Review multivariate normal

- Marginalization of a normal distribution is still a normal distribution
- Conditioning of normal distribution:
  $\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{Y}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$
- Product of normal distribution:
  $\mathcal{N}(\mathbf{y}_1; \mathbf{x}, \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{y}_2; \mathbf{x}, \Sigma_{\mathbf{Y}_2}) =$
  $\mathcal{N}(\mathbf{y}_1; \mathbf{y}_2, \Sigma_{\mathbf{Y}_2} + \Sigma_{\mathbf{Y}_1})\mathcal{N}(\mathbf{x}; (\Lambda_{\mathbf{Y}_1} + \Lambda_{\mathbf{Y}_2})^{-1}(\Lambda_{\mathbf{Y}_2}\mathbf{y}_2 + \Lambda_{\mathbf{Y}_1}y), (\Lambda_{\mathbf{Y}_2} + \Lambda_{\mathbf{Y}_1})^{-1})$
- Mixture of Gaussian
  - Merge components:

$$w \leftarrow \sum_i w_i, \qquad \hat{w}_i = \frac{w_i}{\sum_j w_j}, \qquad \boldsymbol{\mu}_i \leftarrow \sum_i w_i \boldsymbol{\mu}_i,$$

$$\Sigma \leftarrow \sum_{i=1}^n \hat{w}_i(\Sigma_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) - \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i \hat{w}_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j$$

  - Similarity measure

$$Sim(\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p), \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)) = \frac{\mathcal{N}(\boldsymbol{\mu}_p; \boldsymbol{\mu}_q, \Sigma_p + \Sigma_q)}{\sqrt{\mathcal{N}(0; 0, 2\Sigma_p)\mathcal{N}(0; 0, 2\Sigma_q)}}$$

More from last week...

- Bernoulli pdf: $Bern(x|p) = p^x(1-p)^{1-x}$
- Binomial pdf: $Bin(x|p, N) \propto p^x(1-p)^{N-x}$
- Beta pdf: $Beta(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
- Gamma function $\Gamma(z)$
  - $\Gamma(z) = (z-1)\Gamma(z-1)$
  - $\Gamma(n) = (n-1)!$ if n is an integer ≥ 1
- Conjugate prior: a prior with same "form" as its posterior distribution
  - Beta distribution is conjugate prior of Bernoulli and binomial distributions

Summary of Beta distribution

- Pdf:

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

with $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

- Mean:

$$\frac{a}{a+b}$$

- Variance:

$$\frac{ab}{(a+b)^2(a+b+1)}$$

- Mode:

$$\frac{a-1}{a+b-2}$$

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x.

[1] Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$p(p|x, a, b)$$

[1] Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$p(p|x, a, b) = Const1 \cdot Beta(p|a, b) Bern(x|p)$$

[1]Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$\begin{aligned} p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bern(x|p) \\ =& Const2 \cdot p^{a-1+x}(1-p)^{b-1+1-x} \\ =& Beta(p|\tilde{a},\tilde{b}) \end{aligned}$$

So the posterior probability distribution is also beta distributed and the parameters just changed to $\tilde{a} \leftarrow a + x$ and $\tilde{b} \leftarrow b + 1 - x$

[1] Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b.

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b. After the experiment x, we can update the distribution of our estimated p by

$$\begin{aligned} p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bin(x|p,N) \\ =& Const2 \cdot p^{a-1+x}(1-p)^{b-1+N-x} \\ =& Beta(p|\tilde{a},\tilde{b}) \end{aligned}$$

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b. After the experiment x, we can update the distribution of our estimated p by

$$\begin{aligned} p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bin(x|p,N) \\ =& Const2 \cdot p^{a-1+x}(1-p)^{b-1+N-x} \\ =& Beta(p|\tilde{a},\tilde{b}) \end{aligned}$$

Again, the posterior distribution is still beta but with parameters updated to $\tilde{a} \leftarrow a+x$ and $\tilde{b} \leftarrow b+N-x$

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads

Prior and regularization

- One major reason of introducing prior is for the sake of “regularizing” the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - $3/10$, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is $3/10$

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - 3/10, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is 3/10
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - 3/10, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is 3/10
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?
  - 0? Okay, the estimate is a bit extreme. We know that it is very difficult to make a coin that always gives a tail

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - $3/10$, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is $3/10$
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?
  - 0? Okay, the estimate is a bit extreme. We know that it is very difficult to make a coin that always gives a tail
  - How about we first assumed that we actually flipped two times and got 1 head before we did experiment? We will estimate $1/12$ instead of $0/10$

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

  $$Beta(p|2,2)Bin(x = 0|p, N = 10) \sim Beta(0 + a, 10 + b) = Beta(2, 12)$$

  Now, what is the MAP estimate?

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

  $$Beta(p|2,2)Bin(x = 0|p, N = 10) \sim Beta(0 + a, 10 + b) = Beta(2, 12)$$

  Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2, 12)$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

  $$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a, 10+b) = Beta(2,12)$$

  Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$. Thus,

  $$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

  $$Beta(p|2,2)Bin(x = 0|p, N = 10) \sim Beta(0 + a, 10 + b) = Beta(2, 12)$$

  Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2, 12)$. Thus,

  $$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

- Recall that $Beta(1, 1) = 1$ and so likelihood function is equivalent to $Beta(p|1,1)Bin(0|p, 10) \sim Beta(1, 11)$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

  $$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a,10+b) = Beta(2,12)$$

  Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$. Thus,

  $$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

- Recall that $Beta(1,1) = 1$ and so likelihood function is equivalent to $Beta(p|1,1)Bin(0|p,10) \sim Beta(1,11)$. Thus the ML estimate is the mode of $Beta(1,11) \Rightarrow p_{Head}^{(ML)} = \frac{1-1}{1+11-2} = \frac{0}{10} = 0$
  - This indeed is the same as our high school naïve estimate

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p,

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean.

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean. Thus

$$p_{Head}^{(Bayesian)} = \frac{a}{a+b} = \frac{1}{11}$$

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean. Thus

$$p_{Head}^{(Bayesian)} = \frac{a}{a+b} = \frac{1}{11}$$

- Note that Bayesian estimation is "self-regularized" (i.e., giving less extreme results) since it inherently averages out all possible cases

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

$$Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) = \binom{N}{x_1 x_2 \cdots x_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n},$$

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

$$Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) = \binom{N}{x_1 x_2 \cdots x_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n},$$

- Just make sure we are in the same pace. Note that $p_1 + p_2 + \cdots + p_n = 1$ and $x_1 + x_2 + \cdots + x_n = N$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$
- It turns out that the distribution is the so-called Dirichlet distribution. Its pdf is given by

$$\begin{aligned} &Dir(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n) \\ =&\frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}\end{aligned}$$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$
- It turns out that the distribution is the so-called Dirichlet distribution. Its pdf is given by

$$
\begin{aligned}
&Dir(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n)\\
=&\frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}
\end{aligned}
$$

- As usual since pdf should be normalized to 1, we have

$$
\int x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n)}
$$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1 + 1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n + 1)} = \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_n}$$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1 + 1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n + 1)} = \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1+1} x_2^{\alpha_2-1} \cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$.

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$
$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1 + 1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n + 1)} = \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1+1} x_2^{\alpha_2-1} \cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$. Thus, $Var(X_1) = E[X_1^2] - E[X_1^2] = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)} - \frac{\alpha_1^2}{(\alpha_1+\cdots+\alpha_n)^2} = \frac{\alpha_1(\alpha_0-\alpha_1)}{\alpha_0^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1 + \cdots + \alpha_n$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$
$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1 + 1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n + 1)} = \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1+1} x_2^{\alpha_2-1} \cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$. Thus, $Var(X_1) = E[X_1^2] - E[X_1^2] = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)} - \frac{\alpha_1^2}{(\alpha_1+\cdots+\alpha_n)^2} = \frac{\alpha_1(\alpha_0-\alpha_1)}{\alpha_0^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1 + \cdots + \alpha_n$
- Mode: one can show that the mode of $Dir(\alpha_1, \cdots, \alpha_n)$ is

$$\frac{\alpha_i - 1}{\alpha_1 + \cdots + \alpha_n - n}.$$

We will not show it now but will leave as an exercise

Summary of Dirichlet distribution

- Pdf:

$$Dir(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

- Mean:

$$\frac{\alpha_i}{\alpha_1 + \cdots + \alpha_n}$$

- Variance:

$$\frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

- Mode:

$$\frac{\alpha_i - 1}{\alpha_1 + \cdots + \alpha_n - n}$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n)$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$\begin{aligned} & p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n) \\ = & Const1 \cdot Dir(p_1, \cdots, p_n | \alpha_1, \cdots, \alpha_n) Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) \end{aligned}$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$\begin{aligned} & p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n) \\ = & Const1 \cdot Dir(p_1, \cdots, p_n | \alpha_1, \cdots, \alpha_n) Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) \\ = & Const2 \cdot p_1^{x_1 + \alpha_1} \cdots p_n^{x_n + \alpha_n} \\ = & Dir(p_1, \cdots, p_n | \tilde{\alpha}_1, \cdots, \tilde{\alpha}_n) \end{aligned}$$

So the posterior distribution is Dirichlet with parameters updated to $\tilde{\alpha}_1 \leftarrow x_1 + \alpha_1, \cdots, \tilde{\alpha}_n \leftarrow x_n + \alpha_n$

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store.

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store. Its pdf is given by

$$Poisson(k|\lambda T) = \frac{e^{-\lambda T}(\lambda T)^k}{k!},$$

where k is a non-negative integer, λ is rate of arrival and T is the length of the observed period.

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store. Its pdf is given by

$$Poisson(k|\lambda T) = \frac{e^{-\lambda T}(\lambda T)^k}{k!},$$

where k is a non-negative integer, λ is rate of arrival and T is the length of the observed period. It is easy to check that (please verify)

$$Mean = \lambda T$$

$$Variance = \lambda T$$

N.B. the parameters λT comes as a group and so we can consider it as a single parameter

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
    - That is, λ is a constant that does not change with time

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
    - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
   - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
   - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease
   - It makes sense to model say customers to a department store

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
   - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease
   - It makes sense to model say customers to a department store
   - It can be less perfect to model the times my car broke down. The events are likely to be related

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T)$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
$Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
$= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
$Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
$= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} \approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} \approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$
  $= \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} \approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$
  $= \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^N$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} \approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$
  $= \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^N = \frac{(\lambda T)^k}{k!}\exp(-\lambda T)$,
  where we use $(1+a/N)^N = \exp(a)$ for the last equality

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
$Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
$= \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} \approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$
$= \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^N = \frac{(\lambda T)^k}{k!}\exp(-\lambda T)$,
where we use $(1+a/N)^N = \exp(a)$ for the last equality

Note that indeed $Pr(k \text{ arrivals in } T) = Poisson(k|\lambda T)$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$.

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  Pr(next event happened within in time $[t, t+\Delta]$)
  $= Pr$(next event happened within in time $[n\Delta, (n+1)\Delta]$)

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  $Pr(\text{next event happened within in time } [t, t + \Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n + 1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals})Pr(\text{event happened in } n + 1 \text{ interval})$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  $Pr(\text{next event happened within in time } [t, t + \Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals})Pr(\text{event happened in } n+1 \text{ interval})$
  $= (1 - \lambda\Delta)^n(\lambda\Delta)$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  $Pr(\text{next event happened within in time } [t, t+\Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n+1 \text{ interval})$
  $= (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then,
  $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta}$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  Pr(next event happened within in time $[t, t + \Delta]$)
  $= Pr$(next event happened within in time $[n\Delta, (n+1)\Delta]$)
  $= Pr$(no event in first n intervals)Pr(event happened in $n+1$ interval)
  $= (1 - \lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then,
  $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1 - \lambda\frac{t}{n})^n$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  $Pr(\text{next event happened within in time } [t, t + \Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n + 1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals})Pr(\text{event happened in } n + 1 \text{ interval})$
  $= (1 - \lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then,
  $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1 - \lambda\frac{t}{n})^n = \lambda\exp(-\lambda t)$, where we use
  $(1 + a/n)^n = \exp(a)$ again for $n \to \infty$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now,
  Pr(next event happened within in time $[t, t+\Delta]$)
  $= Pr$(next event happened within in time $[n\Delta, (n+1)\Delta]$)
  $= Pr$(no event in first n intervals)Pr(event happened in $n+1$ interval)
  $= (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then,
  $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1-\lambda\frac{t}{n})^n = \lambda\exp(-\lambda t)$, where we use
  $(1+a/n)^n = \exp(a)$ again for $n \to \infty$

Exponential distribution

$f_T(t) = \lambda\exp(-\lambda t) \triangleq Exp(t|\lambda)$ is the pdf of the exponential distribution with parameter λ. It is easy to verify that (as exercise)

- $E[T] = 1/\lambda$
- $Var(T) = 1/\lambda^2$

Normal distribution revisit

For a univariate normal random variable, the pdf is given by

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

with

$$
E[X|\mu, \sigma^2] = \mu,
$$

$$
E[(X-\mu)^2|\mu, \sigma^2] = \sigma^2,
$$

Recall that $\lambda = \frac{1}{\sigma^2}$ is the precision parameter that simplifies computations in many cases

Conjugate prior of normal distribution for fixed σ_2

Consider σ^2 fixed and μ as the model parameter, then the posterior probability is given by

$$p(\mu|x;\sigma^2) \propto p(\mu, x;\sigma^2)$$

Conjugate prior of normal distribution for fixed σ_2

Consider σ^2 fixed and μ as the model parameter, then the posterior probability is given by

$$p(\mu|x;\sigma^2) \propto p(\mu,x;\sigma^2)$$
$$= p(\mu) Norm(x|\mu;\sigma^2)$$

Conjugate prior of normal distribution for fixed σ_2

Consider σ^2 fixed and μ as the model parameter, then the posterior probability is given by

$$p(\mu|x;\sigma^2) \propto p(\mu,x;\sigma^2)$$
$$=p(\mu)Norm(x|\mu;\sigma^2)$$
$$\propto p(\mu)exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Conjugate prior of normal distribution for fixed σ_2

Consider σ^2 fixed and μ as the model parameter, then the posterior probability is given by

$$
\begin{aligned}
&p(\mu|x;\sigma^2) \propto p(\mu, x; \sigma^2) \\
=&p(\mu) Norm(x|\mu;\sigma^2) \\
\propto&p(\mu) exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

It is apparent that the posterior will keep the same form if $p(\mu)$ is also normal. Therefore, normal distribution is the conjugate prior of itself for fixed variance

Posterior distribution of normal variable for fixed σ^2

Given prior $p(\mu) = Norm(\mu|\mu_0, \sigma_0^2)$ and likelihood $Norm(x|\mu; \sigma^2)$. Let's find the posterior probability,

$$\begin{aligned} & p(\mu|x; \sigma^2, \mu_0, \sigma_0^2) \\ = & Const \cdot Norm(\mu|\mu_0, \sigma_0^2) Norm(x|\mu; \sigma^2) \end{aligned}$$

Posterior distribution of normal variable for fixed σ^2

Given prior $p(\mu) = Norm(\mu|\mu_0, \sigma_0^2)$ and likelihood $Norm(x|\mu; \sigma^2)$. Let's find the posterior probability,

$$\begin{aligned} & p(\mu|x; \sigma^2, \mu_0, \sigma_0^2) \\ =& Const \cdot Norm(\mu|\mu_0, \sigma_0^2) Norm(x|\mu; \sigma^2) \\ =& Const2 \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ =& Norm\left(\mu; \tilde{\mu}, \tilde{\sigma}^2\right), \end{aligned}$$

Posterior distribution of normal variable for fixed σ^2

Given prior $p(\mu) = Norm(\mu|\mu_0, \sigma_0^2)$ and likelihood $Norm(x|\mu; \sigma^2)$. Let's find the posterior probability,

$$\begin{aligned} & p(\mu|x; \sigma^2, \mu_0, \sigma_0^2) \\ = & Const \cdot Norm(\mu|\mu_0, \sigma_0^2) Norm(x|\mu; \sigma^2) \\ = & Const2 \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ = & Norm\left(\mu; \tilde{\mu}, \tilde{\sigma}^2\right), \end{aligned}$$

where $\tilde{\mu} = \frac{\sigma_0^2 x + \mu_0 \sigma^2}{\sigma_0^2 + \sigma^2}$ and $\tilde{\sigma}^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}$. Alternatively, $\tilde{\lambda} = \lambda_0 + \lambda$ and $\tilde{\mu} = \frac{\lambda}{\tilde{\lambda}} x + \frac{\lambda_0}{\tilde{\lambda}} \mu_0$. Note that we have already came across the more general expression when we studied product of multivariate normal distribution

Conjugate prior of normal distribution for fixed μ

Consider μ fixed and λ as the model parameter

$$p(x|\lambda;\mu) \propto p(x,\lambda;\mu) = p(\lambda) Norm(x|\lambda;\mu)$$

Conjugate prior of normal distribution for fixed μ

Consider μ fixed and λ as the model parameter

$$
\begin{aligned}
p(x|\lambda;\mu) &\propto p(x,\lambda;\mu) = p(\lambda) Norm(x|\lambda;\mu) \\
&\propto p(\lambda)\sqrt{\lambda}\exp\left(-\frac{\lambda(x-\mu)^2}{2}\right)
\end{aligned}
$$

Conjugate prior of normal distribution for fixed μ

Consider μ fixed and λ as the model parameter

$$
\begin{aligned}
p(x|\lambda;\mu) \propto & p(x,\lambda;\mu) = p(\lambda) Norm(x|\lambda;\mu) \\
\propto & p(\lambda)\sqrt{\lambda}\exp\left(-\frac{\lambda(x-\mu)^2}{2}\right)
\end{aligned}
$$

More generally, when we have N observations from the same source,

$$
\begin{aligned}
p(x_1,\cdots,x_N,\lambda;\mu) = & p(\lambda)\prod_{i=1}^{N} Norm(x_i|\lambda;\mu) \\
\propto & p(\lambda)\lambda^{\frac{N}{2}}\exp\left(-\lambda\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2}\right)
\end{aligned}
$$

Conjugate prior of normal distribution for fixed μ

Consider μ fixed and λ as the model parameter

$$
\begin{aligned}
p(x|\lambda;\mu) \propto & p(x,\lambda;\mu) = p(\lambda) Norm(x|\lambda;\mu) \\
\propto & p(\lambda)\sqrt{\lambda}\exp\left(-\frac{\lambda(x-\mu)^2}{2}\right)
\end{aligned}
$$

More generally, when we have N observations from the same source,

$$
\begin{aligned}
p(x_1,\cdots,x_N,\lambda;\mu) = & p(\lambda)\prod_{i=1}^{N} Norm(x_i|\lambda;\mu) \\
\propto & p(\lambda)\lambda^{\frac{N}{2}}\exp\left(-\lambda\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2}\right)
\end{aligned}
$$

From inspection, the conjugate prior should have a form $\lambda^a \exp(-b\lambda)$

Gamma distribution

The distribution with the desired form described in previous slide turns out to be the Gamma distribution. Its pdf, mean, and variance (please verify the mean and variance) are given by

$$Gamma(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)$$

$$E[\lambda] = \frac{a}{b}$$

$$Var[\lambda] = \frac{a}{b^2},$$

where $a, b > 0$ and $\lambda \geq 0$

Gamma distribution

The distribution with the desired form described in previous slide turns out to be the Gamma distribution. Its pdf, mean, and variance (please verify the mean and variance) are given by

$$Gamma(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)$$

$$E[\lambda] = \frac{a}{b}$$

$$Var[\lambda] = \frac{a}{b^2},$$

where $a, b > 0$ and $\lambda \geq 0$

N.B. when $a = 1$, Gamma reduces to the exponential distribution. When a is integer, it reduces to Erlang distribution

Posterior distribution of normal variable for fixed μ

Posterior probability given Normal likelihood (fixed mean) and Gamma prior

$$p(\lambda|x, a, b; \mu) = Const1 \cdot Gamma(\lambda|a, b) Norm(x|\lambda; \mu)$$

Posterior distribution of normal variable for fixed μ

Posterior probability given Normal likelihood (fixed mean) and Gamma prior

$$\begin{aligned}
p(\lambda|x, a, b; \mu) &= Const1 \cdot Gamma(\lambda|a, b) Norm(x|\lambda; \mu) \\
&= Const2 \cdot \lambda^{a-1} \exp(-b\lambda)\sqrt{\lambda} \exp\left(-\lambda \frac{(x-\mu)^2}{2}\right) \\
&= Gamma\left(\lambda; \tilde{a}, \tilde{b}\right),
\end{aligned}$$

where $\tilde{a} \leftarrow a + \frac{1}{2}$ and $\tilde{b} \leftarrow b + \frac{(x-\mu)^2}{2}$

Conjugate prior summary

Distribution	Likelihood $p(\mathbf{x}\|\theta)$	Prior $p(\theta)$	Distribution
Bernoulli	$(1-\theta)^{(1-x)}\theta^{x}$	$\propto (1-\theta)^{(a-1)}\theta^{(b-1)}$	Beta
Binomial	$\propto (1-\theta)^{(N-x)}\theta^{x}$	$\propto (1-\theta)^{(a-1)}\theta^{(b-1)}$	Beta
Multinomial	$\propto \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}$	$\propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1}$	Dirichlet
Normal (fixed σ^2)	$\propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$	$\propto \exp\left(-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right)$	Normal
Normal (fixed μ)	$\propto \sqrt{\theta}\exp\left(-\frac{\theta(x-\mu)^2}{2}\right)$	$\propto \theta^{a-1}exp(-b\theta)$	Gamma
Poisson	$\propto \theta^{x}\exp(-\theta)$	$\propto \theta^{a-1}exp(-b\theta)$	Gamma

Lagrange multiplier

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) = 0$$

Consider $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$ and let $\tilde{f}(\mathbf{x}) = \min_{\lambda} L(\mathbf{x}, \lambda)$.

Lagrange multiplier

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) = 0$$

Consider $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$ and let $\tilde{f}(\mathbf{x}) = \min_\lambda L(\mathbf{x}, \lambda)$. Note that

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) \text{ if } g(\mathbf{x}) = 0 \\ -\infty \text{ otherwise} \end{cases}$$

Lagrange multiplier

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) = 0$$

Consider $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$ and let $\tilde{f}(\mathbf{x}) = \min_\lambda L(\mathbf{x}, \lambda)$. Note that

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) \text{ if } g(\mathbf{x}) = 0 \\ -\infty \text{ otherwise} \end{cases}$$

Therefore, the problem is identical to $\max_{\mathbf{x}} \tilde{f}(\mathbf{x})$ or

$$\max_{\mathbf{x}} \min_{\lambda} f(\mathbf{x}) - \lambda g(\mathbf{x}),$$

where λ is known to be the Lagrange multiplier.

Lagrange multiplier (con't)

Assume the optimum is a saddle point,

$$\max_{\mathbf{x}} \min_{\lambda} f(\mathbf{x}) - \lambda g(\mathbf{x}) = \min_{\lambda} \max_{\mathbf{x}} f(\mathbf{x}) - \lambda g(\mathbf{x}),$$

the R.H.S. implies

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

Inequality constraint

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) \leq 0$$

Consider $\tilde{f}(\mathbf{x}) = \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x})$,

Inequality constraint

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) \leq 0$$

Consider $\tilde{f}(\mathbf{x}) = \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x})$, note that

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq 0 \\ -\infty & \text{otherwise} \end{cases}$$

Inequality constraint

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$

$$g(\mathbf{x}) \leq 0$$

Consider $\tilde{f}(\mathbf{x}) = \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x})$, note that

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq 0 \\ -\infty & \text{otherwise} \end{cases}$$

Therefore, we can rewrite the problem as

$$\max_{\mathbf{x}} \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Inequality constraint (con't)

Assume

$$\max_{\mathbf{x}} \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x}) = \min_{\lambda \geq 0} \max_{\mathbf{x}} f(\mathbf{x}) - \lambda g(\mathbf{x})$$

The R.H.S. implies

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

Inequality constraint (con't)

Assume

$$\max_{\mathbf{x}} \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x}) = \min_{\lambda \geq 0} \max_{\mathbf{x}} f(\mathbf{x}) - \lambda g(\mathbf{x})$$

The R.H.S. implies

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

Moreover, at the optimum point $(\mathbf{x}^*, \lambda^*)$, we should have

$$\lambda^* g(\mathbf{x}^*) = 0$$

since

$$\max_{\substack{\mathbf{x} \\ g(\mathbf{x}) \leq 0}} f(\mathbf{x}) \equiv \max_{\mathbf{x}} \min_{\lambda \geq 0} f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Karush-Kuhn-Tucker conditions

Problem

$$\max_{\mathbf{x}} f(\mathbf{x})$$
$$g(\mathbf{x}) \leq 0, \quad h(\mathbf{x}) = 0$$

Conditions

$$\begin{aligned} \nabla f(\mathbf{x}^*) - \mu^* \nabla g(\mathbf{x}^*) - \lambda^* \nabla h(\mathbf{x}^*) &= 0 \\ g(\mathbf{x}^*) &\leq 0 \\ h(\mathbf{x}^*) &= 0 \\ \mu^* &\geq 0 \\ \mu^* g(\mathbf{x}^*) &= 0 \end{aligned}$$

Kraft's Inequality

Let $l_1, l_2, \cdots, l_K$ satisfy $\sum_{k=1}^{K} 2^{-l_k} \leq 1$. Then, there exists a uniquely decodable code for symbols $x_1, x_2, \cdots, x_K$ such that $l(x_1) = l_1$, $l(x_2) = l_2, \cdots, l(x_K) = l_K$.

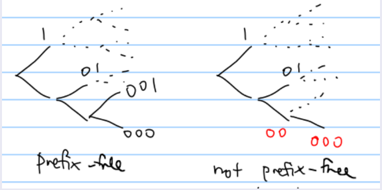
Kraft's Inequality

Let $l_1, l_2, \cdots, l_K$ satisfy $\sum_{k=1}^{K} 2^{-l_k} \leq 1$. Then, there exists a uniquely decodable code for symbols $x_1, x_2, \cdots, x_K$ such that $l(x_1) = l_1$, $l(x_2) = l_2, \cdots, l(x_K) = l_K$.

Intuition

Consider # "descendants" of each codeword at the "l_{max}"-level, then for prefix-free code, we have

$$\sum_{k=1}^{K} 2^{l_{max}-l} \leq 2^{l_{max}}$$



Kraft's Inequality

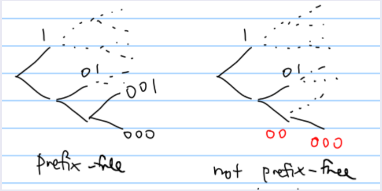
Let $l_1, l_2, \cdots, l_K$ satisfy $\sum_{k=1}^{K} 2^{-l_k} \leq 1$. Then, there exists a uniquely decodable code for symbols $x_1, x_2, \cdots, x_K$ such that $l(x_1) = l_1$, $l(x_2) = l_2, \cdots, l(x_K) = l_K$.

Intuition

Consider # "descendants" of each codeword at the "l_{max}"-level, then for prefix-free code, we have

$$\sum_{k=1}^{K} 2^{l_{max}-l} \leq 2^{l_{max}}$$

$$\Rightarrow \sum_{k=1}^{K} 2^{-l_k} \leq 1$$



Forward Proof

Given $l_1, l_2, \cdots, l_K$ satisfy $\sum_{k=1}^{K} 2^{-l_k} \leq 1$, we can assign nodes on a tree as previous slides. More precisely,

- Assign i-th node as a node at level l_i, then cross out all its descendants
- Repeat the procedure for i from 1 to K
- We know that there are sufficient tree nodes to be assigned since the Kraft's inequaltiy is satisfied

The corresponding code is apparently prefix-free and thus is uniquely decodable

Converse Proof

Consider message from coding k symbols $\mathbf{x} = x_1, x_2, \cdots, x_k$

$$
\begin{aligned}
\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k &= \left(\sum_{x_1\in\mathcal{X}} 2^{-l(x_1)}\right)\left(\sum_{x_2\in\mathcal{X}} 2^{-l(x_2)}\right)\cdots\left(\sum_{x_3\in\mathcal{X}} 2^{-l(x_3)}\right) \\
&= \sum_{x_1,x_2,\cdots,x_k\in\mathcal{X}^k} 2^{-(l(x_1)+l(x_2)+\cdots+l(x_k))} \\
&= \sum_{\mathbf{x}\in\mathcal{X}^k} 2^{-l(\mathbf{x})}
\end{aligned}
$$

Converse Proof

Consider message from coding k symbols $\mathbf{x} = x_1, x_2, \cdots, x_k$

$$\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k = \left(\sum_{x_1\in\mathcal{X}} 2^{-l(x_1)}\right)\left(\sum_{x_2\in\mathcal{X}} 2^{-l(x_2)}\right)\cdots\left(\sum_{x_3\in\mathcal{X}} 2^{-l(x_3)}\right)$$
$$= \sum_{x_1,x_2,\cdots,x_k\in\mathcal{X}^k} 2^{-(l(x_1)+l(x_2)+\cdots+l(x_k))}$$
$$= \sum_{\mathbf{x}\in\mathcal{X}^k} 2^{-l(\mathbf{x})} = \sum_{m=1}^{kl_{max}} a(m)2^{-m},$$

where $a(m)$ is the number of codeword with length m. However, for the code to be uniquely decodable, $a(m) \leq 2^m$, where 2^m is the number of available codewords with length m.

Converse Proof

Consider message from coding k symbols $\mathbf{x} = x_1, x_2, \cdots, x_k$

$$\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k = \left(\sum_{x_1\in\mathcal{X}} 2^{-l(x_1)}\right)\left(\sum_{x_2\in\mathcal{X}} 2^{-l(x_2)}\right)\cdots\left(\sum_{x_3\in\mathcal{X}} 2^{-l(x_3)}\right)$$
$$= \sum_{x_1,x_2,\cdots,x_k\in\mathcal{X}^k} 2^{-(l(x_1)+l(x_2)+\cdots+l(x_k))}$$
$$= \sum_{\mathbf{x}\in\mathcal{X}^k} 2^{-l(\mathbf{x})} = \sum_{m=1}^{kl_{max}} a(m)2^{-m},$$

where $a(m)$ is the number of codeword with length m. However, for the code to be uniquely decodable, $a(m) \leq 2^m$, where 2^m is the number of available codewords with length m. Therefore,

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq (kl_{max})^{1/k}$$

Converse Proof

Consider message from coding k symbols $\mathbf{x} = x_1, x_2, \cdots, x_k$

$$\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k = \left(\sum_{x_1\in\mathcal{X}} 2^{-l(x_1)}\right)\left(\sum_{x_2\in\mathcal{X}} 2^{-l(x_2)}\right)\cdots\left(\sum_{x_3\in\mathcal{X}} 2^{-l(x_3)}\right)$$
$$= \sum_{x_1,x_2,\cdots,x_k\in\mathcal{X}^k} 2^{-(l(x_1)+l(x_2)+\cdots+l(x_k))}$$
$$= \sum_{\mathbf{x}\in\mathcal{X}^k} 2^{-l(\mathbf{x})} = \sum_{m=1}^{kl_{max}} a(m)2^{-m},$$

where $a(m)$ is the number of codeword with length m. However, for the code to be uniquely decodable, $a(m) \leq 2^m$, where 2^m is the number of available codewords with length m. Therefore,

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq (kl_{max})^{1/k} \approx 1 \text{ as } k \to \infty$$

Minimum rate required to compress a source

$$min_{l_1,l_2,\cdots,l_K} \sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} \leq 1 \text{ and } l_1,\cdots,l_K \geq 0$$

$$\equiv max_{l_1,l_2,\cdots,l_K} - \sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} - 1 \leq 0 \text{ and } -l_1,\cdots,-l_K \leq 0$$

KKT conditions

$$-\nabla\left(\sum_{k=1}^{K} p_k l_k\right) - \mu_0 \nabla\left(\sum_{k=1}^{K} 2^{-l_k} - 1\right) + \sum_{k=1}^{K} \mu_k \nabla l_k = 0$$

Minimum rate required to compress a source

$$min_{l_1,l_2,\cdots,l_K} \sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} \leq 1 \text{ and } l_1,\cdots,l_K \geq 0$$

$$\equiv max_{l_1,l_2,\cdots,l_K} -\sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} - 1 \leq 0 \text{ and } -l_1,\cdots,-l_K \leq 0$$

KKT conditions

$$-\nabla\left(\sum_{k=1}^{K} p_k l_k\right) - \mu_0 \nabla\left(\sum_{k=1}^{K} 2^{-l_k} - 1\right) + \sum_{k=1}^{K} \mu_k \nabla l_k = 0$$

$$\sum_{k=1}^{K} 2^{-l_k} - 1 \leq 0, \quad l_1,\cdots,l_K \geq 0, \quad \mu_0,\mu_1,\cdots,\mu_K \geq 0$$

Minimum rate required to compress a source

$$min_{l_1,l_2,\cdots,l_K} \sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} \leq 1 \text{ and } l_1,\cdots,l_K \geq 0$$

$$\equiv max_{l_1,l_2,\cdots,l_K} -\sum_{k=1}^{K} p_k l_k \text{ subject to } \sum_{k=1}^{K} 2^{-l_k} - 1 \leq 0 \text{ and } -l_1,\cdots,-l_K \leq 0$$

KKT conditions

$$-\nabla\left(\sum_{k=1}^{K} p_k l_k\right) - \mu_0 \nabla\left(\sum_{k=1}^{K} 2^{-l_k} - 1\right) + \sum_{k=1}^{K} \mu_k \nabla l_k = 0$$

$$\sum_{k=1}^{K} 2^{-l_k} - 1 \leq 0, \quad l_1,\cdots,l_K \geq 0, \quad \mu_0,\mu_1,\cdots,\mu_K \geq 0$$

$$\mu_0\left(\sum_{k=1}^{K} 2^{-l_k} - 1\right) = 0, \quad \mu_k l_k = 0$$

Minimum rate required to compress a source

Since we expect $l_k > 0$, $\mu_k = 0$.

Minimum rate required to compress a source

Since we expect $l_k > 0$, $\mu_k = 0$. Expand the first equation, we get

$$-p_j + \mu_0 2^{-l_j} \log 2 = 0 \Rightarrow 2^{-l_j} = \frac{p_j}{\mu_0 \log 2}$$

Minimum rate required to compress a source

Since we expect $l_k > 0$, $\mu_k = 0$. Expand the first equation, we get

$$-p_j + \mu_0 2^{-l_j} \log 2 = 0 \Rightarrow 2^{-l_j} = \frac{p_j}{\mu_0 \log 2}$$

And by $\sum_{k=1}^{K} 2^{-l_k} \leq 1$, we have

$$\sum_{k=1}^{K} \frac{p_j}{\mu_0 \log 2} = \frac{1}{\mu_0 \log 2} \leq 1 \Rightarrow \mu_0 \geq \frac{1}{\log 2}$$

Minimum rate required to compress a source

Since we expect $l_k > 0$, $\mu_k = 0$. Expand the first equation, we get

$$-p_j + \mu_0 2^{-l_j} \log 2 = 0 \Rightarrow 2^{-l_j} = \frac{p_j}{\mu_0 \log 2}$$

And by $\sum_{k=1}^{K} 2^{-l_k} \leq 1$, we have

$$\sum_{k=1}^{K} \frac{p_j}{\mu_0 \log 2} = \frac{1}{\mu_0 \log 2} \leq 1 \Rightarrow \mu_0 \geq \frac{1}{\log 2}$$

Note that as $\mu_0 \downarrow$, $\frac{p_j}{\mu_0 \log 2} \uparrow$ and $l_j \downarrow$.

Minimum rate required to compress a source

Since we expect $l_k > 0$, $\mu_k = 0$. Expand the first equation, we get

$$-p_j + \mu_0 2^{-l_j} \log 2 = 0 \Rightarrow 2^{-l_j} = \frac{p_j}{\mu_0 \log 2}$$

And by $\sum_{k=1}^{K} 2^{-l_k} \leq 1$, we have

$$\sum_{k=1}^{K} \frac{p_j}{\mu_0 \log 2} = \frac{1}{\mu_0 \log 2} \leq 1 \Rightarrow \mu_0 \geq \frac{1}{\log 2}$$

Note that as $\mu_0 \downarrow$, $\frac{p_j}{\mu_0 \log 2} \uparrow$ and $l_j \downarrow$. Therefore, if we want to decrease code rate, we should reduce μ_0 as much as possible. Thus, take $\mu_0 = \frac{1}{\log 2}$. Then $2^{-l_j} = p_j \Rightarrow l_j = -\log_2 p_j$. Thus, the minimum rate becomes

$$\sum_{k=1}^{K} p_k l_k = -\sum_{k=1}^{K} p_k \log_2 p_k \triangleq H(p_1, \cdots, p_K)$$

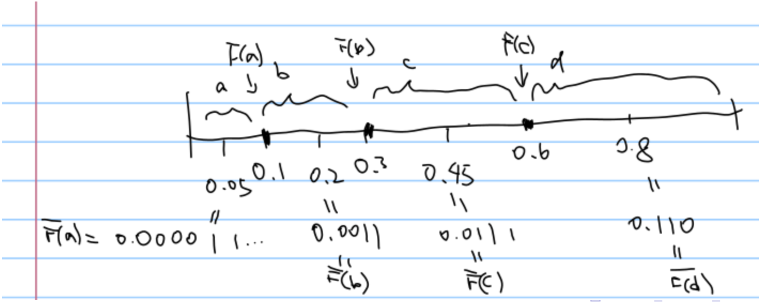
Shannon-Fano-Elias code

Key idea

Each codeword corresponds to an intervel of $[0, 1]$.

Example

110 corresponds to $[0.110, 0.110\underset{\cdot}{1}] = [0.11, 0.111) = [0.75, 0.875)$



Shannon-Fano-Elias code

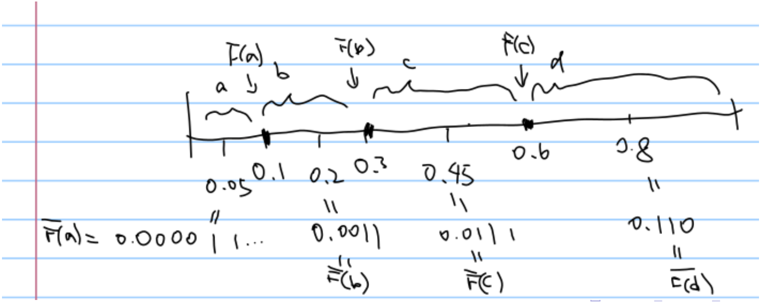
Key idea

Each codeword corresponds to an intervel of $[0, 1]$.

Example

110 corresponds to $[0.110, 0.1101] = [0.11, 0.111) = [0.75, 0.875)$

011 corresponds to $[0.011, 0.0111] = [0.011, 0.1) = [0.375, 0.5)$



Example

Consider a source that
$p(x_1) = 0.25, p(x_2) = 0.25, p(x_3) = 0.2, p(x_4) = 0.15, p(x_5) = 0.15$

x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ in Binary	$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$	Codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.25	0.5	0.375	0.011	3	011
3	0.2	0.7	0.6	$0.1\overline{0011}$	4	1001
4	0.15	0.85	0.775	$0.110\overline{0011}$	4	1100
5	0.15	1.0	0.925	$0.111\overline{0110}$	4	1110

Property

- The length of the codeword of x is $\lceil \log_2 \frac{1}{p(x)} \rceil + 1$. This ensures that the "code interval" of each codeword does not overlap

Property

- The length of the codeword of x is $\lceil \log_2 \frac{1}{p(x)} \rceil + 1$. This ensures that the "code interval" of each codeword does not overlap
- SFE code is prefix-free $\rightarrow$ uniquely decodable

Property

- The length of the codeword of x is $\lceil \log_2 \frac{1}{p(x)} \rceil + 1$. This ensures that the "code interval" of each codeword does not overlap
- SFE code is prefix-free $\rightarrow$ uniquely decodable
  - If a codeword is prefix of another (say 10 and 1010), the corresponding intervals must overlap each other (consider $[0.10, 0.11)$ and $[0.101, 0.11)$)

Property

- The length of the codeword of x is $\lceil \log_2 \frac{1}{p(x)} \rceil + 1$. This ensures that the "code interval" of each codeword does not overlap
- SFE code is prefix-free $\rightarrow$ uniquely decodable
  - If a codeword is prefix of another (say 10 and 1010), the corresponding intervals must overlap each other (consider $[0.10, 0.11)$ and $[0.101, 0.11)$)
  - Since no codeword can overlap in SFE, no code word can be prefix of another

Property

- The length of the codeword of x is $\lceil \log_2 \frac{1}{p(x)} \rceil + 1$. This ensures that the "code interval" of each codeword does not overlap
- SFE code is prefix-free $\rightarrow$ uniquely decodable
  - If a codeword is prefix of another (say 10 and 1010), the corresponding intervals must overlap each other (consider $[0.10, 0.11)$ and $[0.101, 0.11)$)
  - Since no codeword can overlap in SFE, no code word can be prefix of another
- Average code rate is upper bounded by $H(X) + 2$

$$\begin{aligned}\sum_{x \in \mathcal{X}} p(x) l(x) &= \sum_{x \in \mathcal{X}} p(x) \left(\left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1 \right) \\ &\leq \sum_{x \in \mathcal{X}} p(x) \left(\log_2 \frac{1}{p(x)} + 2 \right) = H(X) + 2\end{aligned}$$

"Symbol grouping" trick

- Let's consider two symbols as a super-symbol and compress the pair at each time with SFE code
- The code rate is bounded by $H(X_S) + 2$, where

"Symbol grouping" trick

- Let's consider two symbols as a super-symbol and compress the pair at each time with SFE code
- The code rate is bounded by $H(X_S) + 2$, where

$$H(X_S) = -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 p(x_1, x_2)$$

"Symbol grouping" trick

- Let's consider two symbols as a super-symbol and compress the pair at each time with SFE code
- The code rate is bounded by $H(X_S) + 2$, where

$$\begin{aligned} H(X_S) &= -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 p(x_1, x_2) \\ &= -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 (p(x_1) p(x_2)) \end{aligned}$$

"Symbol grouping" trick

- Let's consider two symbols as a super-symbol and compress the pair at each time with SFE code
- The code rate is bounded by $H(X_S) + 2$, where

$$\begin{aligned} H(X_S) &= -\sum_{x_1,x_2 \in \mathcal{X}^2} p(x_1,x_2) \log_2 p(x_1,x_2) \\ &= -\sum_{x_1,x_2 \in \mathcal{X}^2} p(x_1,x_2) \log_2 (p(x_1)p(x_2)) \\ &= -\sum_{x_1,x_2 \in \mathcal{X}^2} p(x_1,x_2) \log_2 p(x_1) - \sum_{x_1,x_2 \in \mathcal{X}^2} p(x_1,x_2) \log_2 p(x_2) \end{aligned}$$

"Symbol grouping" trick

- Let's consider two symbols as a super-symbol and compress the pair at each time with SFE code
- The code rate is bounded by $H(X_S) + 2$, where

$$\begin{aligned}
H(X_S) &= -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 p(x_1, x_2) \\
&= -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 (p(x_1) p(x_2)) \\
&= -\sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 p(x_1) - \sum_{x_1, x_2 \in \mathcal{X}^2} p(x_1, x_2) \log_2 p(x_2) \\
&= -\sum_{x_1 \in \mathcal{X}} p(x_1) \log_2 p(x_1) - \sum_{x_2 \in \mathcal{X}} p(x_2) \log_2 p(x_2) \\
&= 2H(X)
\end{aligned}$$

Therefore, the code rate per original symbol is upper bounded by

$$\frac{1}{2}(H(X_S) + 2) = H(X) + 1$$

Forward proof of Source Coding Theorem

In theory, we can group as many symbol as we want (we want do it in practice, why?), say we group N symbols at a time and compress it using SFE code.

Forward proof of Source Coding Theorem

In theory, we can group as many symbol as we want (we want do it in practice, why?), say we group N symbols at a time and compress it using SFE code. The code rate per original symbol is upper bounded by

$$\frac{1}{N}(H(X_S)+2)=\frac{1}{N}(NH(X)+2)=H(X)+\frac{2}{N}$$

Forward proof of Source Coding Theorem

In theory, we can group as many symbol as we want (we want do it in practice, why?), say we group N symbols at a time and compress it using SFE code. The code rate per original symbol is upper bounded by

$$\frac{1}{N}(H(X_S)+2)=\frac{1}{N}(NH(X)+2)=H(X)+\frac{2}{N}$$

Therefore as long as a given rate $R > H(X)$, we can always find a large enough N such that the code rate using the "grouping trick" and SFE code is below R. This concludes the forward proof.