HW 2: due tomorrow
No quiz → next week (exponential families)

COMPSCI 688: Probabilistic Graphical Models

Lecture 11: Continuous Distributions and Exponential Families

Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Big Picture

The Big Picture

Summary of course so far

- compact **representations** of high-dimensional distributions
  - Bayes nets, MRFs, CRFs
  - conditional independence, graph structure, factorization
- **inference**
  - conditioning, marginalization
  - variable elimination, message passing
    - trees only, efficient!
    - extensions: junction tree, loopy belief propagation
- **learning**
  - Bayes nets: counting
  - MRFs/CRFs: numerical optimization of log-likelihood, inference is key subroutine

$\frac{\partial}{\partial \theta} Z(\theta) =$ marginals

What's left?

- Inference (and therefore learning) not tractable for many models
  → approximate inference
    - MCMC
    - variational
- Other types of probability distributions (**continuous**, parametric, ...)
  - statistical models, ML models

Today

- A bit of probability: continuous distributions, expectations
- Exponential families: very general class of distributions
  - includes MRFs
  - "redo" learning in much more general way

Continuous Distributions

Continuous Random Variables and Density Functions

How to define the distribution of a random variable $X \in \mathbb{R}^d$?

The random variable $X \in \Omega$ has **density function** $p : \Omega \to \mathbb{R}^+$ if

$$P(X \in A) = \int_A p(x)dx$$

$$P(X \in A) = \sum_{w \in A} p(w)$$

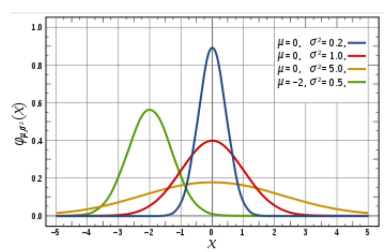
Implies $p(x) \geq 0$, $\int_\Omega p(x) = 1$.

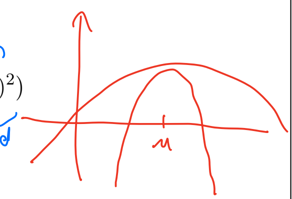
Note: a pmf is a density function (integral over finite set $\equiv$ sum)

Example: Normal Distribution

The univariate normal (or Gaussian) distribution is the most well known continuous distribution. It has density

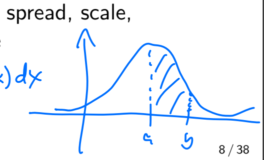
$$p(x;\mu,\sigma^2) = \overset{\frac{1}{Z}}{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}} \underbrace{\exp\big(-\overbrace{\frac{1}{2\sigma^2}(x-\mu)^2}^{\text{log-density}}\big)}_{\text{unnormalized prob}}$$



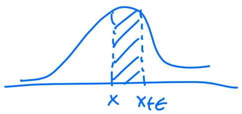


- $\mu \in \mathbb{R}$: location, mean, mode
- $\sigma^2 \geq 0$: spread, scale, variance

$P(a \leq X \leq b) = \int_a^b p(x)dx$



How to Think About a Density



A density is "like" a probability. For $X \in \mathbb{R}$ with density $p(x)$

$$P(X \in [x, x+\epsilon]) = \int_x^{x+\epsilon} p(x)dx \approx \epsilon p(x)$$

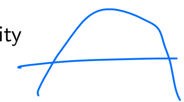
$$p(x) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} P(X \in [x, x+\epsilon])$$

The density can be though of as the probability of X landing in a tiny interval around x (divided the width of the interval).

The standard rules of probability (conditioning, marginalization) usually translate to densities in a straightforward way.

Example: Multivariate Normal Distribution

A multivariate normal (or Gaussian) random variable $\mathbf{X} \in \mathbb{R}^n$ has density



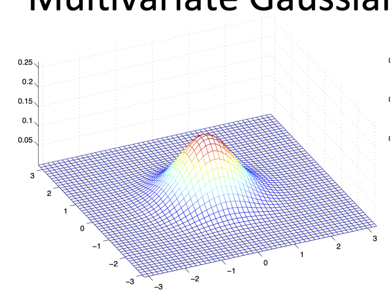
$$p(\mathbf{x}; \mu, \Sigma) = \cancel{\frac{1}{|2\pi\Sigma|^{1/2}}} \underbrace{\exp\big(\overbrace{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}^{\text{neg. quadratic}}\big)}_{\text{unnormalized prob}}$$

- $\mu \in \mathbb{R}^n$: mean, mode
- $\Sigma \in \mathbb{R}^{n \times n}$: covariance matrix, defines scale and orientation
  - Must be positive definite (PSD): $\mathbf{x}^\top \Sigma \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$. (Equivalently, all eigenvalues positive). PD

Visualization

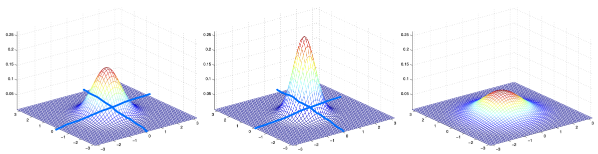
Sequence of examples due to Andrew Ng / Stanford

Multivariate Gaussian



$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

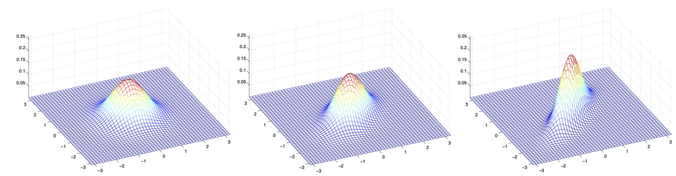
Examples: Symmetric



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = 0.6I; \qquad \Sigma = 2I.$$

$$\Sigma = I$$

Examples: Non-Symmetric



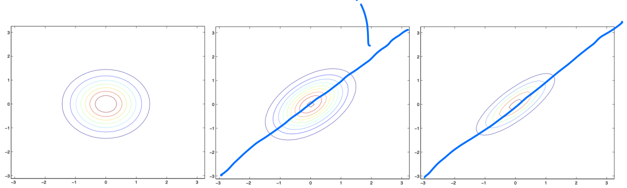
$Cov(X_1, X_2) = 0.5$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \ \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \ .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

$\Sigma_{ij} = Cov(X_i, X_j)$

Contours

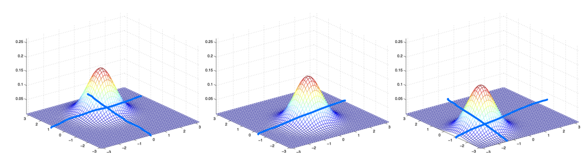
$X_1 = X_2$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \ \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \ .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

Mean

- Change mu: move mean of density around



$$\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \ \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}; \ \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}.$$

Marginal and Conditional Densities

- Definitions from pmfs usually translate to densities
- Suppose $p(\mathbf{x}, \mathbf{y})$ is a density for $(\mathbf{X}, \mathbf{Y})$. The marginal and conditional densities are

$$p(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x}$$

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{x}, \mathbf{y})}{\int p(\mathbf{x}, \mathbf{y}) d\mathbf{x}}$$

Expectations

Expectations

Given a random variable $\mathbf{X}$ with pmf or density $p(\mathbf{x})$ and a function $f(X)$, the expected value $\mathbb{E}[f(\mathbf{X})]$ is

$$\mathbb{E}[f(\mathbf{X})] = \sum_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}) \quad \text{discrete}$$

$$\mathbb{E}[f(\mathbf{X})] = \int p(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad \text{continuous}$$

The sum/integral is over all possible values of $\mathbf{x}$.

We often write this as $\boxed{\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{X})]}$ to make the distribution clear.

Mean and Variance

$f(x) = x$

The moments of a distribution are expectations of polynomials, e.g. $f(x) = (x - c)^d$ for scalars.

The mean is

$$\mu \leftarrow \mathbb{E}[\mathbf{X}] = \int p(\mathbf{x}) \mathbf{x} \, d\mathbf{x}$$

$\Sigma_{ij} = \mathbb{E}[z_i z_j]$
$= \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$
$:= \mathrm{Cov}(X_i, X_j)$

Let $\mu = \mathbb{E}[X]$. The *variance* is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] \qquad X \text{ scalar}$$

$$\mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top] \quad X \text{ vector}$$

covariance matrix

$$[X] - [\mu] = [z]$$

$$[z]\,[z^\top] = \begin{bmatrix} z_1 z_1 & \cdots & z_1 z_n \\ z_2 z_1 & & \\ z_n z_1 & & \end{bmatrix} = \Sigma$$

Marginal and conditional means use marginal and conditional densities:

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{p(\mathbf{y})}[\mathbf{Y}] \qquad \text{marginal}$$
$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\mathbf{X}|\mathbf{Y}=\mathbf{y}] = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[\mathbf{X}] \quad \text{conditional}$$

In the vector case, $\mathrm{Var}(\mathbf{X})$ is the *covariance matrix*.

Linearity of Expectation

For $X, a, b \in \mathbb{R}$:

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$$

For vectors $\mathbf{X}$ and b and matrix A

$$\mathbb{E}[A\mathbf{X} + b] = A\,\mathbb{E}[\mathbf{X}] + b$$

Proof: write out expectation, use linearity of sum/integral

Variance is Positive (Semi-Definite)

A covariance matrix $\mathrm{Var}(\mathbf{X})$ is always positive semi-definite.

Proof (scalar): $\mathbb{E}[(X-\mu)^2] \geq 0$ because the integrand is non-negative

Proof (vector): let $\mathbf{z}$ be any vector and $\mu = \mathbb{E}[\mathbf{X}]$. Then

$$z^T \mathrm{Var}(X) z = z^T \mathbb{E}\left[(X-\mu)(X-\mu)^T\right] z$$
$$= \mathbb{E}\left[z^T (X-\mu)(X-\mu)^T z\right] \qquad v = (X-\mu)^T z$$
$$= \mathbb{E}\left[v^T v\right]$$
$$= \mathbb{E}\left[\|v\|^2\right]$$
$$\geq 0$$

Significance

Expectations are important, but can be hard to compute!

Example: suppose $p(\mathbf{x})$ is an MRF. A marginal is an expectation:

$$P(X_u = a, X_v = b) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{I}[X_u = a, X_v = b]\right]$$

Inference = computing expectations = hard in general

We will come back to approximating expectations and approximate inference

Exponential Families

Exponential Families

An exponential family defines a set of distributions with densities of the form

$$p_\theta(x) = h(x)\exp(\theta^\top T(x) - A(\theta))$$

- θ: "(natural) parameters"
- $T(x)$: "sufficient statistics"
- $A(\theta)$: "log-partition function"
- $h(x)$: "base measure" (we'll usually ignore)

Interpretation ($h(x) = 1$)

$$p_\theta(x) = \exp(\theta^\top T(x) - A(\theta))$$

- $\theta^\top T(x)$ is a real-valued "score" (positive or negative), defined in terms of "features" $T(x)$ and parameters θ
- $\exp(\theta^\top T(x))$ is an unnormalized probability
- The log-partition $A(\theta) = \log Z(\theta)$ function ensures normalization

$$p_\theta(x) = \frac{\exp(\theta^\top T(x))}{\exp(A(\theta))}, \quad A(\theta) = \log Z(\theta) = \log \int \exp(\theta^\top T(x))dx$$

- Valid parameters are the ones for which $A(\theta)$ is finite.

Applications and Importance

- We can get *many* different families of distributions by selecting different "features" $T(x)$ for a variable x in some sample space:
  - Bernoulli, Binomial, Multinomial, Beta, Gaussian, Poisson, MRFs, ...
- There is a general theory that covers learning and other properties of all of these distributions!
- A good trick to seeing that a distribution belongs to an exponential family is to match its log-density to

$$\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta)$$

Preview: Graphical Models

For some intuition why exponential families could be relevant for graphical models, observe that the unnormalized probability factors over "simpler" functions, just like graphical models:

$$\exp(\theta^\top T(x)) = \exp \sum_i \theta_i T_i(x) = \prod_i \exp(\theta_i T_i(x))$$

(Think: what could $T(x)$ look like to recover a graphical model?)

Example: Bernoulli Distribution

The Bernoulli distribution with parameter $\mu \in [0, 1]$ has density (pmf)

$$p_\mu(x) = \begin{cases} \mu & x = 1 \\ 1 - \mu & x = 0 \end{cases}$$

One way to write the log-density is

$$\log p_\mu(x) = \mathbb{I}[x = 1] \log \mu + \mathbb{I}[x = 0] \log(1 - \mu)$$

To match this to an exponential family

$$\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta),$$

Review: Bernoulli Distribution

To match this to an exponential family $\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta)$, take

- $h(x) = 1$
- $T(x) = (\mathbb{I}[x = 1], \mathbb{I}[x = 0])$
- $\theta = (\log \mu, \log(1 - \mu))$
- $\exp(\theta^\top T(x)) = \begin{cases} e^{\theta_1} & x = 1 \\ e^{\theta_2} & x = 0 \end{cases}$
- $A(\theta) = \log(e^{\theta_1} + e^{\theta_2})$
- It's easy to check that $A(\theta) = 0$ when $\theta = (\log \mu, \log(1 - \mu))$

Example: Bernoulli, Single Parameter

We can also write the Bernoulli as a single-parameter exponential family. Rewrite the log-density as

$$\log p_\mu(x) = \log(1-\mu) + x \log \frac{\mu}{1-\mu}$$

Review: Bernoulli, Single Parameter

- $h(x) = 1$
- $T(x) = \mathbb{I}[x = 1] = x$
- $\theta = \log \frac{\mu}{1-\mu}$
- $\exp(\theta^\top x) = \begin{cases} e^\theta & x = 1 \\ 1 & x = 0 \end{cases}$
- $A(\theta) = \log(1 + e^\theta)$
- It's easy to check that $\log(1 + e^\theta) = -\log(1-\mu)$ when $\theta = \log \frac{\mu}{1-\mu}$

Example: Normal Distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2\right)$$

Review: Normal Distribution

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\right) \end{aligned}$$

$$\log p_{\mu,\sigma^2}(x) = x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})$$

- $h(x) = 1$
- $T(x) = (x^2, x)$
- $\theta = (\frac{-1}{2\sigma^2}, \frac{\mu}{\sigma^2})$
- $A(\theta) = \log \int \exp(x^2\theta_1 + x\theta_2)dx = \ldots = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})$

Note: we need $\theta_1 < 0$; why?

Pairwise Markov Random Field

Will revisit later...

Next Time

- graphical models are exponential families
- derive important properties of exponential families
- general treatment of maximum likelihood learning in exponential families