shannon entropy:

$$H(p) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x).$$

Result: The solution is a member p_θ of the exponential family associated with ϕ, for some vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$ of canonical parameters:

$$p_\theta(x) = e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m.$$

We now prove this result, and we will explain later how to choose the parameters θ.

Sketch of proof using Lagrange multipliers

Assume $\mathcal{X}^m$ is finite, so that a distribution p is a vector $p = (p(x), x \in \mathcal{X}^m) \in \mathbb{R}_+^{|\mathcal{X}^m|}$.

Sketch of proof using Lagrange multipliers

Assume $\mathcal{X}^m$ is finite, so that a distribution p is a vector $p = (p(x), x \in \mathcal{X}^m) \in \mathbb{R}_+^{|\mathcal{X}^m|}$.

We have to solve the following optimization problem:

$$\text{Maximize}_{p} \qquad H(p) = -\sum_{x \in \mathcal{X}^m} (\log p(x)) p(x),$$

$$\text{Subject to} \qquad \sum_{x \in \mathcal{X}^m} p(x) - 1 = 0 \text{ and } \sum_{x \in \mathcal{X}^m} \phi_i(x) p(x) - \mu_i = 0, \; i = 1, 2, \dots, n.$$

Sketch of proof using Lagrange multipliers

Assume $\mathcal{X}^m$ is finite, so that a distribution p is a vector $p = (p(x), x \in \mathcal{X}^m) \in \mathbb{R}_+^{|\mathcal{X}^m|}$.

We have to solve the following optimization problem:

$$\underset{p}{\text{Maximize}} \qquad H(p) = -\sum_{x \in \mathcal{X}^m} (\log p(x)) p(x),$$

$$\text{Subject to} \qquad \sum_{x \in \mathcal{X}^m} p(x) - 1 = 0 \text{ and } \sum_{x \in \mathcal{X}^m} \phi_i(x) p(x) - \mu_i = 0, \; i = 1, 2, \ldots, n.$$

The Lagrange function associated with this problem is

$$\mathcal{L}(p, \eta, \theta) = -\sum_{x \in \mathcal{X}^m} (\log p(x)) p(x) + \eta \left(\sum_{x \in \mathcal{X}^m} p(x) - 1 \right) + \sum_{i=1}^{n} \theta_i \left(\sum_{x \in \mathcal{X}^m} \phi_i(x) p(x) - \mu_i \right),$$

with $p = (p(x), x \in \mathcal{X}^m) \in \mathbb{R}^{|\mathcal{X}^m|}$, $\eta \in \mathbb{R}$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$.

Sketch of proof using Lagrange multipliers

The Lagrange function associated with this problem is

$$\mathcal{L}(p,\eta,\theta) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x) + \eta\left(\sum_{x\in\mathcal{X}^m} p(x) - 1\right) + \sum_{i=1}^{n} \theta_i \left(\sum_{x\in\mathcal{X}^m} \phi_i(x)p(x) - \mu_i\right).$$

Sketch of proof using Lagrange multipliers

The Lagrange function associated with this problem is

$$\mathcal{L}(p,\eta,\theta) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x) + \eta\left(\sum_{x\in\mathcal{X}^m} p(x) - 1\right) + \sum_{i=1}^{n}\theta_i\left(\sum_{x\in\mathcal{X}^m} \phi_i(x)p(x) - \mu_i\right).$$

We look for the stationary point(s) of this function:

$$0 = \frac{\partial\mathcal{L}}{\partial\eta} = \sum_{x\in\mathcal{X}^m} p(x) - 1,$$

Sketch of proof using Lagrange multipliers

The Lagrange function associated with this problem is

$$\mathcal{L}(p,\eta,\theta) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x) + \eta\left(\sum_{x\in\mathcal{X}^m} p(x) - 1\right) + \sum_{i=1}^{n}\theta_i\left(\sum_{x\in\mathcal{X}^m} \phi_i(x)p(x) - \mu_i\right).$$

We look for the stationary point(s) of this function:

$$0 = \frac{\partial\mathcal{L}}{\partial\eta} = \sum_{x\in\mathcal{X}^m} p(x) - 1, \qquad 0 = \frac{\partial\mathcal{L}}{\partial\theta_i} = \sum_{x\in\mathcal{X}^m} \phi_i(x)p(x) - \mu_i, \quad i = 1,2,\ldots,n,$$

Sketch of proof using Lagrange multipliers

The Lagrange function associated with this problem is

$$\mathcal{L}(p,\eta,\theta) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x) + \eta\left(\sum_{x\in\mathcal{X}^m} p(x) - 1\right) + \sum_{i=1}^{n}\theta_i\left(\sum_{x\in\mathcal{X}^m}\phi_i(x)p(x) - \mu_i\right).$$

We look for the stationary point(s) of this function:

$$0 = \frac{\partial\mathcal{L}}{\partial\eta} = \sum_{x\in\mathcal{X}^m} p(x) - 1, \qquad 0 = \frac{\partial\mathcal{L}}{\partial\theta_i} = \sum_{x\in\mathcal{X}^m}\phi_i(x)p(x) - \mu_i, \quad i = 1,2,\dots,n,$$

$$0 = \frac{\partial\mathcal{L}}{\partial p(x)} = -(1+\log p(x)) + \eta + \sum_{i=1}^{n}\theta_i\phi_i(x),$$

Sketch of proof using Lagrange multipliers

The Lagrange function associated with this problem is

$$\mathcal{L}(p,\eta,\theta) = -\sum_{x\in\mathcal{X}^m} (\log p(x))p(x) + \eta\left(\sum_{x\in\mathcal{X}^m} p(x) - 1\right) + \sum_{i=1}^{n}\theta_i\left(\sum_{x\in\mathcal{X}^m}\phi_i(x)p(x) - \mu_i\right).$$

We look for the stationary point(s) of this function:

$$0 = \frac{\partial\mathcal{L}}{\partial\eta} = \sum_{x\in\mathcal{X}^m} p(x) - 1, \qquad 0 = \frac{\partial\mathcal{L}}{\partial\theta_i} = \sum_{x\in\mathcal{X}^m}\phi_i(x)p(x) - \mu_i, \quad i = 1,2,\ldots,n,$$

$$0 = \frac{\partial\mathcal{L}}{\partial p(x)} = -(1+\log p(x)) + \eta + \sum_{i=1}^{n}\theta_i\phi_i(x), \quad \text{so that } p(x) = e^{-1+\eta}\cdot e^{\langle\theta,\phi(x)\rangle}.$$

Sketch of proof using Lagrange multipliers

What we sweep under the carpet:

- We can verify that such a stationary point is indeed a maximum of the entropy.

Sketch of proof using Lagrange multipliers

What we sweep under the carpet:

- We can verify that such a stationary point is indeed a maximum of the entropy.
- We can show *a priori* that a maximum-entropy distribution has maximum support.

Sketch of proof using Lagrange multipliers

What we sweep under the carpet:

- We can verify that such a stationary point is indeed a maximum of the entropy.
- We can show *a priori* that a maximum-entropy distribution has maximum support.
- The maximum-entropy distribution is unique because the representation is minimal.

Sketch of proof using Lagrange multipliers

What we sweep under the carpet:

- We can verify that such a stationary point is indeed a maximum of the entropy.
- We can show *a priori* that a maximum-entropy distribution has maximum support.
- The maximum-entropy distribution is unique because the representation is minimal.
- The log-partition function $A(\theta)$ may tend to infinity as μ approaches the boundary of $\mathcal{M}$, so this reasoning is valid only when when μ is in the interior of $\mathcal{M}$.

Sketch of proof using Lagrange multipliers

What we sweep under the carpet:

- We can verify that such a stationary point is indeed a maximum of the entropy.
- We can show *a priori* that a maximum-entropy distribution has maximum support.
- The maximum-entropy distribution is unique because the representation is minimal.
- The log-partition function $A(\theta)$ may tend to infinity as μ approaches the boundary of $\mathcal{M}$, so this reasoning is valid only when when μ is in the interior of $\mathcal{M}$.
- The continuous variant of this result is proved with *calculus of variations*.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

3 – Variational inference

Calculating the expectation of the sufficient statistics requires calculating the log-partition function $A(\theta)$.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

3 – Variational inference

Calculating the expectation of the sufficient statistics requires calculating the log-partition function $A(\theta)$.

Calculating the log-partition function $A(\theta)$ is difficult:

- Discrete finite case: Combinatorial explosion.

TU/e

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

3 – Variational inference

Calculating the expectation of the sufficient statistics requires calculating the log-partition function $A(\theta)$.

Calculating the log-partition function $A(\theta)$ is difficult:

- Discrete finite case: Combinatorial explosion.
- Discrete infinite case: Calculate an infinite sum.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

3 – Variational inference

Calculating the expectation of the sufficient statistics requires calculating the log-partition function $A(\theta)$.

Calculating the log-partition function $A(\theta)$ is difficult:

- Discrete finite case: Combinatorial explosion.
- Discrete infinite case: Calculate an infinite sum.
- Continuous case: Calculate a high-dimensional integral.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

3 – Variational inference

Calculating the expectation of the sufficient statistics requires calculating the log-partition function $A(\theta)$.

Calculating the log-partition function $A(\theta)$ is difficult:

- Discrete finite case: Combinatorial explosion.
- Discrete infinite case: Calculate an infinite sum.
- Continuous case: Calculate a high-dimensional integral.

Variational methods will give us a principled way of evaluating or approximating $A(\theta)$. These include sum-product algorithms, the Bethe approximation, and mean-field methods.

3 – Variational inference

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

According to (Wainwright and Jordan, 2008):

> *The general idea is to express a quantity of interest as the solution of an optimization problem. The optimization problem can then be "relaxed" in various ways, either by approximating the function to be optimized or by approximating the set over which the optimization takes place. Such relaxations, in turn, provide a means of approximating the original quantity of interest.*

Outline

Outline

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Convexity

Proposition 3.1:

1. The function A has derivatives of all orders on its domain Ω.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

Convexity

Proposition 3.1:

1. The function A has derivatives of all orders on its domain Ω.
   The first two derivatives yield the mean and covariance of $\phi(X)$:

$$\frac{\partial A}{\partial \theta_i} = \mathbb{E}_{p_\theta} \left(\phi_i(X) \right), \qquad \frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = \mathrm{Cov}_{p_\theta} \left(\phi_i(X), \phi_j(X) \right).$$

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle \theta, \phi(x)\rangle}\right)$$

Convexity

Proposition 3.1:

1. The function A has derivatives of all orders on its domain Ω.
   The first two derivatives yield the mean and covariance of $\phi(X)$:

$$\frac{\partial A}{\partial \theta_i} = \mathbb{E}_{p_\theta}\left(\phi_i(X)\right), \qquad \frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = \mathrm{Cov}_{p_\theta}\left(\phi_i(X), \phi_j(X)\right).$$

   In vector notation, we obtain $\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$ and $\nabla^2 A(\theta) = \mathrm{Cov}_{p_\theta}\left(\phi(X)\right)$.

$$p_\theta(x) = e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$
$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle\theta,\phi(x)\rangle}\right)$$

Convexity

Proposition 3.1:

1. The function A has derivatives of all orders on its domain Ω.
   The first two derivatives yield the mean and covariance of $\phi(X)$:

$$\frac{\partial A}{\partial\theta_i} = \mathbb{E}_{p_\theta}(\phi_i(X)), \qquad \frac{\partial^2 A}{\partial\theta_i\partial\theta_j} = \mathrm{Cov}_{p_\theta}(\phi_i(X), \phi_j(X)).$$

   In vector notation, we obtain $\nabla A(\theta) = \mathbb{E}_{p_\theta}(\phi(X))$ and $\nabla^2 A(\theta) = \mathrm{Cov}_{p_\theta}(\phi(X))$.
2. The function A is strictly convex on its domain Ω.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}}$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}} = \sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x\in\mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x\in\mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}} = \sum_{x\in\mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}$$

$$= \sum_{x\in\mathcal{X}^m} \phi_i(x) p_\theta(x)$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\textstyle\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}} = \sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}$$

$$= \sum_{x \in \mathcal{X}^m} \phi_i(x) p_\theta(x) = \mathbb{E}_{p_\theta} \left(\phi_i(X) \right).$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}} = \sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}$$

$$= \sum_{x \in \mathcal{X}^m} \phi_i(x) p_\theta(x) = \mathbb{E}_{p_\theta} \left(\phi_i(X) \right).$$

The calculation for the second partial derivative is similar.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Sketch of proof

1. For the first partial derivative, we have

$$\frac{\partial A}{\partial \theta_i} = \frac{\sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle}}{\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}} = \sum_{x \in \mathcal{X}^m} \phi_i(x) e^{\langle \theta, \phi(x) \rangle - A(\theta)}$$

$$= \sum_{x \in \mathcal{X}^m} \phi_i(x) p_\theta(x) = \mathbb{E}_{p_\theta} \left(\phi_i(X) \right).$$

The calculation for the second partial derivative is similar.

2. The Hessian matrix $\nabla^2 A(\theta)$ is the covariance matrix of the vector $\phi(X)$ when $X \sim p_\theta$, and a covariance matrix is positive semi-definite. This shows that A is convex. (Strict convexity: minimality of the representation.)

TU/e

Outline

Outline

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{\langle \theta, \mu \rangle - A(\theta)\}$.

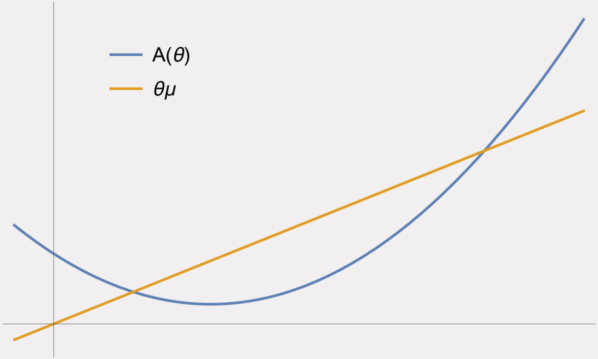
$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{\langle \theta, \mu \rangle - A(\theta)\}$.



$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$.

TU/e

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$.

Theorem 3.4 (Part 1):

1. For each $\mu \in \mathcal{M}^\circ$, the supremum in $A^*(\mu)$ is attained by the vector $\theta \in \Omega$ that satisfies the moment-matching condition

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$.

Theorem 3.4 (Part 1):

1. For each $\mu \in \mathcal{M}^\circ$, the supremum in $A^*(\mu)$ is attained by the vector $\theta \in \Omega$ that satisfies the moment-matching condition, and $A^*(\mu) = -H(p_\theta)$.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} (\phi(X))$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$.

Theorem 3.4 (Part 1):

1. For each $\mu \in \mathcal{M}^\circ$, the supremum in $A^*(\mu)$ is attained by the vector $\theta \in \Omega$ that satisfies the moment-matching condition, and $A^*(\mu) = -H(p_\theta)$.
2. For each $\mu \notin \overline{\mathcal{M}}$, we have $A^*(\mu) = +\infty$.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

Conjugate dual function

For each $\mu \in \mathbb{R}^n$, let $A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$.

Theorem 3.4 (Part 1):

1. For each $\mu \in \mathcal{M}^\circ$, the supremum in $A^*(\mu)$ is attained by the vector $\theta \in \Omega$ that satisfies the moment-matching condition, and $A^*(\mu) = -H(p_\theta)$.
2. For each $\mu \notin \overline{\mathcal{M}}$, we have $A^*(\mu) = +\infty$.
3. For each $\mu \in \overline{\mathcal{M}} \setminus \mathcal{M}^\circ$, we have $A^*(\mu) = \lim_{n \to +\infty} A^*(\mu^n)$ taken over any sequence $(\mu^n)_{n \in \mathbb{N}} \subseteq \mathcal{M}^\circ$ converging to μ.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

$$A^*(\mu) = \sup_{\theta \in \Omega} \left\{ \langle \theta, \mu \rangle - A(\theta) \right\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle \theta, \mu \rangle - A(\theta)$ is strictly concave.

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle \theta, \phi(x)\rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

$$A^*(\mu) = \sup_{\theta\in\Omega}\left\{\langle \theta, \mu\rangle - A(\theta)\right\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle \theta, \mu\rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial \theta_i}(\langle \theta, \mu\rangle - A(\theta)), \quad i = 1, 2, \ldots, n,$$

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle \theta, \phi(x)\rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

$$A^*(\mu) = \sup_{\theta\in\Omega}\{\langle \theta, \mu\rangle - A(\theta)\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle \theta, \mu \rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial \theta_i}(\langle \theta, \mu\rangle - A(\theta)), \quad i = 1, 2, \ldots, n, \quad \text{i.e.,} \quad 0 = \mu_i - \frac{\partial}{\partial \theta_i} A(\theta), \quad i = 1, 2, \ldots, n,$$

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} (\phi(X))$$

$$A^*(\mu) = \sup_{\theta \in \Omega} \{ \langle \theta, \mu \rangle - A(\theta) \}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle \theta, \mu \rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial \theta_i} (\langle \theta, \mu \rangle - A(\theta)), \quad i = 1, 2, \ldots, n, \quad \text{i.e.,} \quad 0 = \mu_i - \frac{\partial}{\partial \theta_i} A(\theta), \quad i = 1, 2, \ldots, n,$$

that is, $\mu = \nabla A(\theta)$.

$$p_\theta(x) = e^{\langle \theta, \phi(x) \rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log \left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x) \rangle} \right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} \left(\phi(X) \right)$$

$$A^*(\mu) = \sup_{\theta \in \Omega} \left\{ \langle \theta, \mu \rangle - A(\theta) \right\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle \theta, \mu \rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial \theta_i} (\langle \theta, \mu \rangle - A(\theta)), \quad i = 1, 2, \ldots, n, \quad \text{i.e.,} \quad 0 = \mu_i - \frac{\partial}{\partial \theta_i} A(\theta), \quad i = 1, 2, \ldots, n,$$

that is, $\mu = \nabla A(\theta)$.

If $\mu \in \mathcal{M}^\circ$, there is a unique $\theta \in \Omega$ that satisfies this moment-matching condition because A is strictly convex

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle\theta,\phi(x)\rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}(\phi(X))$$

$$A^*(\mu) = \sup_{\theta\in\Omega}\{\langle\theta,\mu\rangle - A(\theta)\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle\theta,\mu\rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial\theta_i}(\langle\theta,\mu\rangle - A(\theta)), \quad i = 1,2,\ldots,n, \quad \text{i.e.,} \quad 0 = \mu_i - \frac{\partial}{\partial\theta_i}A(\theta), \quad i = 1,2,\ldots,n,$$

that is, $\mu = \nabla A(\theta)$.

If $\mu \in \mathcal{M}^\circ$, there is a unique $\theta \in \Omega$ that satisfies this moment-matching condition because A is strictly convex, and we have

$$H(p_\theta) = -\sum_{x\in\mathcal{X}^m}(\log p_\theta(x))p_\theta(x)$$

TU/e

$$p_\theta(x) = e^{\langle\theta,\phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x\in\mathcal{X}^m} e^{\langle\theta,\phi(x)\rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}(\phi(X))$$

$$A^*(\mu) = \sup_{\theta\in\Omega}\{\langle\theta,\mu\rangle - A(\theta)\}$$

Sketch of proof

Since the function A is strictly convex, the function $\theta \in \Omega \mapsto \langle\theta,\mu\rangle - A(\theta)$ is strictly concave.

Therefore, $\theta \in \Omega$ is a supremum if and only if

$$0 = \frac{\partial}{\partial\theta_i}(\langle\theta,\mu\rangle - A(\theta)), \quad i = 1,2,\ldots,n, \quad \text{i.e.,} \quad 0 = \mu_i - \frac{\partial}{\partial\theta_i}A(\theta), \quad i = 1,2,\ldots,n,$$

that is, $\mu = \nabla A(\theta)$.

If $\mu \in \mathcal{M}^\circ$, there is a unique $\theta \in \Omega$ that satisfies this moment-matching condition because A is strictly convex, and we have

$$H(p_\theta) = -\sum_{x\in\mathcal{X}^m}(\log p_\theta(x))p_\theta(x) = \langle\theta,\mu\rangle - A(\theta).$$

Variational representation

$$p_\theta(x) = e^{\langle \theta, \phi(x)\rangle - A(\theta)}, \quad x \in \mathcal{X}^m$$

$$A(\theta) = \log\left(\sum_{x \in \mathcal{X}^m} e^{\langle \theta, \phi(x)\rangle}\right)$$

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left(\phi(X)\right)$$

$$A^*(\mu) = \sup_{\theta \in \Omega} \{\langle \theta, \mu\rangle - A(\theta)\}$$

Theorem 3.4 (Part 2):

1. The log-partition function has the following variational representation:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu\rangle - A^*(\mu)\}.$$

2. For each $\theta \in \Omega$, the above supremum is attained uniquely at the vector $\mu \in \mathcal{M}^\circ$ that satisfies the moment-matching condition.

Conclusion

- Exponential families are parametric sets of probability distributions that appear in many applications.

TU/e

Conclusion

- Exponential families are parametric sets of probability distributions that appear in many applications.
- Many classical distributions can be seen as maximum-entropy distributions under a given moment-matching condition.

Conclusion

- Exponential families are parametric sets of probability distributions that appear in many applications.
- Many classical distributions can be seen as maximum-entropy distributions under a given moment-matching condition.
- The (log-)partition function and the expectation of the sufficient statistics are hard to calculate in general, but for exponential families, they can be approximated using variational inference.