21. What is the singular value decomposition of a matrix A.

Singular value decomposition of matrix A is to decompose A to be a product of tranpose of one orthogonal matrix U, one diagonal matrix Σ whose diagonal entry is singular value of A and another othogonal matrix V.

$$A = U^T \Sigma V$$

22. Calculate the gradient of the following function

$$f(X) = tr\{X^T C X\}$$

Solutions starting from definition of derivative: (Refer to Notes)

$$\begin{aligned}
&\lim_{n\to 0} \frac{f(X+n\xi) - f(X)}{n} && (1)\\
&= \lim_{n\to 0} \frac{tr\{(X+n\xi)^T C(X+n\xi)\} - tr\{X^T C X\}}{n} && (2)\\
&= \lim_{n\to 0} \frac{tr\{(X+n\xi)^T C(X+n\xi) - X^T C X\}}{n} && (3)\\
&= \lim_{n\to 0} \frac{tr\{X^T C X + nX^T C\xi + n\xi^T C X + n^2\xi^T C\xi - X^T C X\}}{n} && (4)\\
&= \lim_{n\to 0} \frac{tr\{nX^T C\xi + n\xi^T C X + n^2\xi^T C\xi\}}{n} && (5)\\
&= \lim_{n\to 0} tr\{\frac{nX^T C\xi + n\xi^T C X + n^2\xi^T C\xi}{n}\} && (6)\\
&= \lim_{n\to 0} tr\{X^T C\xi + \xi^T C X + n\xi^T C\xi\} && (7)\\
&= tr\{X^T C\xi + \xi^T C X\} && (8)\\
&= tr\{X^T C\xi + X^T C^T \xi\} && (9)\\
&= tr\{X^T (C + C^T)\xi\} && (10)
\end{aligned}$$

23. Explain the difference of frequentist's approach and Bayesian approach. (Refer to Textook P23)

Aspects: **objective, sequential learning, compromise**

Frequetist's approach is to use maixmum likelihood estimation to do point estimation of parameter w. Does not require preknowledge but need a large amount of data. The sequential learning is based on the sequential update formula

$$w^{(\tau+1)} = w^{(\tau)} - \xi \bigtriangledown E(w)$$

Whereas the bayesian approach is to derive a probability distribution over the parameter based on the Bayes Theorem. This approach does not necesarily require a large amount of data if the prior knowledge is provided. And it is inherently a good framework for sequential learning. The posterior in current iteration works as the prior in the next iteration.

$$p(w^{(\tau+1)}|t) = \frac{p(t|w^{(\tau)})p(w^{(\tau)})}{p(t)}$$

According to the Bayes Theorem, there is a **compromise** between the observation of dataset and the preknowledge. If the amount of data is very large or preknowledge (prior) provides no disparity for various classes, the preknowledge can be ignored. However, when the data is scarce, the preknowledge plays an important role.

24. What is S-fold cross validation? Why we use it? (Refer to P32-33)

Cross validation is Model Selection algorithm, sometimes used for determining the parameters. Randomly group all labeled data into S sets, and for each model run experiment for S times. At each experiment leave one set as testing set and others as training set. Extreme: leave-one-out cross validation, S = N, each set contains only one piece of labeled data.

Reason:

- To assess a model, performance on training set is not sufficient since it cannot be generalised to all data set.
- When data is scarce, and in order to assess models appropriately, we wish to use as much of the available data as possible for training.

Drawbacks:

- The number of training runs that must be performed is increased by a factor of S, and this can prove problematic for models whose training is computationally expensive.
- Exploring combinations of settings for such parameters could require a number of training runs that is exponential in the number of parameters.

25. Derive a least squre solution for linear regression model (with regularisation). (Refer to P144-145)

The error function for least square error with regularisation:

$$\begin{aligned} E(\boldsymbol{w}) &= E_D(\boldsymbol{w}) + \lambda E_w(\boldsymbol{w}) && (11) \\ &= \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - t_n)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} && (12) \end{aligned}$$

λ is the regularization coefficient that controls the relative importance of the data-dependent error $E_D(\boldsymbol{w})$ and the regularization term $E_w(\boldsymbol{w})$.

$$\begin{aligned} \triangledown E(\boldsymbol{w}) &= \sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - t_n)\boldsymbol{\phi}(\boldsymbol{x}_n) + \lambda\boldsymbol{w} && (13) \\ &= \boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{t}) + \lambda\boldsymbol{w} && (14) \\ &= 0 && (15) \end{aligned}$$

$$w = (\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}$$

26. What is stochastic gradient descent? How to apply it to linear regression problem specifically with the sum-of-squares error function? (Refer to Textbook P144)

Stochastic gradient descent is one optimisation algorithm, which consider only the error of one piece of observed data at each iteration of training. That is where it is different from the batch gradient descent, which takes the averaged errors of all the observed data in one training iteration.

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \xi \triangledown E(\boldsymbol{w})$$

In sum-of-square function, we have error function for one observed data as follows,

$$E(\boldsymbol{w}) = (\boldsymbol{w}^T\Phi(x_n) - t_n)^2$$

whose gradient is

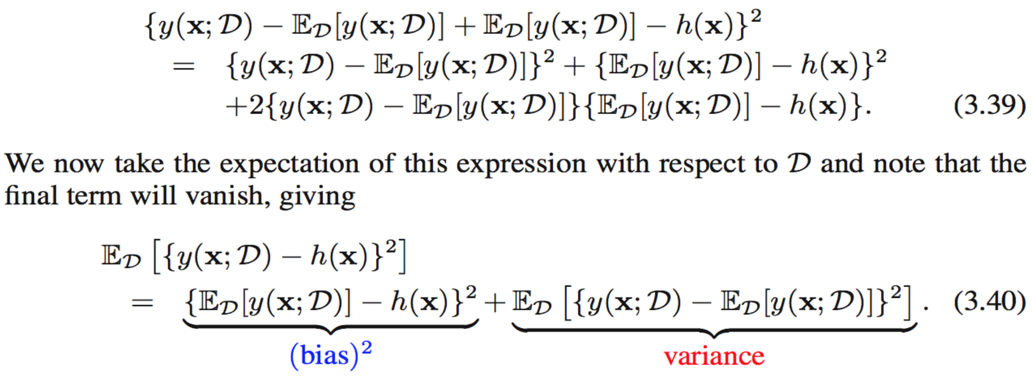
$$\triangledown E(\boldsymbol{w}) = (\boldsymbol{w}^T\Phi(x_n) - t_n)\Phi(x_n)$$

Take this gradient of error function back, we have

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \xi(\boldsymbol{w}^T\Phi(x_n) - t_n)\Phi(x_n)$$

27. What is the bias-variance decomposition? Demonstrate the bias-variance trade-off where the error is measured by the mean squared error. What you can deduce from the result? (Refer to Textbook P149)

Bias-variance trade-off is a frequentist viewpoint of the model complexity issue: **expected squared error** $\mathbb{E}\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2$ can be decomposed as summation of squared bias and the variance.



$$\begin{aligned}&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}. \qquad (3.39)\end{aligned}$$

We now take the expectation of this expression with respect to $\mathcal{D}$ and note that the final term will vanish, giving

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\ &\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}. \qquad (3.40)\end{aligned}$$

Figure 1: Sampling from Standard Distributions.

Deduced Condlusion:

- There is a trade-off between bias and variance, with very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance.
- The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

28. What is a conjugate prior? Why we normally use conjugate prior? What is the conjugate prior for Gaussian distribution? (Refer to P117)

A prior is conjugate to the likilihood function, if the posterior distribution derived from the Bayes Theorem shares the same form (perhaps with different parameters) with prior distribution.

Use conjugate prior to simplify the Bayesian treatment of sequential learning process, in the sense that we have the same update formula at all iterations.

Conjugate prior of gaussian distribution is gaussian distribution.

29. What are the limitations of linear basis function models? (Refer to Textbook P173)

The basis functions are fixed before the data is observed. The chosen basis function may not adapt to the target function. And thus, the lower bound of the error may be quite large.

The number of basis functions needs to grow exponentially, with the dimensionality D of the input space.

30. What is the curse of dimensionality? Why it can be a problem? (Refer to Textbook P36)

The severe difficulty that can arise in high-dimensional space is called the curse of dimensionality.

Not all intuitions or even techniques developed in spaces of low dimensionality will generalize to spaces of many dimensions.

Why problem? the number of parameter in vector w we need to deal with is increased exponentially. Big limitation for the linear basis function model. 10 basis function in one dimension. 10^D in D dimensions.

By the way, in rejection sampling, there is another curse of dimensionality: the probability of rejection increases exponentially with regard to the dimensionality of the sampled space.

31. What are the three models for decision problems? How they are different?

Linear Discriminant. Non-probalistic model. Use discriminant function, directly map the input pattern x into class decision. (e.g. FDA, Perceptron algorithm.)

Discrminative Model. Probalistic model which directly model the posterior distribution $p(C_1|x)$. And make decision based on that posterior distribution. (e.g. logistic regression)

Generative Model. Probalistic model which first model the class-conditional probability distribution $p(x|C_k)$ and prior $p(C_k)$. And finally derive the posterior distribution $p(C_k|x)$ based on Bayes theorem. (e.g. naive bayes for discrete input with conditional independence assumption.)

Generally, the generative model is the most expensive model because of the large number of parameter in class-conditional distribution $p(x|C_k)$ to be fitted.

32. What are the deficiencies of the least squares approach in linear classification? (refer Textbook P185-186)

- Least-squares solutions lack robustness to outliers.
- In some multi-class classification problems that linear decision boundaries can give excellent separation between the classes, the least squares approach will misclassify most of the points from some classes.

33. What is the idea of Fisher's Linear Discriminant? Derive the fisher's linear discriminant that projects $x \in \mathbb{R}^D$ to $x \in \mathbb{R}^{D'}$ where $D > D'$.

ADD something here...

34. What is the perceptron algorithm? Describe it in detail.

Basic idea of perceptron algorithm is to minimise the number of misclassified patterns in two-class classfication problem.

Generalised linear model:

$$y(x) = f(w^T \Phi(x)) = +1 \ (w^T \Phi(x) > 0), -1 \ (w^T \Phi(x) < 0)$$

Use Particular coding scheme: $+1$ for C_1, -1 for C_2.

Error function is defined:

$$E(w) = -\sum_{n \in \mathbb{M}} w^T \Phi(x_n) t_n$$

Intuitively, correctly classified pattern will not contribute to the update, while misclassified pattern do. However, after the update, the correctly classified pattern in the past may turn to be wrong.

This algorithm guaranteed to converge, applied with stochastic gradient descent:

$$-w^{(\tau+1)}\Phi(x_n)t_n = -w^{(\tau)}\Phi(x_n)t_n - (\Phi(x_n)t_n)^T\Phi(x_n)t_n < -w^{(\tau)}\Phi(x_n)t_n$$

35. What is the probalistic generative model? Describe it in detail.

We have Naive Bayes for the generative model with discrete input. The Naive Bayes make a class conditional independence assumption: all features are independent on each other conditioned on the class.

36. What is logistic regression? Describe it in detail.

Logistic Regression, in essence, is a discrminative model to solve the classfication problem. It directly model the posterior distribution $y_n = p(C_1|x_n)$. Its likelihood function is

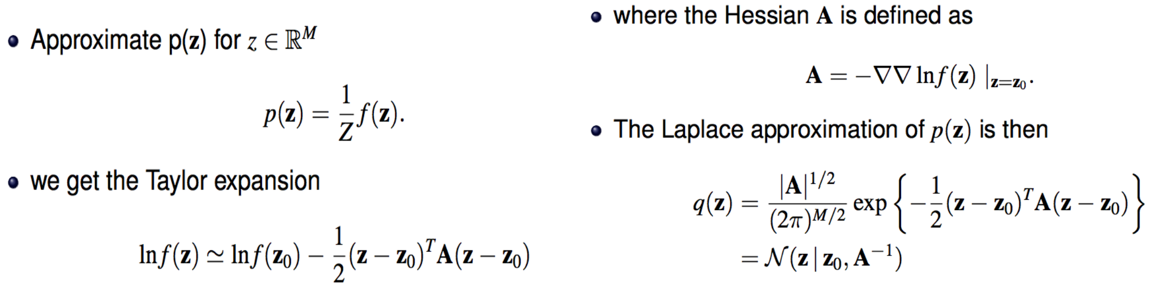
$$p(t, X) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

37. What is the feature mapping in classfication problem? Why we need this feature mapping sometimes?

Map input pattern from input space to feature space, on which the decision evaluation is based.

Sometimes, the patterns in input space is not linearly separable but in feature space it is linearly separable. Note that the overlapping problem cannot be solved by the mapping into feature space.

38. What is Laplace Approximation? Why we need to do Laplace Approximation sometimes? Describe in details. (Refer to Notes)



- Approximate p(**z**) for $z \in \mathbb{R}^M$

$$p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z}).$$

- we get the Taylor expansion

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)$$

- where the Hessian $\mathbf{A}$ is defined as

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z}) \mid_{\mathbf{z}=\mathbf{z}_0}.$$

- The Laplace approximation of $p(\mathbf{z})$ is then

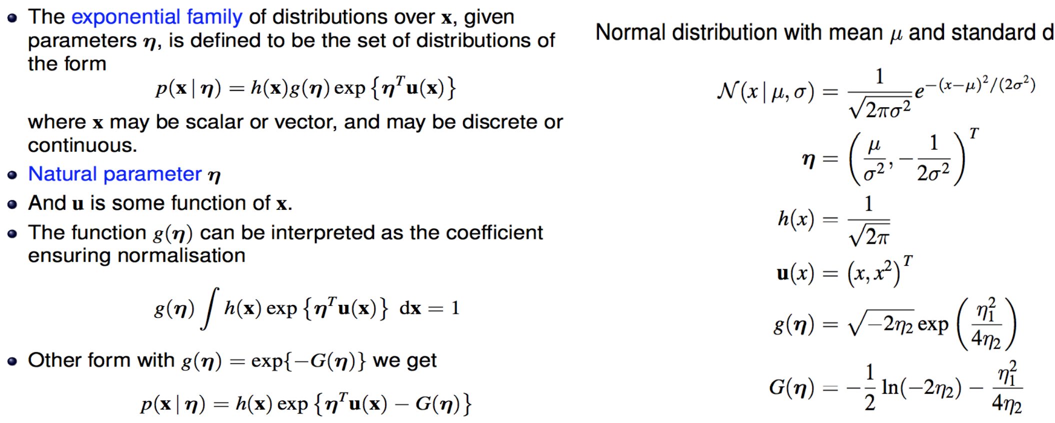
$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\}$$
$$= \mathcal{N}(\mathbf{z} \mid \mathbf{z}_0, \mathbf{A}^{-1})$$

Figure 2: Laplace Approximation in vector space.

39. Describe Baysian Logistic Regression in detail.

ADD something here...

59. Describe K-Means Clustering.
60. Describe EM algorithm for mixture of Gaussians in detail.
61. Show that Gaussian distribution belongs to exponential family.



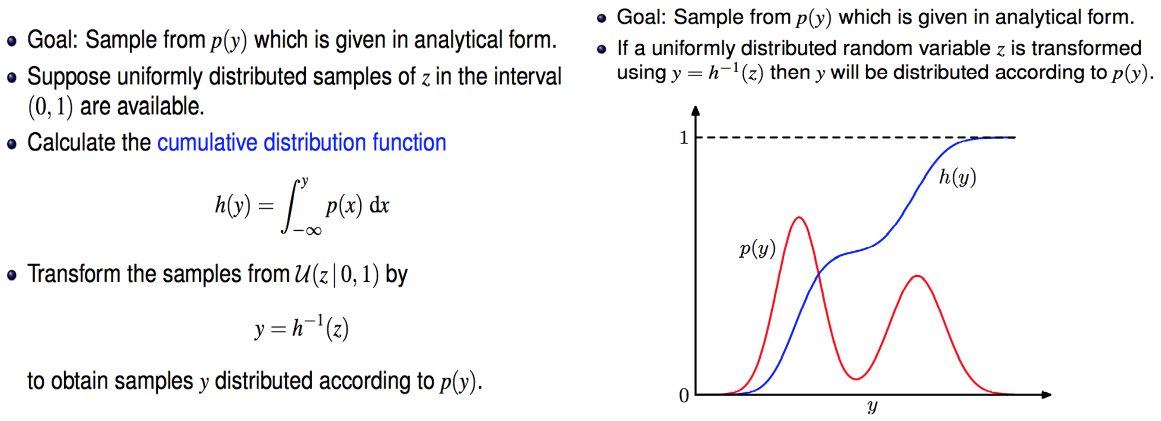
- The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form
$$p(\mathbf{x}\,|\,\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}$$
where $\mathbf{x}$ may be scalar or vector, and may be discrete or continuous.
- Natural parameter $\boldsymbol{\eta}$
- And $\mathbf{u}$ is some function of $\mathbf{x}$.
- The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient ensuring normalisation
$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$
- Other form with $g(\boldsymbol{\eta}) = \exp\{-G(\boldsymbol{\eta})\}$ we get
$$p(\mathbf{x}\,|\,\boldsymbol{\eta}) = h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}) - G(\boldsymbol{\eta})\right\}$$

Normal distribution with mean μ and standard deviation σ

$$\mathcal{N}(x\,|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$$
$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$$
$$h(x) = \frac{1}{\sqrt{2\pi}}$$
$$\mathbf{u}(x) = (x, x^2)^T$$
$$g(\boldsymbol{\eta}) = \sqrt{-2\eta_2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$
$$G(\boldsymbol{\eta}) = -\frac{1}{2}\ln(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}$$

Figure 3: Demonstrate Gaussian distribution belongs to exponential family. (eta η)

62. How do we sample from standard distribution? (Refer to 19_Sampling.pdf)



- Goal: Sample from $p(y)$ which is given in analytical form.
- Suppose uniformly distributed samples of z in the interval $(0,1)$ are available.
- Calculate the cumulative distribution function
$$h(y) = \int_{-\infty}^{y} p(x)\,\mathrm{d}x$$
- Transform the samples from $\mathcal{U}(z\,|\,0,1)$ by
$$y = h^{-1}(z)$$
to obtain samples y distributed according to $p(y)$.

- Goal: Sample from $p(y)$ which is given in analytical form.
- If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$ then y will be distributed according to $p(y)$.

Figure 4: Sampling from Standard Distributions.

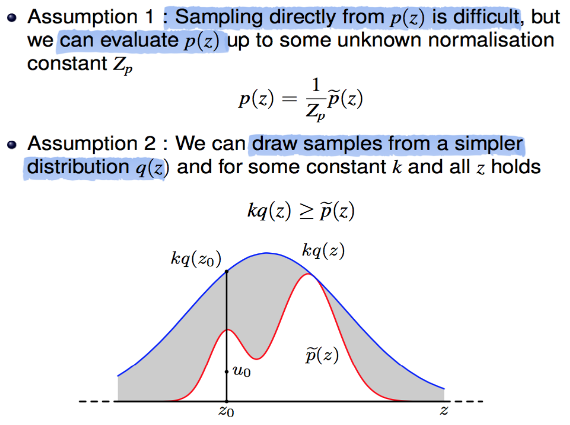
Last step is:

$$p_y(y) = p_z(z)\left|\frac{dz}{dy}\right|$$

63. Describe the method you would use to sample from the following distribution:

$$p(x=i) = \frac{1}{i} - \frac{1}{i+1},\quad i = 1,2,3,4...$$

64. Describe rejection sampling in detail. (Refer to 19_Sampling.pdf)



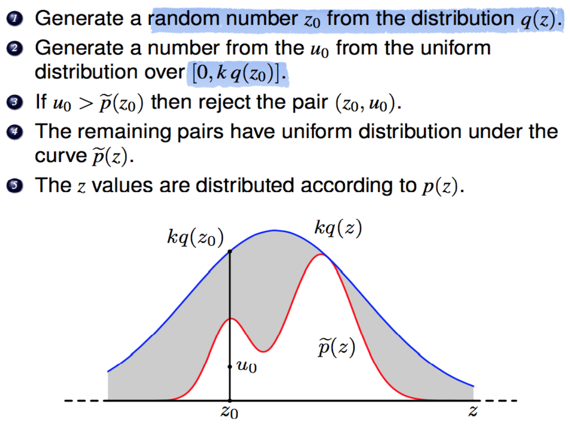
- Assumption 1 : Sampling directly from $p(z)$ is difficult, but we can evaluate $p(z)$ up to some unknown normalisation constant Z_p

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

- Assumption 2 : We can draw samples from a simpler distribution $q(z)$ and for some constant k and all z holds

$$kq(z) \ge \widetilde{p}(z)$$

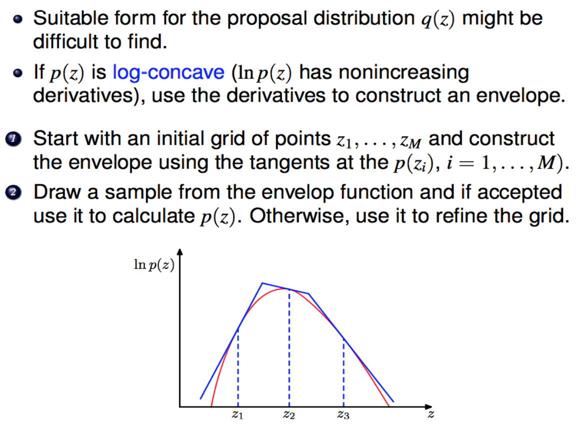
$kq(z_0)$
$kq(z)$
u_0
$\widetilde{p}(z)$
z_0
z



1. Generate a random number z_0 from the distribution $q(z)$.
2. Generate a number from the u_0 from the uniform distribution over $[0, k\,q(z_0)]$.
3. If $u_0 > \widetilde{p}(z_0)$ then reject the pair (z_0, u_0).
4. The remaining pairs have uniform distribution under the curve $\widetilde{p}(z)$.
5. The z values are distributed according to $p(z)$.

$kq(z_0)$
$kq(z)$
u_0
$\widetilde{p}(z)$
z_0
z

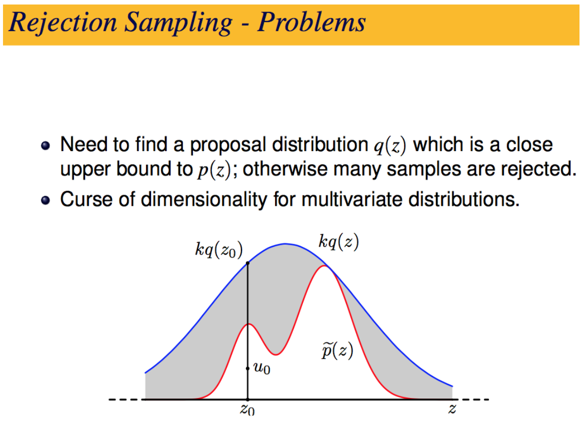
Figure 5: Rejection Sampling Details.



- Suitable form for the proposal distribution $q(z)$ might be difficult to find.
- If $p(z)$ is log-concave ($\ln p(z)$ has nonincreasing derivatives), use the derivatives to construct an envelope.

1. Start with an initial grid of points $z_1, \dots, z_M$ and construct the envelope using the tangents at the $p(z_i)$, $i = 1, \dots, M$).
2. Draw a sample from the envelop function and if accepted use it to calculate $p(z)$. Otherwise, use it to refine the grid.

$\ln p(z)$
z_1
z_2
z_3
z



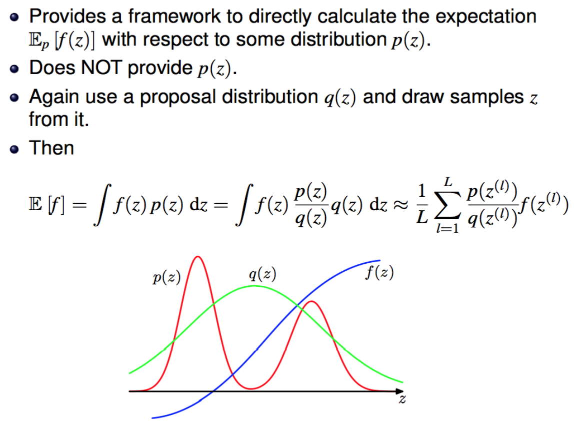
Rejection Sampling - Problems

- Need to find a proposal distribution $q(z)$ which is a close upper bound to $p(z)$; otherwise many samples are rejected.
- Curse of dimensionality for multivariate distributions.

$kq(z_0)$
$kq(z)$
u_0
$\widetilde{p}(z)$
z_0
z

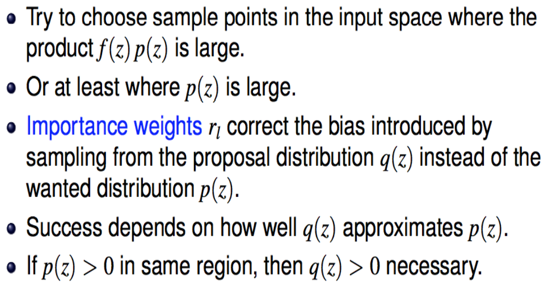
Figure 6: Adaptive Rejection Sampling. Problem of rejection sampling.

65. Describe importance sampling in detail.



- Provides a framework to directly calculate the expectation $\mathbb{E}_p[f(z)]$ with respect to some distribution $p(z)$.
- Does NOT provide $p(z)$.
- Again use a proposal distribution $q(z)$ and draw samples z from it.
- Then

$$\mathbb{E}[f] = \int f(z)\,p(z)\,\mathrm{d}z = \int f(z)\,\frac{p(z)}{q(z)}q(z)\,\mathrm{d}z \approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})$$



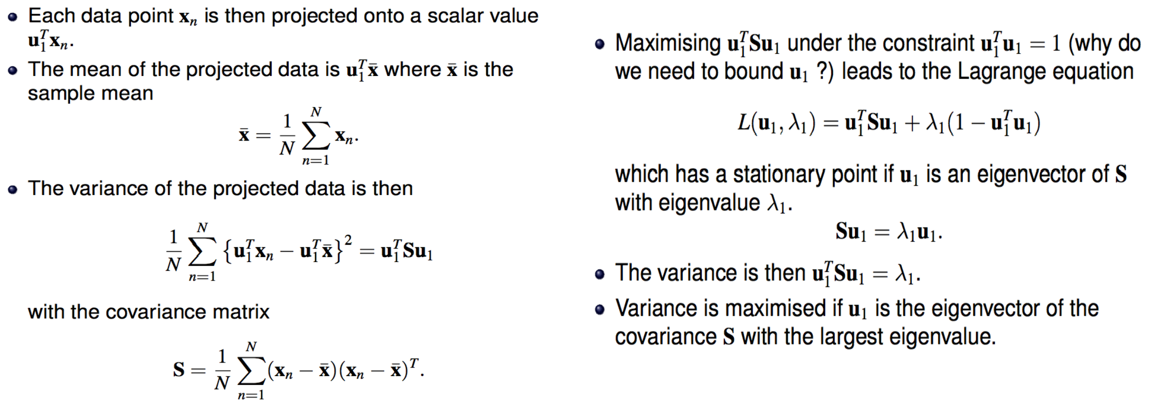
- Try to choose sample points in the input space where the product $f(z)\,p(z)$ is large.
- Or at least where $p(z)$ is large.
- Importance weights r_l correct the bias introduced by sampling from the proposal distribution $q(z)$ instead of the wanted distribution $p(z)$.
- Success depends on how well $q(z)$ approximates $p(z)$.
- If $p(z) > 0$ in same region, then $q(z) > 0$ necessary.

Figure 7: Adaptive Rejection Sampling. Problem of rejection sampling.

66. Describe Metropolis-Hasting algorithm in detail.
67. What is the idea behind principle component analysis?

- PCA is an algorithm for unsupervised learning.
- Linearly project the data points onto a lower dimensional subspace (principle subspace), such that the variance of the projected data is maximised.

- Functionality: Data Decorelation. Transform the coordinate of original space and apply the scaling on each coordinate axis, such that in new coordinate system, the mean of data objects is zero and covariance matrix is identity.



- Each data point $\mathbf{x}_n$ is then projected onto a scalar value $\mathbf{u}_1^T\mathbf{x}_n$.
- The mean of the projected data is $\mathbf{u}_1^T\bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the sample mean

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$

- The variance of the projected data is then

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right\}^2 = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$$

with the covariance matrix

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T.$$

- Maximising $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$ under the constraint $\mathbf{u}_1^T\mathbf{u}_1 = 1$ (why do we need to bound $\mathbf{u}_1$?) leads to the Lagrange equation

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1)$$

which has a stationary point if $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$ with eigenvalue λ_1.

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1.$$

- The variance is then $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1$.
- Variance is maximised if $\mathbf{u}_1$ is the eigenvector of the covariance $\mathbf{S}$ with the largest eigenvalue.

Figure 8: Principle Component Analysis.

68. What is a non-stationary sequential distribution? (Refer to Textbook P606)

- In the stationary case, the data evolves in time, but the distribution from which it is generated remains the same.

- For the more complex nonstationary situation, the generative distribution itself is evolving (varying) with time.

Note that in this course, we just focus on stationary distribution.

69. Use a Bayesian Network to represent a second-order Markov chain. (Refer to Textbook P608)

$$p(x_1, x_2, \cdots, x_n) = p(x_1)\ p(x_2|x_1)\prod_{n=3}^{N} p(x_n|x_{n-1}, x_{n-2})$$

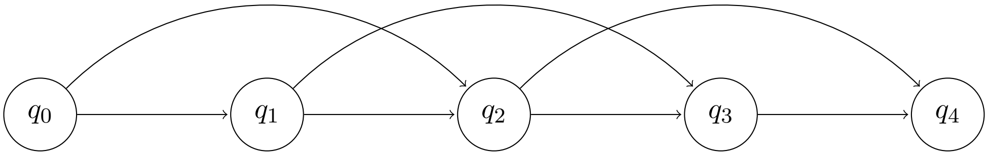


Figure 9: Bayesian Network to represent second-order Markov Chain

70. What is the difference between principle component analysis and Fisher's discriminant? (Refer to Textbook P568)

Similarity: Both methods can be viewed as techniques for linear dimensionality reduction.

Differences:

- Essence: PCA is unsupervised learning which only considers patterns $\boldsymbol{x}_n$. FDA is supervised learning also using class-label information.

- Basic Idea: FDA, given the labels, maximise the between-class covariance and minimise the within-class covariance. PCA consider the maximally reserved data variance (information).

71. Prove that independence of two random variable implies uncorrelatedness. Give counter example to show that uncorrelatedness does not imply independence. (Refer to http://mathforum.org/library/drmath/view/64808.html)

Independence: $P(x, y) = p(x)p(y)$. Uncorrelatedness: $cov(x, y) = 0$
Proof of $independence \rightarrow uncorrelatedness$:

$$\begin{aligned}
Cov(X,Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(x))(y-E(y))p(x,y)dxdy && (16)\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(x))(y-E(y))p(x)p(y)dxdy && (17)\\
&= \big(\int_{-\infty}^{+\infty}(x-E(x))p(x)dx\big)\cdot\big(\int_{-\infty}^{+\infty}(y-E(y))p(y)dy\big) && (18)\\
&= E(x-E(x))\cdot E(y-E(y)) && (19)\\
&= 0 && (20)
\end{aligned}$$

Counter-example:
Suppose X is a normally-distributed random variable with zero mean. Let Y = X^2. Clearly X and Y are not independent. However, The covariance of X and Y is

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0 * E(Y) = E(X^3) = 0$$

72. Assuming K different states for each variable x, how many parameters does M-order Markov chain have? (Refer to textbook P609)
$K^M(K-1)$. K is number of discrete assignments. K^M is the number of possible composite assignments of M conditioned variables. $K-1$ is for each possible assignment, the number of free parameters (given the normalisation property).
Because this grows exponentially with M , it will often render this approach impractical for larger values of order M.
Note that we only talk about stationary case (each factor shares the same probability distribution).
73. What is a homogeneous hidden Markov Model? (Refer to textbook P612)

$$p(x_1, \cdots, x_n, z_1, \cdots, z_n) = p(z_1)\cdot\prod_{n=1}^{N}p(z_n|z_{n-1})\cdot\prod_{n=1}^{N}p(x_n|z_n)$$

Homogeneousness in hidden markov model means that

- all the transition probabilities $p(z_n|z_{n-1})$ share the same parameters A. (A is transition matrix from 1-of-k coding)
- all of the emission distributions $p(x_n|z_n)$ share the same parameters Φ.

The probability distribution of latent variable is **invariant or stationary** if the **Detailed balance** is satisfied. (also called **reversible**.)

$$p^*(z)T(z, z') = p^*(z')T(z', z)$$

74. Describe Viterbi algorithm in detail. to find the most probable sequence of hidden states for a given observation sequence. this can be solved efficiently using the max-sum algorithm the problem of finding the most probable sequence of latent states is not the same as that of finding the set of states that are individually the most probable. max-sum algorithm works with log probabilities and so there is no need to use re-scaled variables The Viterbi algorithm searches this space of paths efficiently to find the most probable path with a computational cost that grows only linearly with the length of the chain.
75. What are the motivations for combining models? (refer to Lecture Notes)

- Motivation: Coming up with a very precise prediction rule can be very hard. It may be easier to come up with a number of not so precise prediction rules.
- Basic Idea: Combine models with different rules to make better prediction.

76. Explain differene between model combination and bayesian averaging.

Model Combination: different data points within the data set can potentially be generated from different values of the latent variable z and hence by different components.

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z})$$

Bayesian Averaging: every hypothesis is responsible for generating the whole set of data.

$$p(\boldsymbol{X}) = \sum_{h} p(\boldsymbol{X}|h)p(h)$$

Distinction Questions:

1. explain how neural network function? including the back propagation with formula.

2. compare online processing

3. explain graph separation in markov random field.

4. explain gaussian mixture model. How do you do inference in gaussian mixture model.

5. explain how hidden markov model is an extended markov chain.

we wish to build a model for sequences that is not limited by the Markov assumption to any order and yet that can be specified using a limited number of free parameters. We can achieve this by introducing additional latent variables to permit a rich class of models to be constructed.

HD Questions:

6. explain three models for classification.

7. What is bias-variance trade-off.

8. what is Fisher discriminant analysis.

HHD Questions:

9. Explain how variational optimisation can be used in finding approximation for posterior distribution.