**Learning and search methods:**

1. A loss function or cost function is a function that maps an event or values of one or

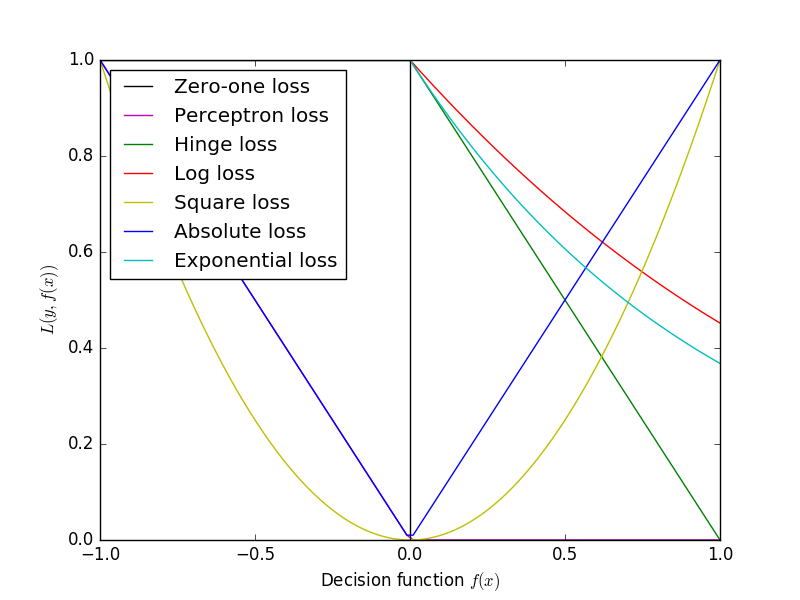
more variables onto a real number intuitively representing some "cost" associated with the event.

Loss function:

Hinge:

Logistic:

Least square:??



1. L1 norm and L2 norm

3.1

3.2 The goal is that, with high probability (the "probably" part), the selected function will have low generalization error (the "approximately correct" part).

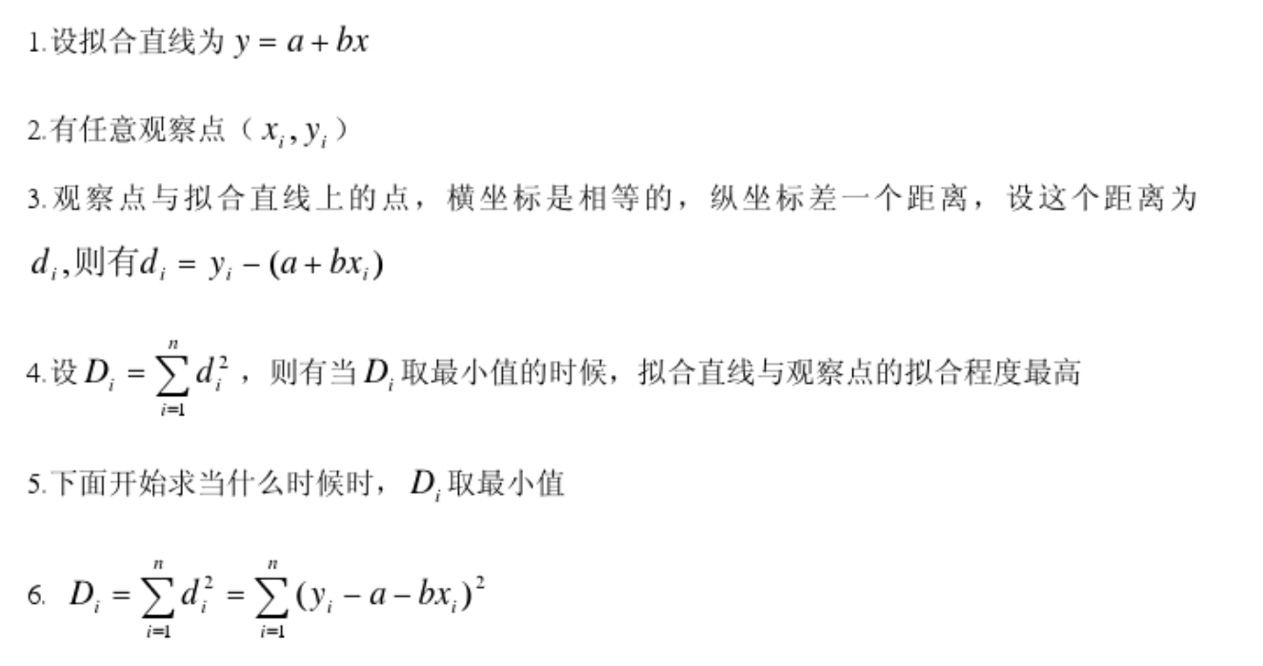
1. Bias: The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

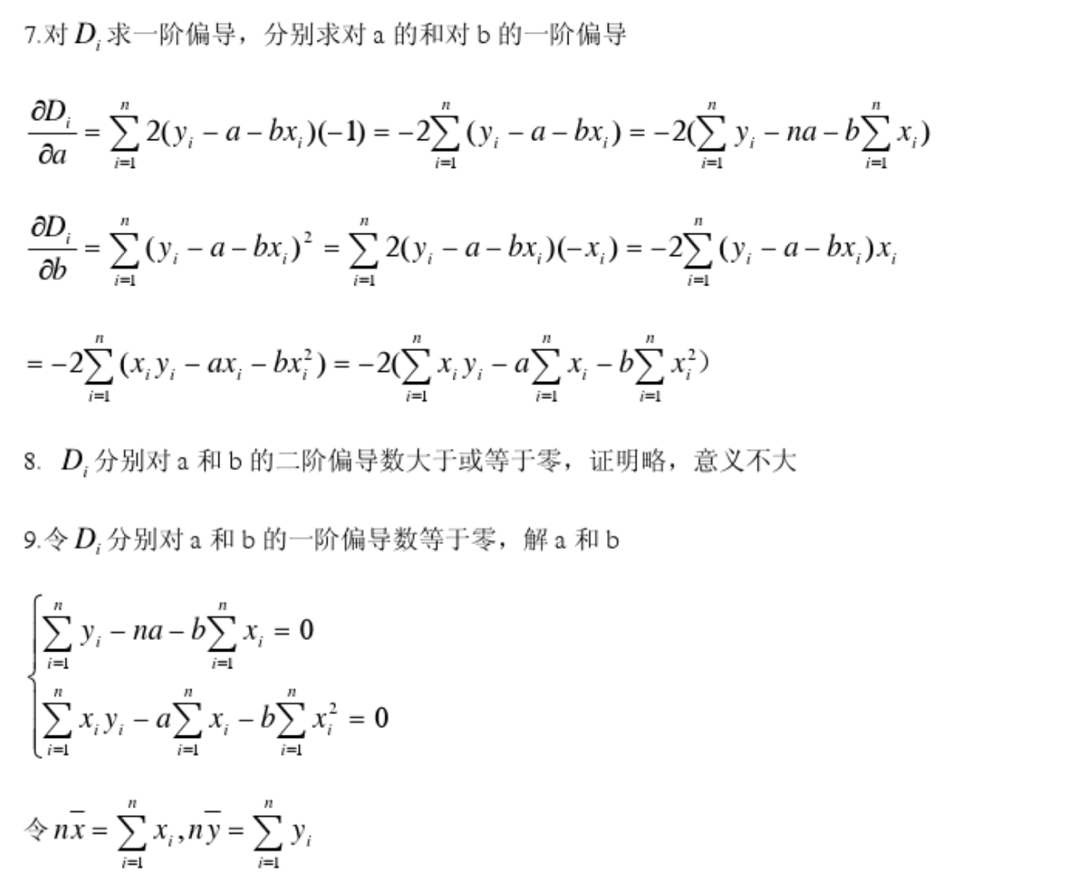
Variance: The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

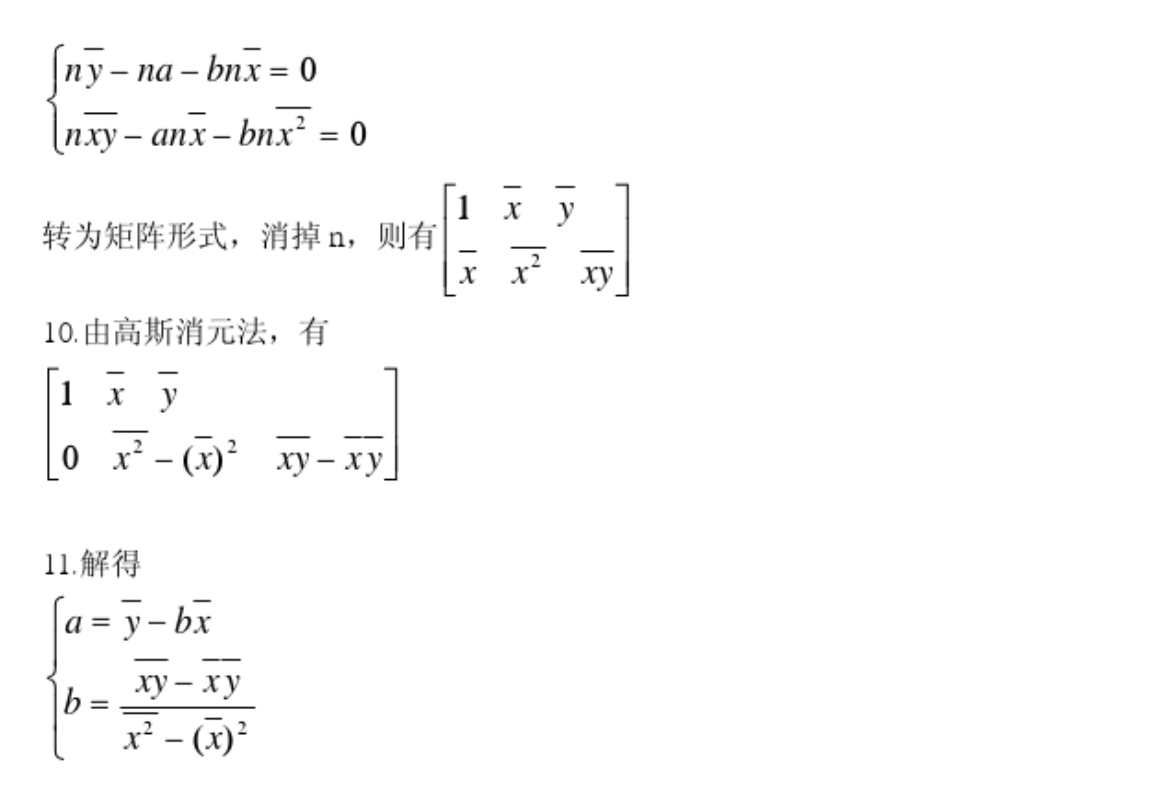
Select when the total error of biases and variances is lowest.

1. Linear regression: when (bias)^2+variance is min

Logistic regression:

1. 





1. ??
2. The learning rate of the algorithm 影响算法效率;

The initial value of parameter, 初始值不同，最小值有可能不同，可能只是局部最小值;

normalization,

1. ??
2. The logistic regression satisfy the Bernoulli distribution.

The objective function of MLE is convex. It is easy to gradient descent to get optimal solution.

1. Discriminative machine learning is essentially training a model to distinguish the correct output among possible output choices. This is typically done by learning model parameters that maximize the conditional probability P(Y/X). 异类数据之间的差异

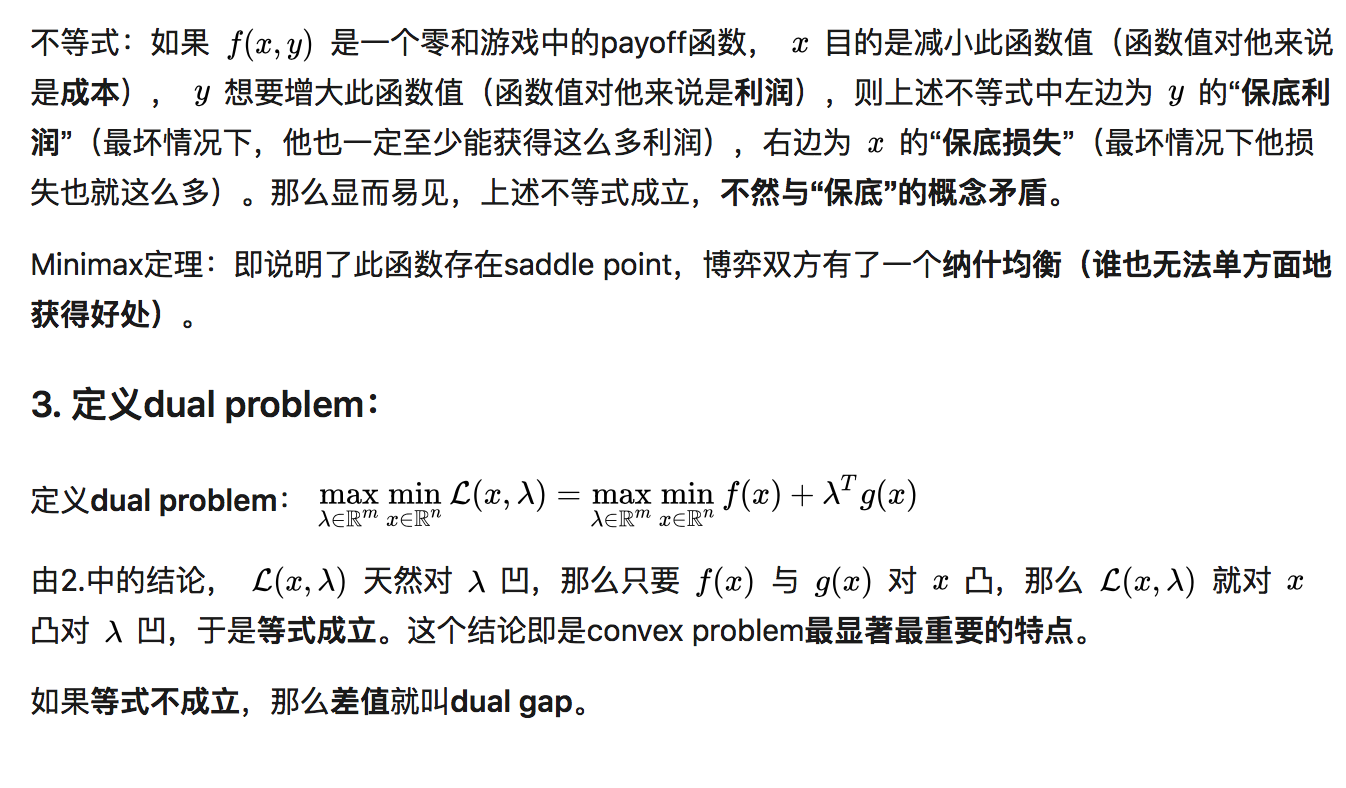
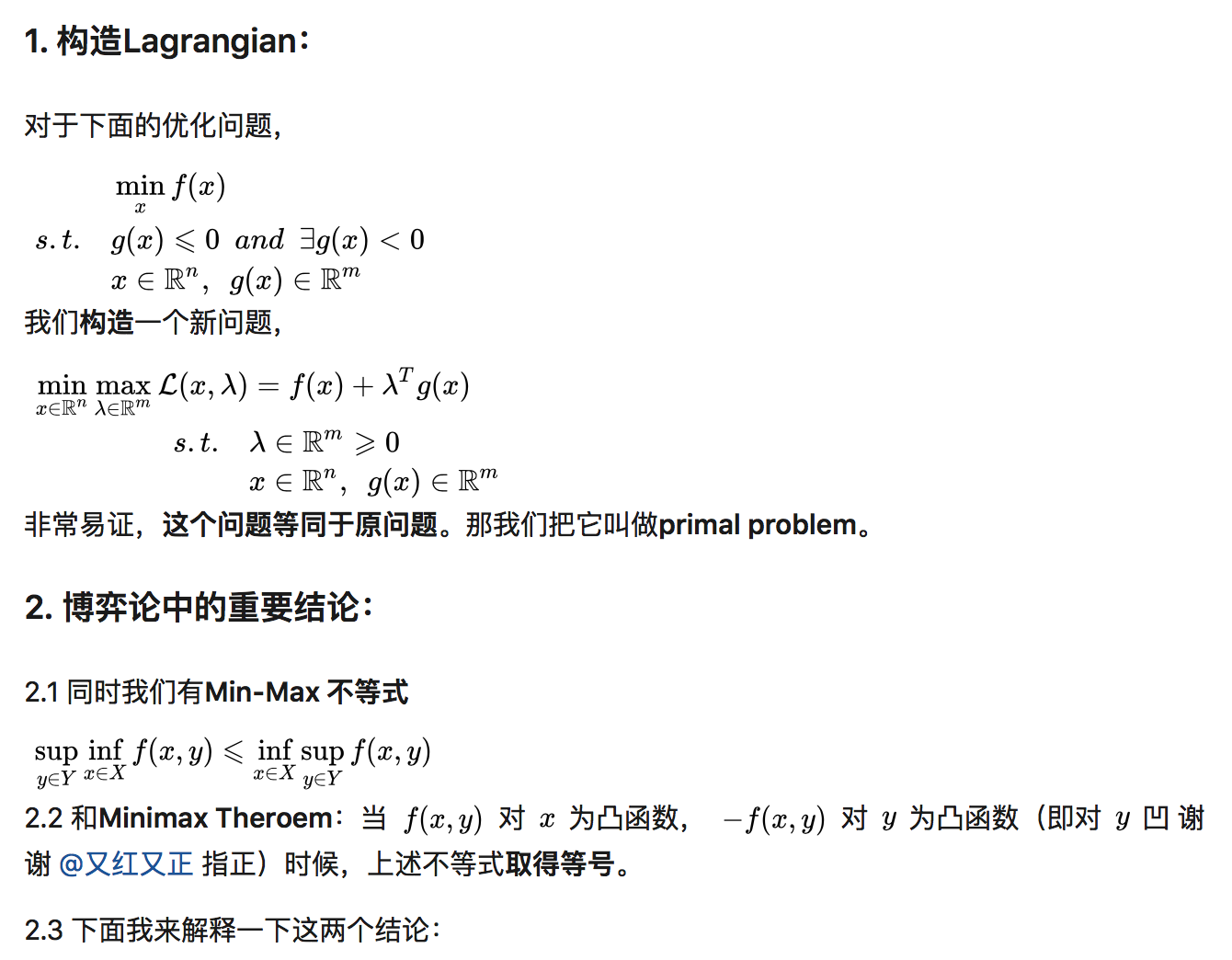
Generative machine learning is training a model to learn parameters maximizing the joint probability of P(X,Y). 同类数据之间的相似度

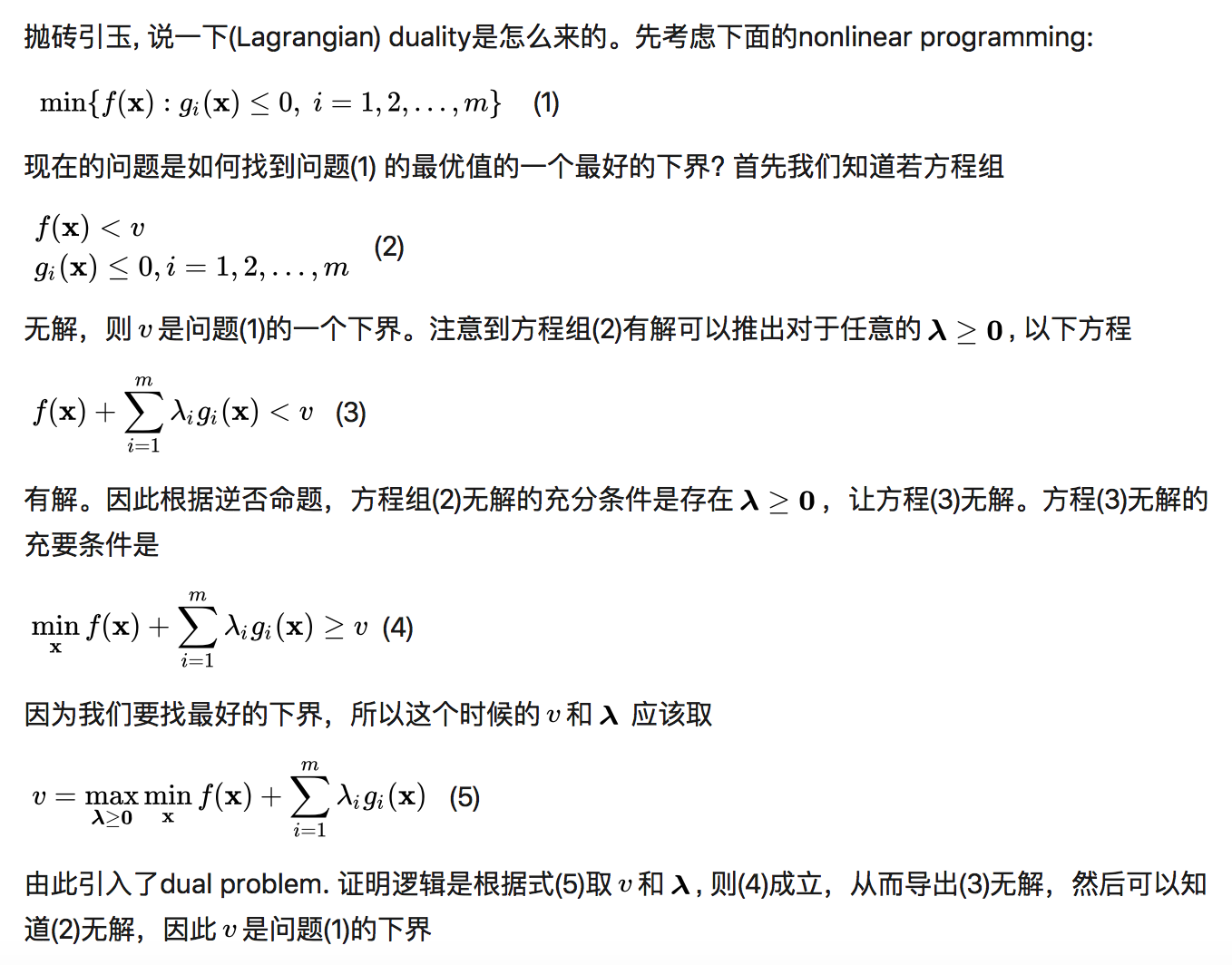
1. ??
2. The parameter of linear models is fewer.

It cannot explain the process of reasoning; cannot work if data size is small

1. ??
2. dual problem is obtained by forming the Lagrangian, using nonnegative Lagrange multipliers to add the constraints to the objective function, and then solving for some primal variable values that minimize the Lagrangian.

Strong duality is a condition in mathematical optimization in which the primal optimal objective and the dual optimal objective are equal.





1. Karush–Kuhn–Tucker (KKT) conditions, also known as the Kuhn–Tucker conditions, are first-order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.

The extreme point of the optimization problem will definitely fix this condition. The necessary condition that the optimization problem has extreme point.

SVM will introduce Lagrangian problem and will have relation with KKT.

1. Some sample of soft margin SVM may not satisfy

but this kind of sample should be minimized, thus

changing the loss function

The regularization of soft margin SVM is

L2 norm tends to have balanced components of w, more non-zero components

L1 norm and L0 norm tends to have scarce components of w, less non-zero components

1. Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space.

Because it can directly calculate the inner products of images , without knowing separately.

1. Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space.

Because it is easy to calculate.

1. Because sometimes the samples in a low-dimensional space are not linear separable, they are only linear separable in higher dimension, so we need a higer-dimensional space.

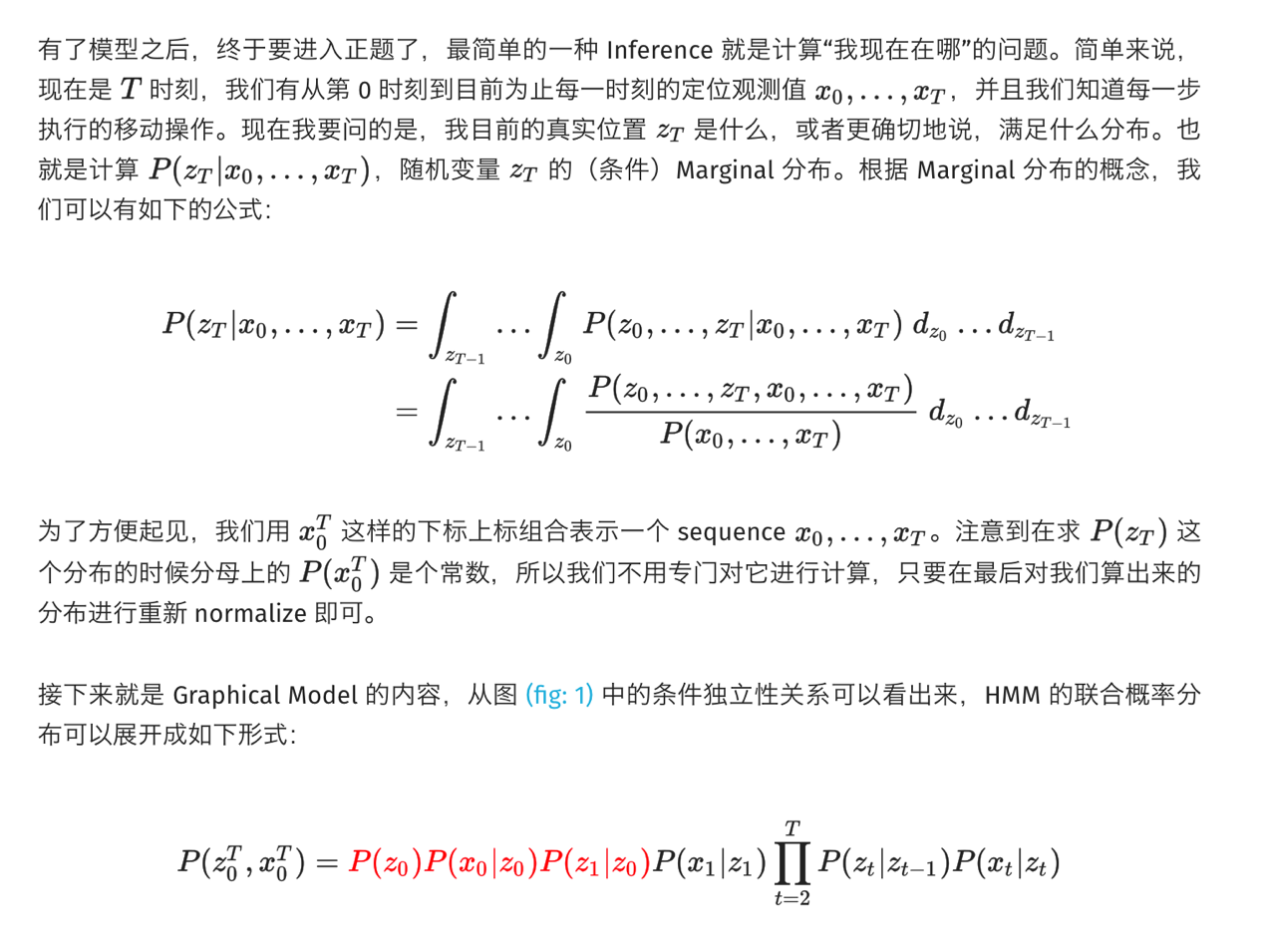
**Probabilistic graphical model:**

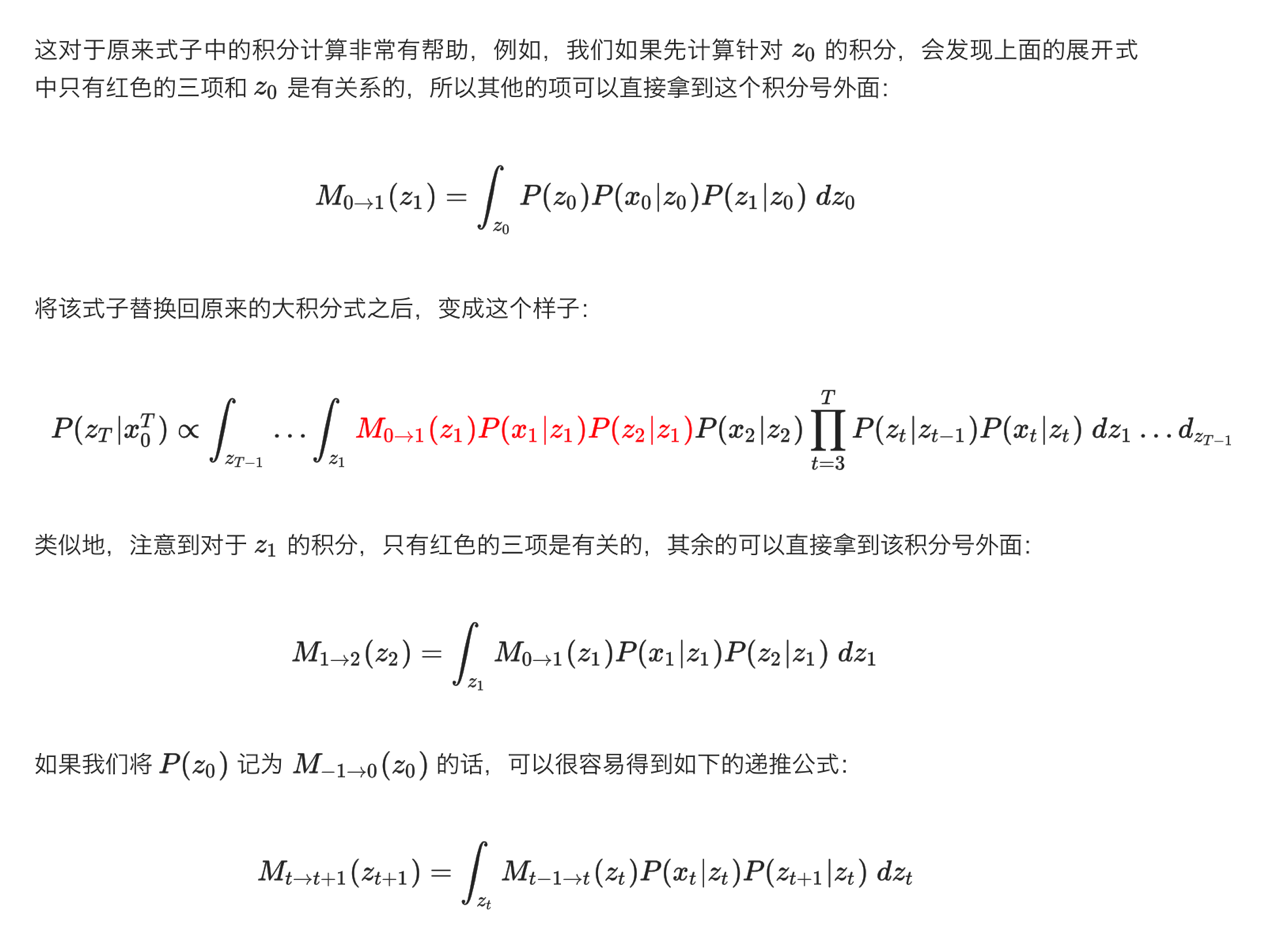
1. ?
2. ?
3. ?
4. ?
5. If the function we use in EM is not convex, we may find the local solution.

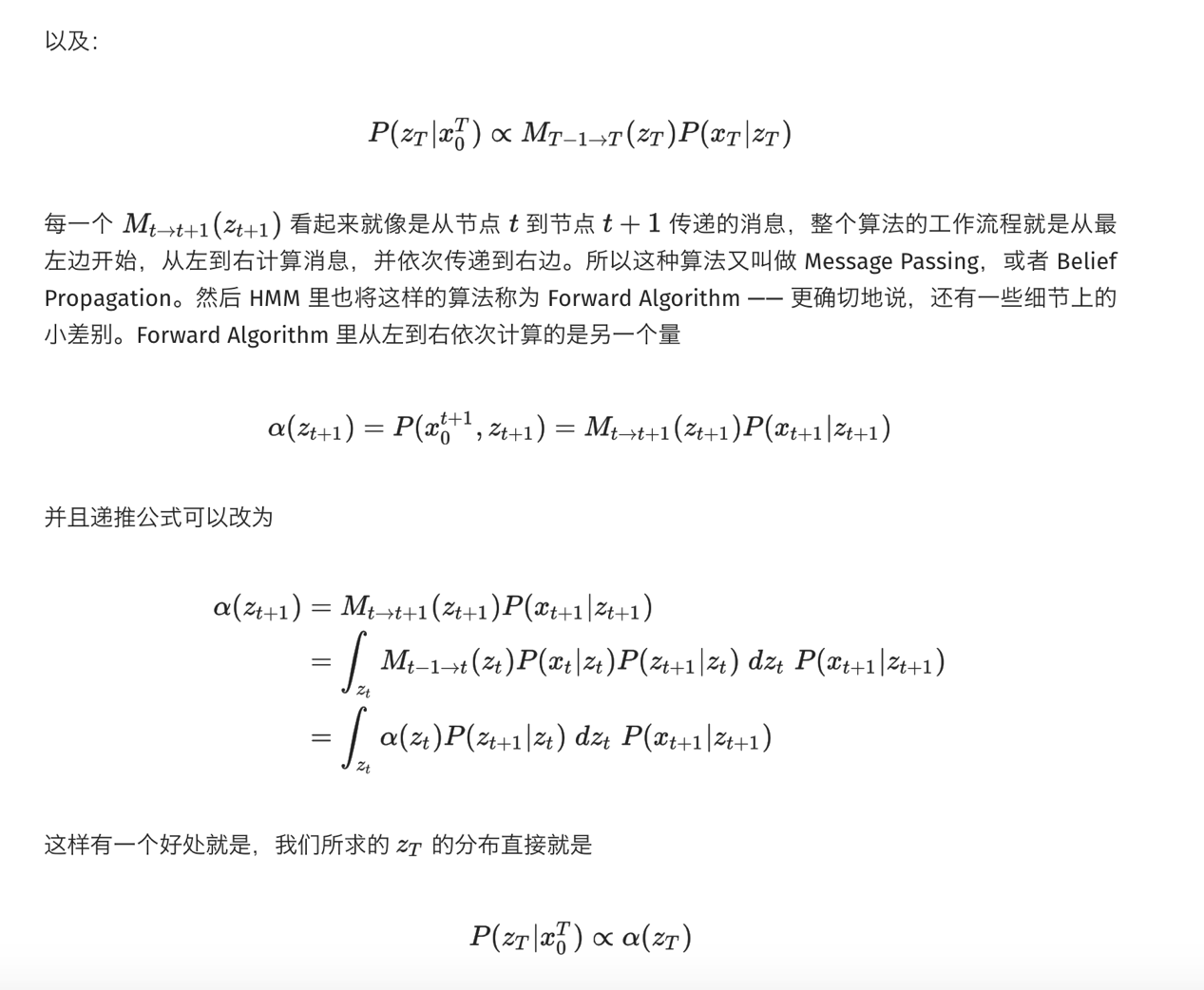
Proof:?

The goal of k-means algorithm is to find the minimum of E, however, it is very difficult because we must know all the clustering of D (sample set). So k-means algorithm is based on greedy algorithm. Since greedy algorithm may not find the best solution, it only finds the local solution.

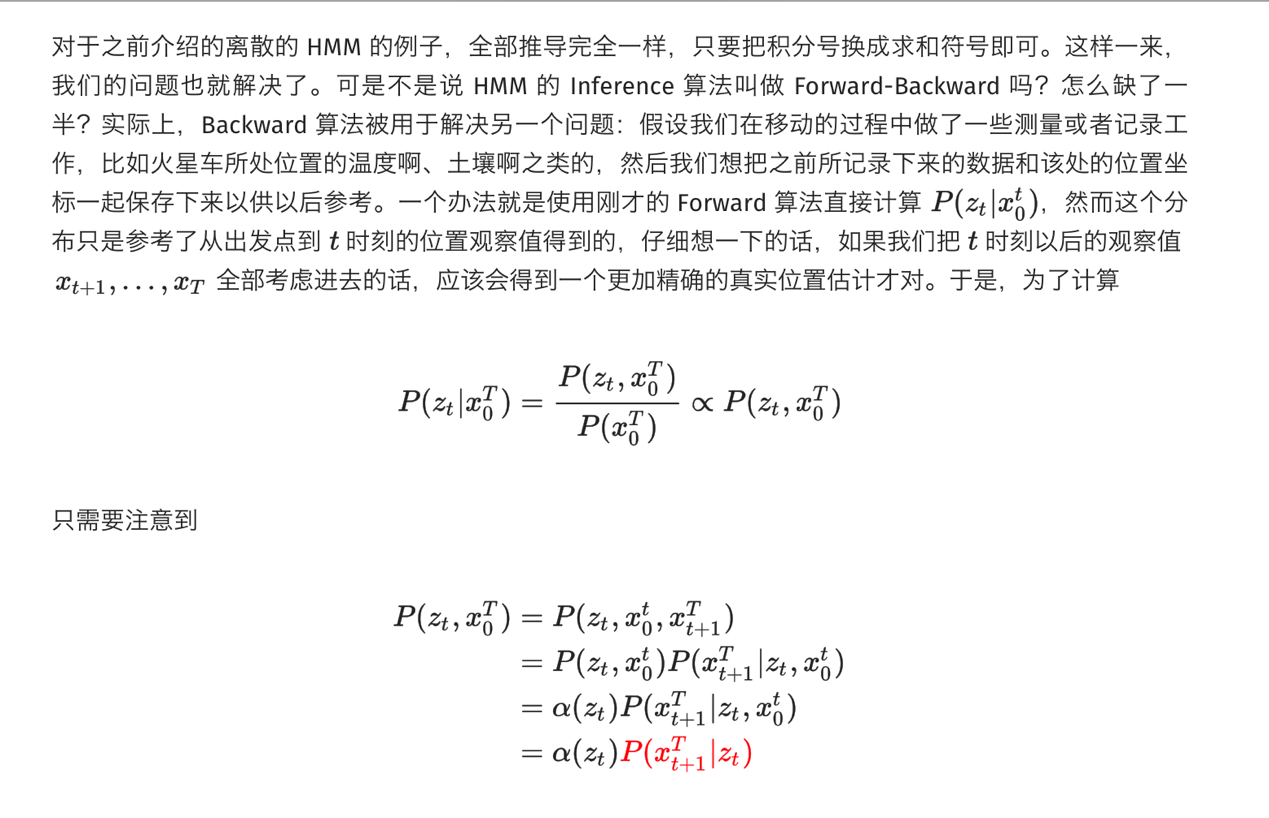
1. Forward:

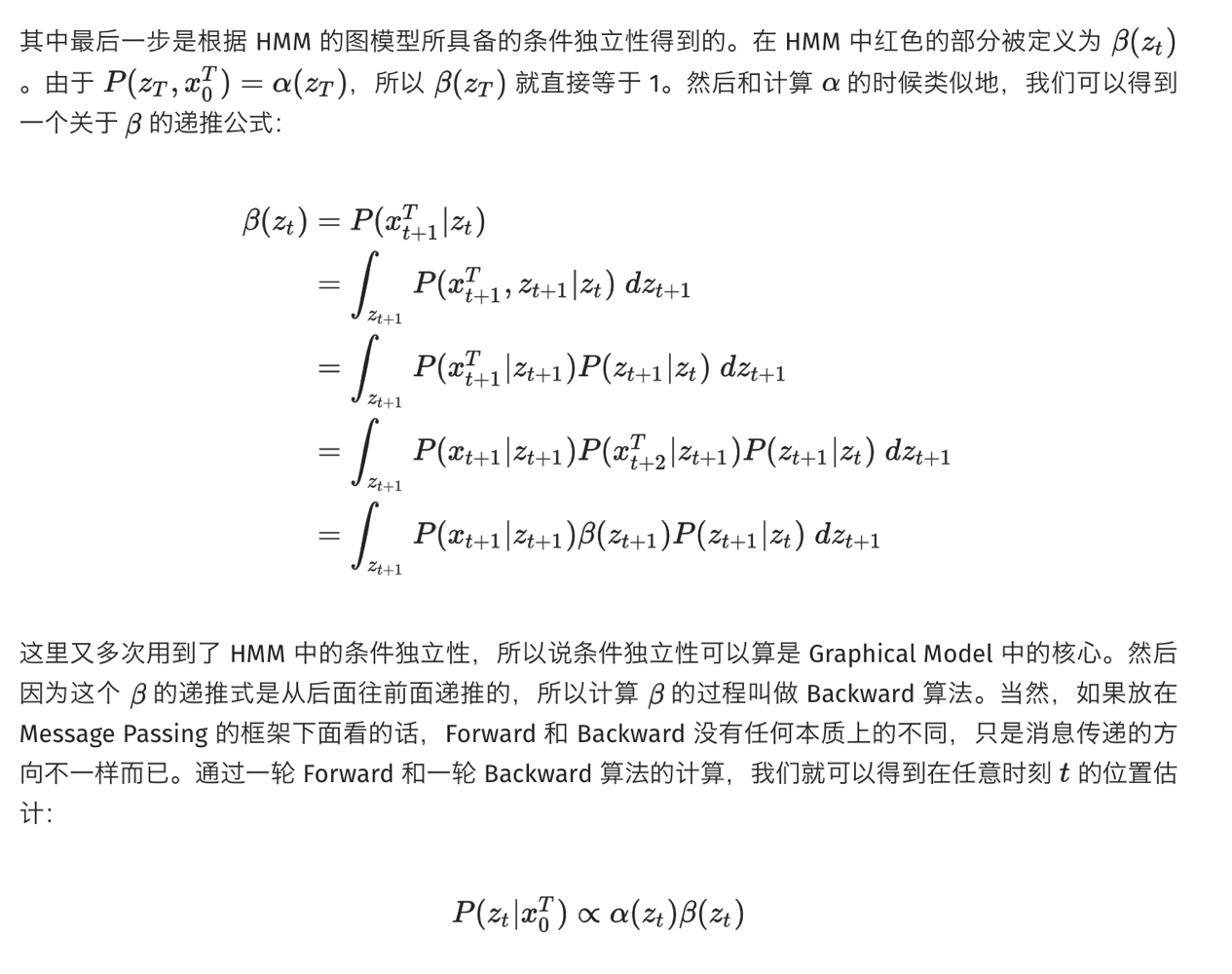


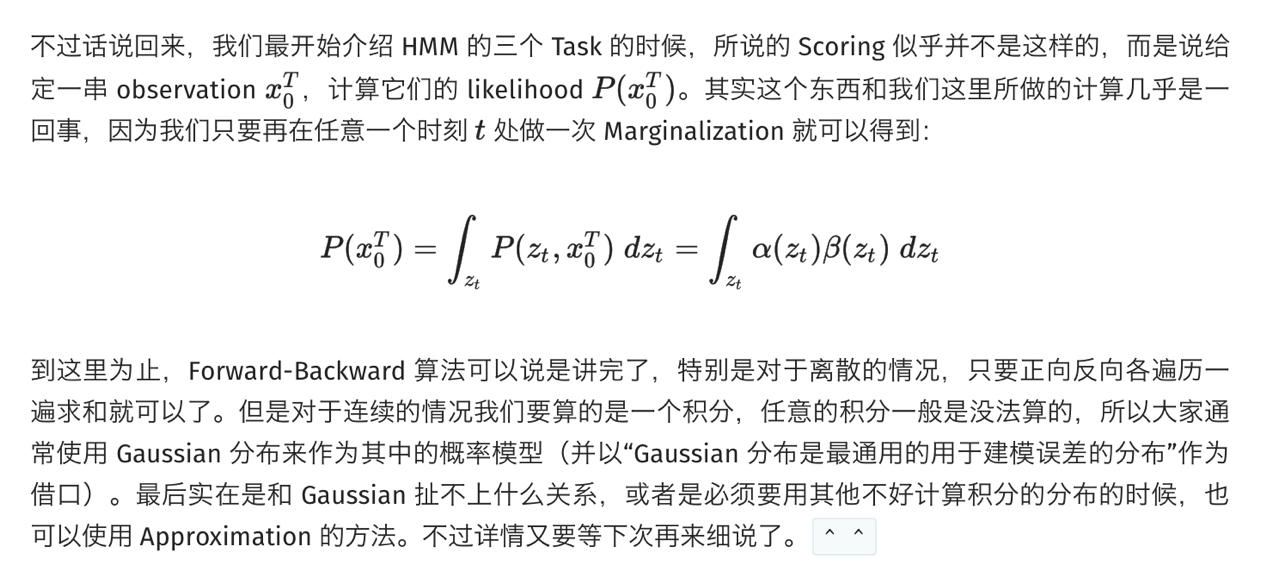




backward:







1. See Question 7
2. ?
3. ?

**Dimension reduction and feature representation:**