# Probability Reference Sheet

By Danny Liu

## 1 Basic Knowledge

Let *E* be an event and *S* the sample space, then  $P(E) \ge 0$ , P(S) = 1,  $P(E) = 1 - P(\overline{E})$  If  $A \perp B$ , then  $P(A \cap B) = P(A)P(B)$  and P(A|B) = P(A), If  $A \cap B = \emptyset$ , then  $P(A \cap B) = 0$   $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$  and  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  If  $S = B_1 \cup ... \cup B_n$  and  $A \subseteq S$ , then  $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$   $P(x_1, ..., x_n) = P(x_n | x_{n-1}, ..., x_1)P(x_{n-1} | x_{n-2}, ..., x_1) ... P(x_1)$ 

## 2 Bayesian Inference

 $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$ , where  $p(D|\theta)$  is the likelihood,  $p(\theta)$  is the prior and p(D) is the marginal

## 2.1 Conjugate Priors

A family F of priors  $p(\theta)$  is conjugate to the likelihood  $p(D|\theta)$  if the posterior  $p(\theta|D)$  is in F Beta $(\alpha, \beta)$  is conjugate to Bin(n, p), Geo(p), Bern $(\phi)$ . Dir $(\alpha)$  is conjugate to Cat(p), Mult(n, p)  $\mathcal{N}(\mu, \sigma^2)$  is self-conjugate. Gamma $(\alpha, \beta)$  is conjugate to Exp $(\lambda)$ , Poi $(\lambda)$ ,  $\mathcal{N}(\mu, \sigma^2)$ 

### 2.2 Maximum Likelihood Estimate

Find  $\theta$  that maximizes the likelihood,  $\theta^* = \operatorname*{argmax}_{\theta} p(D|\theta)$ . Assuming that D is iid, then The conditional likelihood of y|x is  $L(\theta) = \prod_{i=1}^n p(y^{(i)}|x^{(i)};\theta)$ 

## 2.3 Maximum a Posteriori

Given prior  $p(\theta)$ , find  $\theta$  that maximizes the posterior,  $\theta^* = \underset{\theta}{\operatorname{argmax}} p(\theta|D) = \underset{\theta}{\operatorname{argmax}} p(D|\theta)p(\theta)$ When  $n \to \infty$  or  $p(\theta)$  is uniform, then MAP = MLE as  $p(\theta|D) \propto p(D|\theta)$ 

# 3 Expectation and Covariance

$$\begin{split} E[X] &= \int_x x f(x) dx, \, \text{var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2 \\ E[g(X)] &= \int_x g(x) f(x) dx, \, E[aX + b] = a E[X] + b, \, \text{var}(aX + b) = a^2 \text{var}(X) \\ E[\sum_i X_i] &= \sum_i E[X_i], \, \text{var}(\sum_i X_i) = \sum_i \text{var}(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j) \\ \text{cov}(X,Y) &= \langle X,Y \rangle = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \\ \text{Define correlation of } X \text{ and } Y \text{ as } \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\langle X,Y \rangle}{\|X\| \|Y\|} = \cos\theta \\ \rho(X,Y) &= \pm 1 \iff Y = aX + b \text{ for some } a,b \in \mathbb{F} \end{split}$$

#### 3.1 Covariance Matrix

 $\Sigma = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] = E[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T, \text{ where } \Sigma_{ij} = \text{cov}(X_i, X_j)$ \Sigma is symmetric positive semi-definite and \text{cov}(Ax + b) = A\Sigma A^T Suppose  $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$  is jointly Gaussian with parameters  $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$  and  $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ , then

The marginals are  $p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11})$  and  $p(x_2) = \mathcal{N}(x_2|\mu_2, \Sigma_{22})$  with posterior conditional  $p(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2})$ , where  $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ ,  $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ 

## 3.2 Conditional Expectation

The conditional distribution of X given Y is  $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{P[X=x\cap Y=y]}{P[Y=y]}$ The conditional expectation of X given Y is  $E[X|Y=y] = \sum_x x p_{X|Y}(x|y)$  $E[E[X|Y]] = \sum_y E[X|Y=y] p_Y(y) = E[X]$ 

## 4 Inequalities and Limit Theorems

If X only takes on non-negative values, then for any a>0,  $P[X\geq a]\leq \frac{E[X]}{a}$  (Markov) If  $E[X]=\mu$  and  $\mathrm{var}(X)=\sigma^2$ , then for any k>0,  $P[|X-\mu|\geq k\sigma]\leq \frac{1}{k^2}$  (Chebyshev) If f is a convex function, then  $E[fX]\geq f(EX)$  (Jensen)

#### 4.1 Limit Theorems

Let  $X_1, X_2, ..., X_n$  be iid with  $E[X] = \mu$  and  $var(X) = \sigma^2 < \infty$ , then for n sufficiently large

- 1.  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx Z \sim \mathcal{N}(0,1)$ , where  $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$  has mean  $\mu$  and variance  $\frac{\sigma^2}{n}$  (CLT)
- 2.  $P\left[\lim_{n\to\infty} \frac{X_1+X_2+...+X_n}{n} = \mu\right] = 1$  (Law of Large Numbers)

# 5 Moment Generating Functions

$$\begin{split} M_X(t) &= E[e^{tX}] = \int_x e^{tx} f(x) dx, \text{ where } E[X], E[X^2], E[X^3], \dots \text{ are the moments of } X \\ \text{Consider Taylor expansion of } e^{tX} &= 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \dots \\ \text{If two r.vs have the same mgf for all } t \in \mathbb{R}, \text{ then they have the same distribution for all } t \in \mathbb{R}. \\ M_X(t) &= \sum_{k=0}^\infty E[x^k] \frac{t^k}{k!} \text{ and } E[X^k] = \frac{d^k}{dt^k} M_X(t) \big|_{t=0} \text{ for any integer } k > 0 \\ \text{If } X, Y \text{ are r.vs and } Y = aX + b, \text{ then } M_Y(t) = e^{bt} M_X(at) \\ \text{If } X_1, X_2, \dots, X_n \text{ are independent r.vs, then } M_{X_1 + X_2 + \dots + X_n}(t) = M_{X_1}(t) M_{X_2}(t) \dots M_{X_n}(t) \end{split}$$

## 6 Random Variables

A random variable is a function  $X:S\to\mathbb{R}$  where S is the sample space. For discrete X,Y, the joint pmf is  $p_{X,Y}(x,y)=P[X=x\cap Y=y]$ , where  $\sum_x\sum_y p_{X,Y}(x,y)=1$  For continuous X,Y, the joint pdf is  $\int\int_{(x,y)\in R}f(x,y)dxdy=P[(x,y)\in R]$  Note that although  $\int_a^bf(x)dx=P[a\leq X\leq b]$  for continuous X,P[X=a]=0 If  $F(a)=P[X\leq a]=\int_{-\infty}^af(x)dx$  is the cdf, then  $f(x)=\frac{d}{dx}F(x)$  Continuous (discrete) r.vs are independent if and only if  $f_{X,Y}(x,y)=f_X(x)f_Y(y)$  The marginal density of X is  $f_X(x)=\int_y f(x,y)dy$ 

## 6.1 Functions of Random Variables

Suppose X is a random variable and Y = g(X). If g is increasing and differentiable, then

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y) & \text{if } g(a) \le y \le g(b) \\ 0 & \text{otherwise} \end{cases}$$

## 6.2 Continuity Correction

If a discrete distribution (e.g Binomial, Poisson) is approximated by a continuous distribution (e.g normal), then the following adjustments must be made

x = n becomes n - 0.5 < x < n + 0.5

x > n becomes x > n + 0.5 and x < n becomes x < n - 0.5

## 7 Common Distributions

#### 7.1 Bernoulli Distribution

$$X \sim \text{Bernoulli}(\phi), \ p(x) = \phi^x (1 - \phi)^x, \ M_X(t) = (1 - \phi) + \phi e^t, \ E[X] = \phi, \ \text{var}(X) = \phi (1 - \phi)$$

## 7.2 Categorial Distribution

 $X \sim \text{Cat}(\boldsymbol{p}), p(\boldsymbol{x}) = \sum_{i=1}^{k} p_i^{x_i}$  is a generalization of the Bernoulli/special case of multinomial

### 7.3 Binomial Distribution

 $X \sim \text{Bin}(n,p)$  models the number of n Bernoulli trials that end up being successes  $p(x) = p^x (1-p)^{n-x} \binom{n}{x}$ ,  $M_X(t) = (pe^t + 1 - p)^n$ , E[X] = np, var(X) = np(1-p)

### 7.4 Multinomial Distribution

 $X \sim \text{Multinomial}(n, \mathbf{p}) \text{ models the outcome of } n \text{ categorial trials}$ 

$$p(\mathbf{x}) = \binom{n}{x_1, \dots, x_k} \prod_{i=1}^k p_i^{x_i}, M_X(\mathbf{t}) = (\sum_{i=1}^k p_i e^{t_i})^n, \text{ where } \sum_i p_i = 1 \text{ and } \sum_i x_i = n$$

#### 7.5 Geometric Distribution

 $X \sim \text{Geo}(p)$  models the number of Bernoulli trials needed to get one success  $p(x) = (1-p)^{x-1}p, \ M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \ E[X] = \frac{1}{p}, \ \text{var}(X) = \frac{1-p}{p^2}$ 

$$P(w) \qquad (1-p) \qquad P(1-p)e^{t} \qquad (1-p)e^{t} \qquad p(1-p)e^{t}$$

## 7.6 Poisson Distribution

 $X \sim \operatorname{Poi}(\lambda)$  models the number of times a given event occurs independently in a fixed interval  $p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$ ,  $M_X(t) = e^{\lambda(e^t - 1)}$ ,  $E[X] = \operatorname{var}(X) = \lambda = \operatorname{rate}$  at which event occurs Approximates binomial well when n is large, p is small and  $\lambda = np$  is moderate

#### 7.7 Uniform Distribution

$$X \sim U(a,b), f(x) = \frac{1}{b-a}$$
 for  $a \le x \le b, M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}, E[X] = \frac{1}{2}(a+b), var(X) = \frac{(b-a)^2}{12}$ 

### 7.8 Exponential Distribution

 $X \sim \operatorname{Exp}(\lambda)$  models the time between events in a Poisson process with rate  $\lambda$   $f(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$ ,  $M_X(t) = \frac{\lambda}{\lambda - t}$ ,  $E[X] = \frac{1}{\lambda}$ ,  $\operatorname{var}(X) = \frac{1}{\lambda^2}$  Key property memoryless and continuous analogue of the geometric distribution

#### 7.9 Gamma Distribution

 $X \sim \text{Gamma}(\alpha, \beta)$  models the amount of time until  $\alpha$  events occur in a Poisson process with rate  $\beta$   $f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ ,  $M_X(t) = (\frac{\beta}{\beta - t})^{\alpha}$ ,  $E[X] = \frac{\alpha}{\beta}$ ,  $var(X) = \frac{\alpha}{\beta^2}$ 

## 7.10 Gaussian Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2), \ f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x-\mu)^2}{2\sigma^2}), \ M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}, \ E[X] = \mu, \ \text{var}(X) = \sigma^2$$
  
 $P(a \leq X \leq b) = P(Z \leq \frac{b-\mu}{\sigma}) - P(Z \leq \frac{a-\mu}{\sigma}), \ \text{where} \ Z \sim \mathcal{N}(0, 1)$   
Approximates binomial well (with continuity correction) when  $np$  and  $n(1-p) > 5$ 

#### 7.11 Multivariate Gaussian

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
, where  $f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$   
The diagonal entries of  $\boldsymbol{\Sigma}$  are the variances and stretch the density in the *i*th direction(-ish)  $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ji}$  for  $i \neq j$  are the covariances and linearly correlate the *i*th and *j*th directions If the joint  $f(\boldsymbol{x})$  is radially symmetric, then  $\boldsymbol{\Sigma}_{ij} = 0$  for  $i \neq j$  and  $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\Sigma}_{jj}$  for  $i = j$ 

## 7.12 Laplace Distribution

$$X \sim \text{Laplace}(\mu, b), f(x) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b}), M_X(t) = \frac{\exp(\mu t)}{1-b^2t^2}, E[X] = \mu, \text{var}(X) = 2b^2$$
  
If  $X, Y \sim \text{Exp}(\lambda)$ , then  $X - Y \sim \text{Lap}(0, \lambda^{-1})$ . If  $X \sim \text{Lap}(\mu, b)$ , then  $kX + c \sim \text{Lap}(k\mu + c, kb)$   
Laplace has a fatter tail and taller head than the Gaussian due to abs. value and constant term

### 7.13 Beta Distribution

 $X \sim \text{Beta}(\alpha, \beta)$  represents a distribution of probabilities of an uncertain binomial variable  $f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$ , where  $x \in [0, 1]$ ,  $\alpha, \beta > 0$  and  $E[X] = \frac{\alpha}{\alpha+\beta}$ Note  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$  is the beta function and  $\Gamma(n) = (n-1)!$  is the gamma function Well suited to represent a prior, as the posterior after n Bernoulli trials is  $\text{Beta}(\alpha + \text{success}, \beta + \text{fail})$ 

#### 7.14 Dirichlet Distribution

 $X \sim \mathrm{Dir}(\boldsymbol{\alpha})$  represents a distribution of probabilities of an uncertain multinomial variable  $f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1}$ , where  $x_i \in [0,1], \sum_i x_i = 1, \ \alpha_i > 0$  and  $E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ 

### 8 Calculus

Let 
$$f$$
 be a continuous real-valued function on  $[a,b]$  and  $F$  be the antiderivative of  $f$ . Then  $\int_a^b f(t)dt = F(b) - F(a)$  and  $\frac{d}{dx} \int_a^x f(t)dt = f(x)$ . 
$$\int u dv = uv - \int v du, \int_s^\infty \lambda e^{-\lambda x} dx = e^{-\lambda s}, \int x e^{-\lambda x} dx = \frac{1}{\lambda^2} (-\lambda e^{-\lambda x} x - e^{-\lambda x}) + C$$
. 
$$\int \frac{1}{1+x^2} dx = \arctan(x) + C, \int \frac{x^2}{1+x^2} dx = \int 1 - \frac{1}{1+x^2} dx, \int \frac{x^3}{1+x^2} dx = \int x - \frac{x}{1+x^2} dx$$