1. **Why all learning problems are inverse problems, requiring unbounded exhaustive searches, thus ill-posed?**

* We don’t actually know the model and we would like to achieve something through some kind of search and optimize it to some situation. We also want the ability we gain from search to be generalizable.
* In order to make sure there is an existing solution, we need to search through unlimited alternative mathematical frameworks. Even though given a framework, we need to make sure there is a unique solution and is the best through search. And the frameworks, dimensions and data involved are unbounded. Moreover, we need to make our model be generalizable, so that it can be applied to the data we have never seen while training. Thus, we will never know whether the model is stable and generalizable forever.

1. **Why gradient is the key mathematical assumption that we could count on in order to search? What would be the general implications for the existence of at least some continuality or locality?**

* Gradient provide us a shortest path from the present state to the optimal, otherwise we have to exhaust all the parameters to fit the model. And gradient is the only sign showing the process of convergence.
* We assume the processes we are modeling are something classic, as Newton described. There are momentum and continuation, instead of jumping. This makes the way of searching much easier.

1. **What is the generalizability of a mathematical process, from both e****xpressive (smoothness) and inclusive (capacity) point of views?**

* expressive (smoothness): we want the input be described as detailed as possible and contains more variability. For example, instead of using 0-1, we use sigmoid so that we can learn something more than 0-1 like the overlapping.
* inclusive (capacity): we want the input be expanded to a wider space, thus we can capture the universal framework which can be applied to all the inputs.

1. **What make a mathematical process more generalizable in terms of both structural representation and model complexity?**

* Structural:

1. Right Space. A more even/fair space which is normalized and like a Euclidian space.
2. Proper Measurement.

* model complexity:

sparsity, low rank, dimension reduction, low complexity solution……（我TM真不懂啊）

1. **What would be some of the solutions for such an ill-posed problem in order to yield at least some reasonable (generalizable) results?**

* Constraint search: L1, L2 Norm
* Dimension Reduction
* Markov assumption
* Graphical models
* Normalization, Regularization

1. **What are some of the mathematical hurdles that have prevented us from obtaining more generalizable solutions?**

* We can not guarantee that we are dealing with something Euclidian. However, all out mathematical systems are established based on Euclidian.
* Increase dimension: not only leads to a more complex model but also some local solutions
* Interaction/Concordance/Parallelism: We don’t have a way to truly resolve a truly concordant problem, we need to take some orders. We cannot max x and y at the same time in a parallel way

1. **Why variable dependences (interactions) could become an extremely difficult and even an impossible problem? Give philosophical, mathematical, physical, computational, and numerical examples for such a singularity.**

* Dependences (interactions) are very universal, but we can not solve the interactions at the same time. This is the limit of math and computing. We don’t have a mathematical framework to deal with interactions.（YB认为这是数学本身的局限性还是啥）
* Philosophical: Russell's paradox
* Mathematical: Ferma's theorem
* Physical: Three body problem
* Computational: Turing halting problem
* Numerical: sensitive problem/initial value problem

1. **Why a Euclidian-based measure would be most favored but usually impossible to obtain for a real world issue?**

* All our numerical methods are derived in Euclidian space: [differential](javascript:;), space, [geometry](javascript:;), norm, distance
* We need a standard and justified space to form the problem in continues/dynamic way, instead of describing things purely structural and logical.

1. **What are some of the key requirements for a real issue to be formulated as a Euclidian problem?**

* Dimension: Features need to be Orthogonal(正交)
* Space: space can be projected to low/high/curve space
* Measurement: how individual features can be compared among each other in a normalized way

1. **What would be the mathematical alternative frameworks to translate a non-Euclidian problem to mathematically appropriate solutions?**

* The orthogonal normalization(正交归一)
* Model selection, feature selection
* Dimension reduction：PCA ect.
* Projection: To low/high dimension space or even curved space

1. **Why in general the complex and high-dimensional data (the so-called big data problem, n<<p) from the same “class” tend to have a low dimensional representation?**

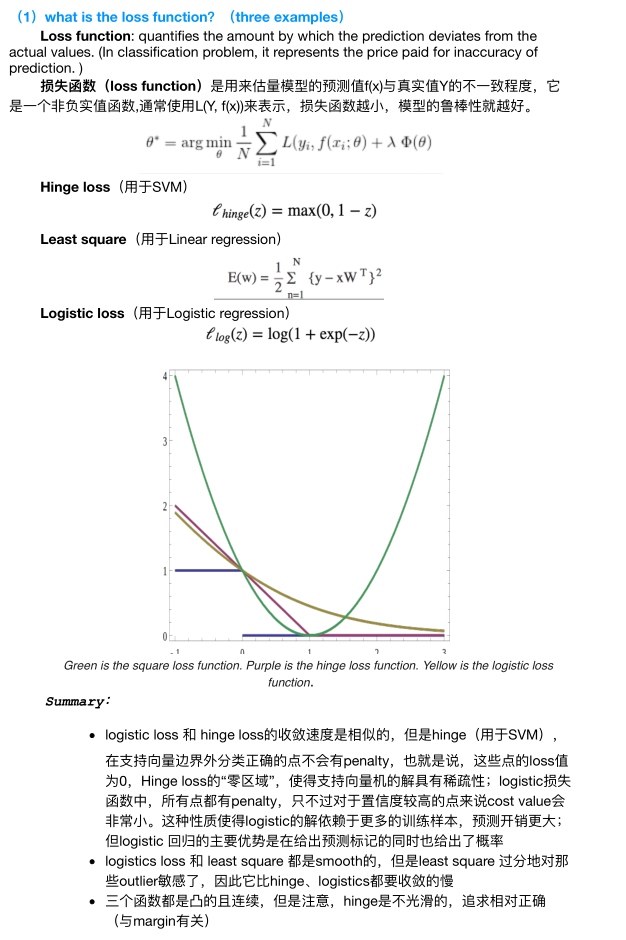
* There is a boundary O(logP). Features have nature to cancel each other so that only logP become outstanding
* Data may be redundant: Even a picture have many pixels, we humans can’t sense all of them, thus many of them are redundant.

1. **Why we would prefer a low complexity model for a high complex problem?**

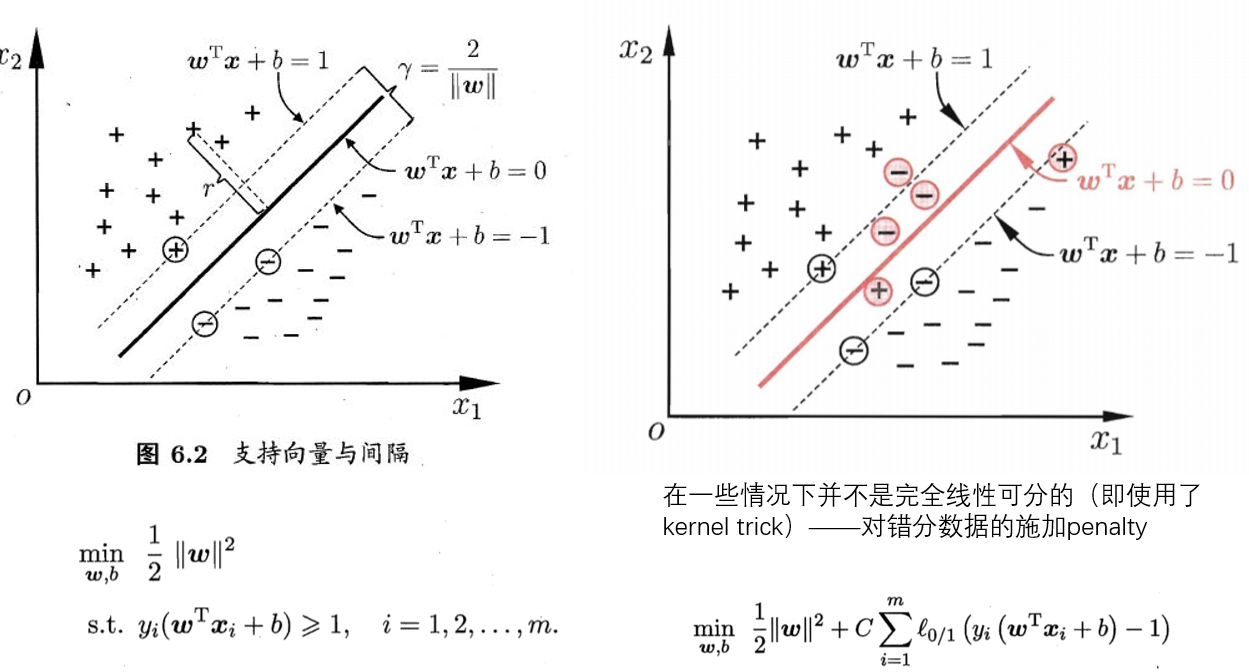
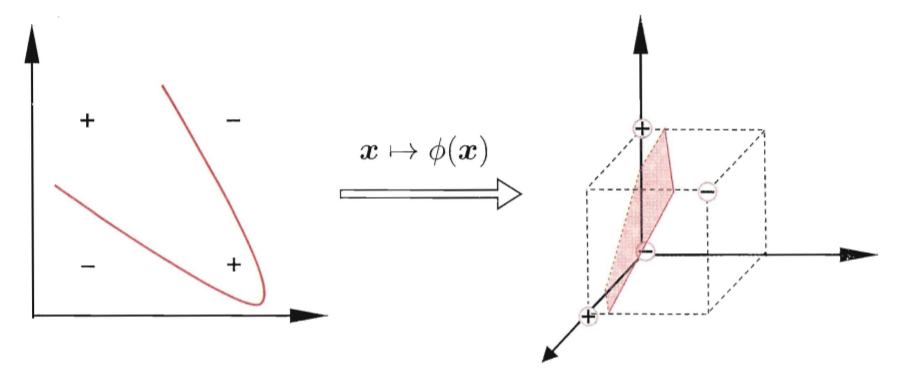
* Low complexity models are less likely to be overfitting to training data(may include noises). Lowering the model complexity makes the error increase linearly but raising the complexity makes the error increase exponentially
* We want our model to capture the main/ principle feature instead of the too detailed, so that our model can be robust.
* Easier to compute: less time and space

1. **What is a loss function? Give three examples (Square, Log, Hinge) and describe their shapes and behaviors**

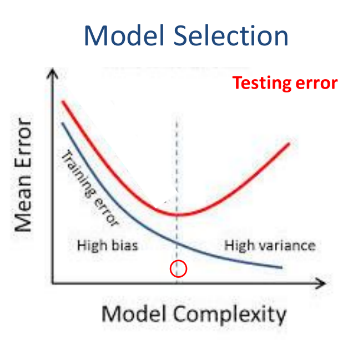
* Square(Used on regression): based on gaussian, evaluate the Euclidian distance, encourage the process in a polynomial way, sensitive
* Hinge(Used on SVM): Linear, encourage the process to be optimized equally, most robust
* Log(Used on Logistic Regression): based on MLE of probability distribution, more robust than square



1. **Using these losses to approach the linear boundary of an overlapping problem, inevitably some risks will be incurred; give two different approaches to remedy the risk using the SVM-based hinge loss as an example.**

* Soft margin SVM，对于错误分类的数据施加penalty, C控制margin，C越小margin越大（可能造成underfitting）
* Kernel trick(depend on data itself)—project the low dimensional data to high dimensional space（可能造成overfitting, very sensitive）
* SVM itself is a dimension reduction process, throwing some non-efficient dimension and data

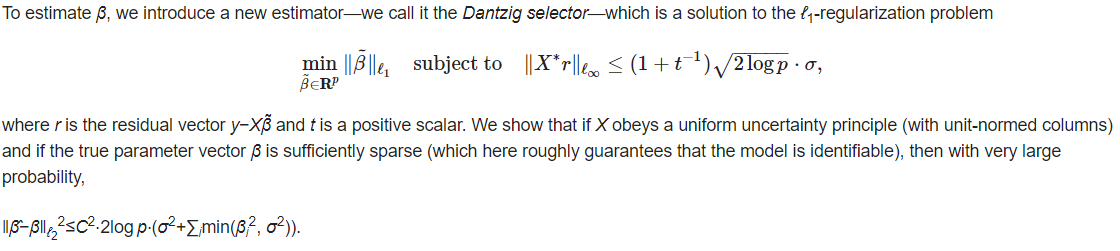
1. **Describe biases (under-fitting) and variance (over-fitting) issue in learning, and how can we select and validate an appropriate model?**

* When underfitting the error decreases linearly, but when overfitting the error increases exponentially, thus overfitting is more risky. (并不知道有啥依据，但是YB反复强调)
* Cross Validation

将数据分成K份，对于每个模型用其中的K – 1份进行训练，剩下1份进行测试，将测试的error求和，对不同模型的求和error进行比较，选取最小的。

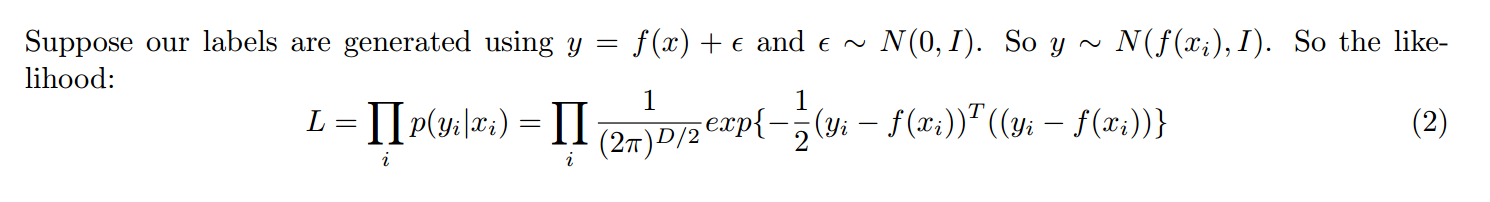
* 即使使用了Cross Validation也不能保证找到最好的结果，因为所选取的模型范围还是有限的，因此问题还是ill-posed and exhausted的。

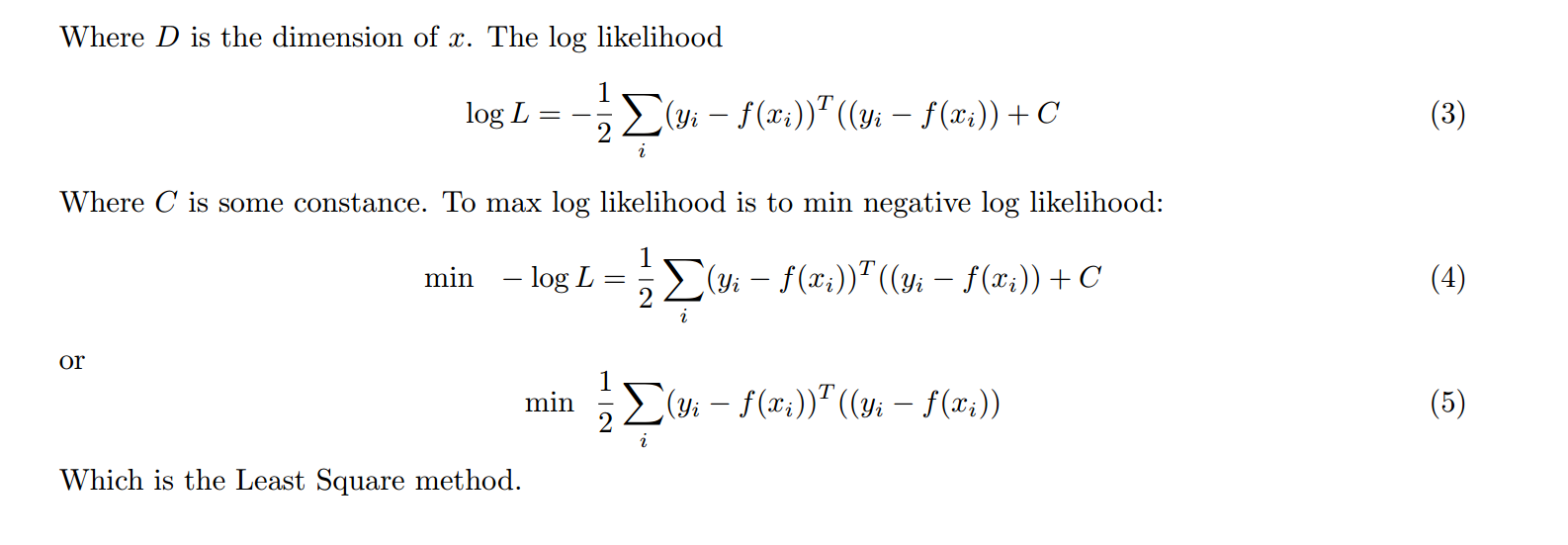
1. **How to control model complexity in the regression of a linear model? Are there supposed to be a unique low-dimensional model for a given high dimensional problem?**

* Control model complexity：use L1 and L2 norm, especially Lasso
* We can’t tell where the unique solution is but there exists an error bound. In practice, we still need to search through several to determine the parameter.

Candes, Emmanuel; Tao, Terence. The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35 (2007), no. 6, 2313--2351.

1. **Using the Least Square as the objective function, we try to find the best set of parameters; what is the statistical justification for the Lease Square if the underlying distribution is Gaussian?**





1. Could you describe the convexity as to how it would facilitate a search? Using the Least Square-based regression and Likelihood-based estimation as the examples?

* If the problem is convex, then through gradient-based search, no step is redundant and every step is toward the convergence. We can find a shortest and guaranteed path toward the solution.
* 第二问没讲

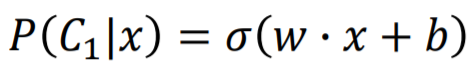
1. **Gradient Decent has a number of different implementations, including SMO, stochastic methods, as well as a more aggressive Newton method, what are some of the key issues when using any Gradient-based searching algorithm?**

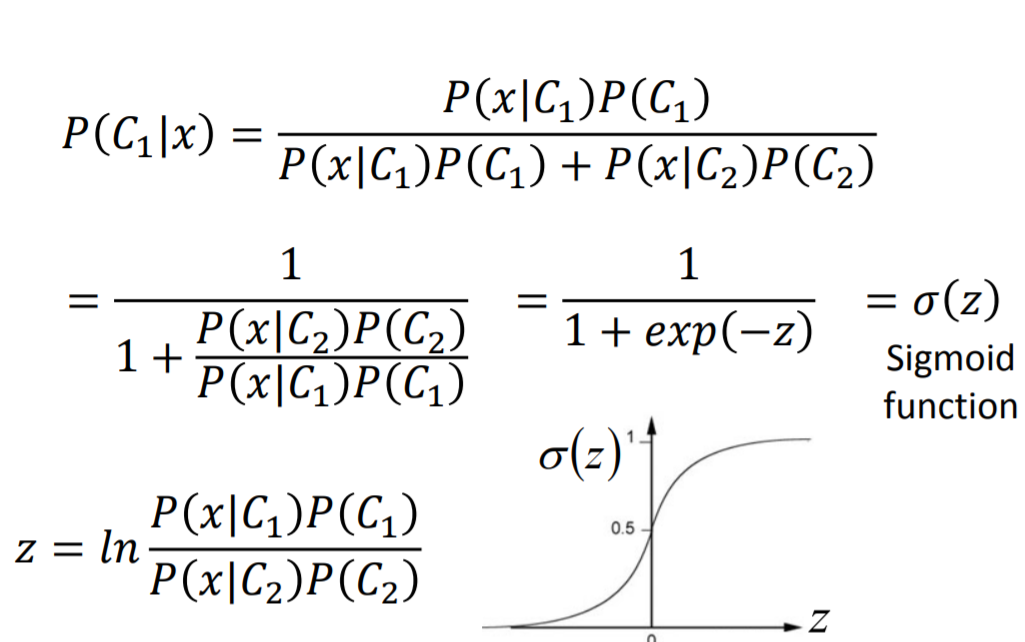
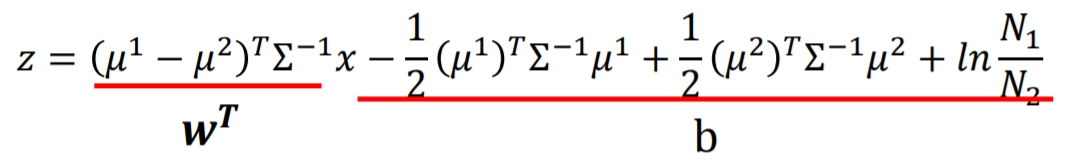
* Parallel issue: how numbers of parameters can be decided at the same time.

1. **What are the five key problems whenever we are talking about a learning process (Existence, Uniqueness, Convexity, Complexity,** **Generalizability)? Why are they so important?**

* Existence: We could have some kind of approximate or partial model, but this does not mean there exists an ultimate truth for our problem.
* Uniqueness: We can have alternative ways, depending on model selection, data. We can’t touch the physical truth if there isn’t a unique solution. Without uniqueness, our model can’t be generalizable.
* Convexity: Convexity makes the process of search much efficient. Otherwise, the solution may not be stable and be trapped into local optimal.
* Complexity: Not only computation complexity but also related to robustness, underfitting, overfitting, sparsity.
* Generalizability: expressive (smoothness) and inclusive (capacity)详见第三题

1. **Give a probabilistic interpretation for logistic regression, how is it related to the MLE-based generative methods from a Bayesian perspective?**

* ****在logistic regression中我们直接假设y的后验概率符合：
* 若采用generative model，我们通过先验概率和贝叶斯公式来计算后验概率：

对其先验概率和做几种不同的假设最终都可以得到的结果，如假设和都符合高斯分布，则：

使用其他分布如伯努利分布等也可以得到同样的结果。

* logistic regression直接计算得到w和b，而generative model先计算出先验分布的参数再根据对应关系得到w和b。因此LR更为robust，而当先验假设与数据更为吻合时generative model表现得更好。

1. **What are the mathematical bases for the logics regression being the universal posterior for the data distributed in any kinds of exponential family members?**

* 直观感受和21是一样的，这YB要是能讲清我把名字倒过来写

1. **What are the key advantages of linear models? But why linear model tends not expressive?**

* Low complexity, convexity(no local solution)
* Linear model is too simple to describe the process in detail, may cause the problem of underfitting.

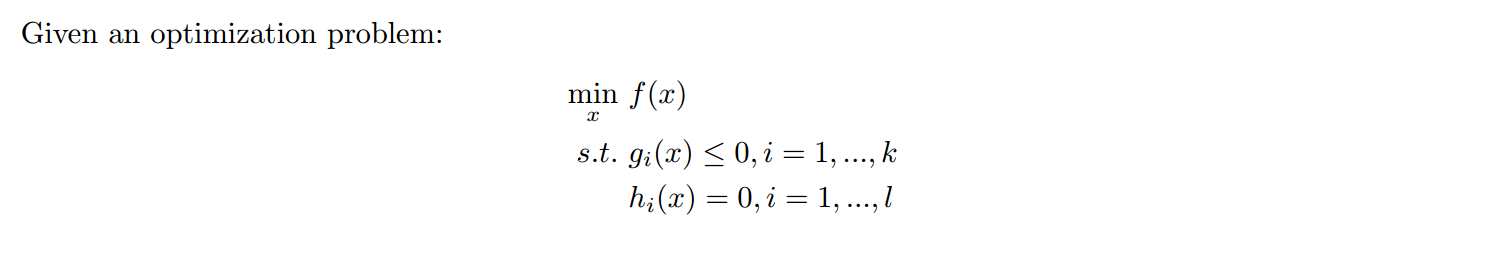
1. **What are the key problems with the complex Neural Network with complex integrations of non-linear model?**

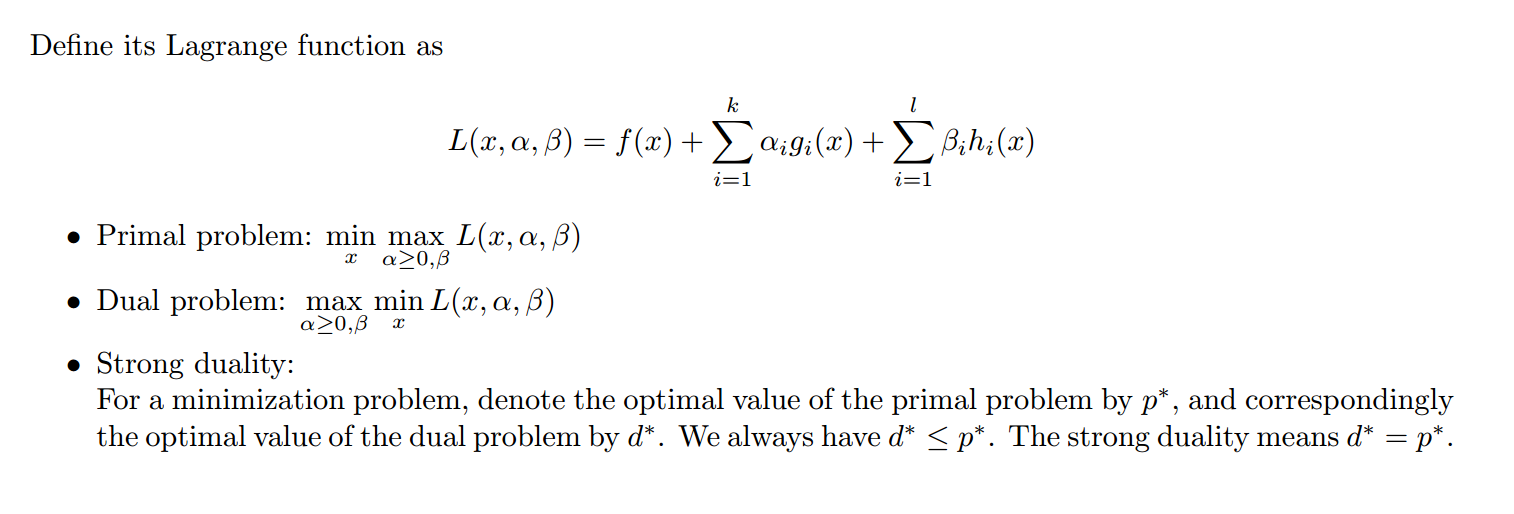
* Structure: topology, width, depth, how each layer is connected with others
* Activation function
* Loss function

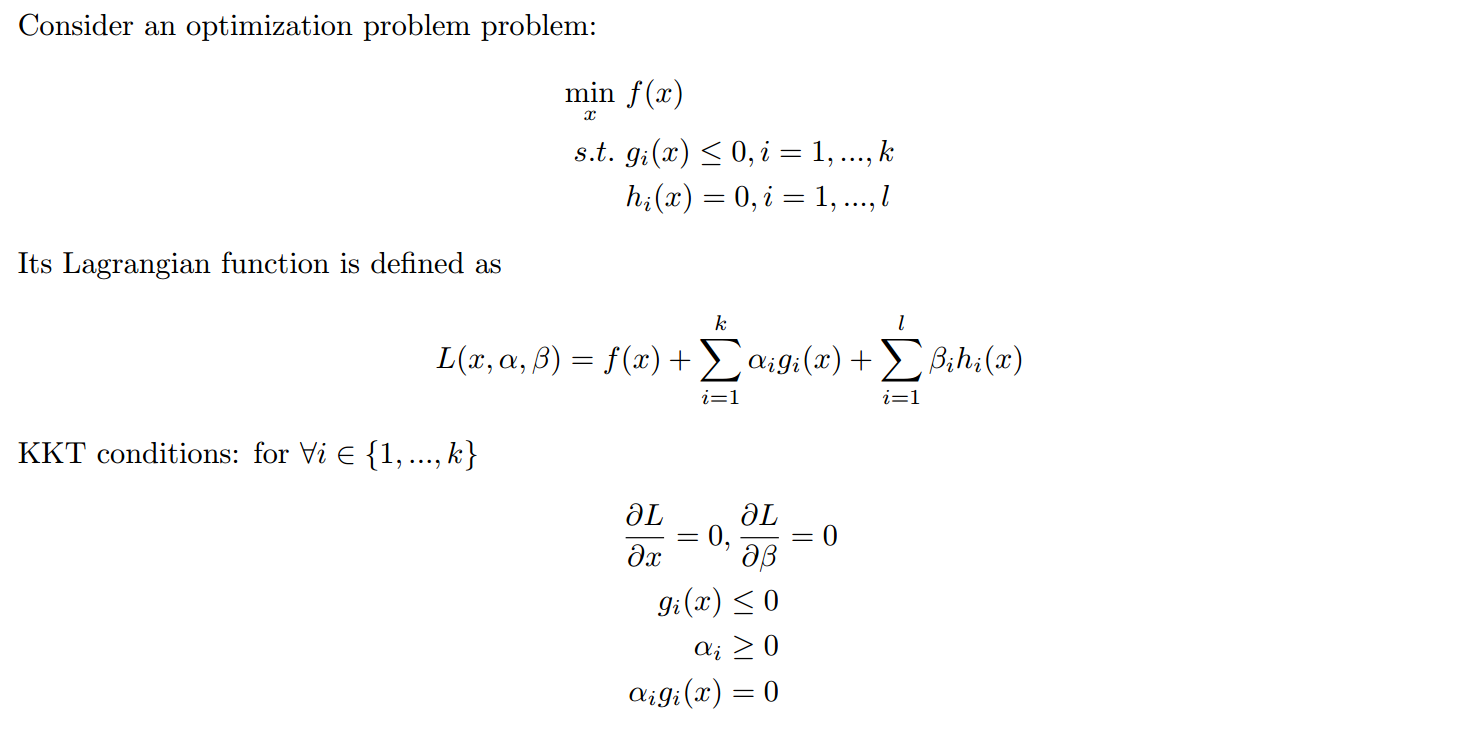
1. **What are three alternatives to approach a constrained maximization problem?**

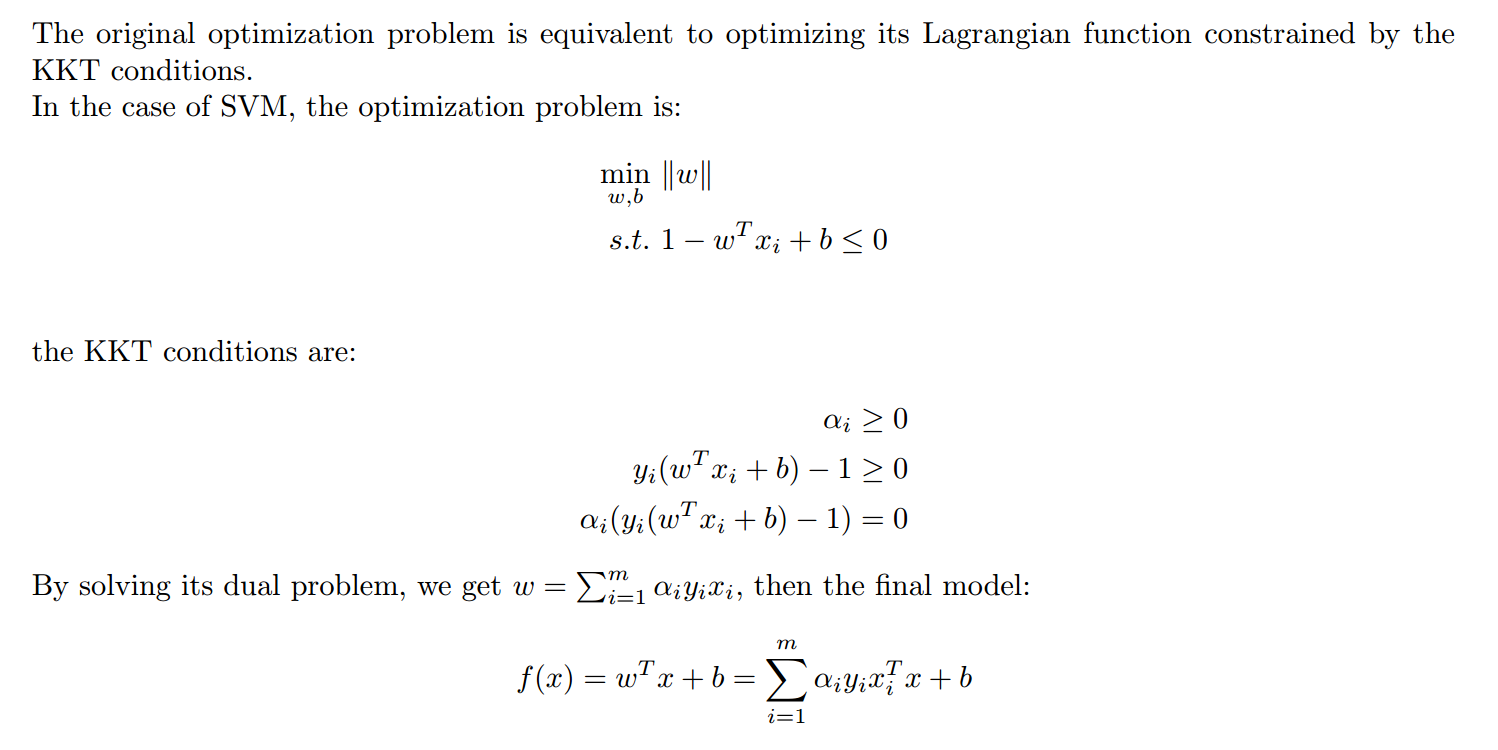
* Find its equivalent problems (Modify the objective function)
* Solving its dual problem (Lagrange Multiplier)
* Using kernel tricks(这个应该不对吧)

1. **What is the dual problem? What is strong duality?**

****

****

1. **What are the KKT conditions? What is the key implication of them? Including the origin of SV?**



若则该数据不影响结果，若则该data在边界上，是support vector。

1. **What is the idea of soft margin SVM, how it is a nice example of regularization?**

* Introducing penalty on the misclassified cases, which makes the model more be robust and stable. Using the constant C to control the width of the margin, in other words, how deeper the margin can go into the data (inverse relation).
* Due to the use of hinge loss:, many data are removed, which utilize the sparsity.

1. **What is the idea of kernel? Why not much additional computational complexity?**

* Kernel: project the space into high dimensional space, sometimes infinite dimensional space non-linearly.
* Kernel trick: do inner product first and get a scalar(number), then plug the scalar into a function. (We still need to choose the kernel).

1. **Why we often want to project a distance “measure” to a different space?**

* Our problem itself is not formed in the true Euclidian space. We want to project it into some space at least some kind of normalized and orthogonal, then we can do some math computation.

1. **What a Turin machine can do? What some of the key computable problems a Turin machine can do?**

* Forward problems (branches, logical……)
* State machine
* Recurrent

1. **What a Turin machine cannot do? What some of the key computable problems a Turin machine cannot do?**

* Interaction/ concurrence/ Parallelism
* Stochastic

1. **Give your own perspectives on the Hilbert No.10 problem in the context of computational limit.**

* 1970年, 苏联数学家马蒂塞维奇最终证明：在一般情况下，答案是否定的。因此，存在着大量数学问题人们永远无法知道其答案是否存在，自然也就无法去找到解决它的办法。人们面对这样的问题只能束手无策。因此一个可以计算的机器可能从诞生之初就有其无法逾越的极限。

1. **Give your own perspectives on the Hilbert No.13 problem in the context of mathematical limit.**

* Any complex mathematical problem can be described as some basis functions and their combination. That’s the fundamental of deep learning and neural network.

1. **Discuss human intelligence vs. Turin machine as to whether they are mutually complementary, exclusive, or overlapping, or contained into each other one way or another.**

**Human machine**

1.Parallel/

Concurrence

2. Stochastic

branches

1. **Explain Bayesian from a recursive point of view, to explain the evolution of human intelligence.**

* Prior is not stationary but something can recurve—the posterior can become the new prior then further constrain the search.
* If a problem is too complex, we can divide it into smaller process to recurve.
* Take best advantage of the prior to constrain the probable solution.

1. **What is computational evolutional basis of instinct, attention, inspiration, and imagination?**

* Prior, Bayesian, pre-existing knowledge

1. **Explain the core idea of machine learning to decompose a complex problem into an integration of individual binary problems, its mathematical and computational frame works.**

* 阿诺德表示奠定了理论基础。Kolmogorov–Arnold Representation Theorem (1956) – Every multivariate continuous function can be represented as a superposition of continuous functions of two variables，在机器学习中这个二元边界问题可以用sigmoid function这种概率模型来表示。
* A complex decision problem can be decomposed into small decision problems as 0-1. What we really need to do is how this parts are put together(decision tree……).

1. **What is the limitation of Euclidian (Newtonian) basis, from the space, dimension, measure point of view?**

* **讲了个啥？？？应该和第六题差不多吧**
* Even though its perfect Euclidian, in high dimensional space(even 3D), because of differential, initial value sensitive issue is very usual.

**Probabilistic graphical model**

1. **Compare the graphical representation with feature vector-based and kernel-based representations.**

* feature vector-based:非欧式空间的，维度之间彼此不独立、不归一，需要进行降维、归一化等操作。
* kernel-based representations：将低维空间非线性的投影到高维，希望投影后可以线性可分。
* graphical representation：承认变量之间的依赖关系，通过图结构来描述变量之间的关系。

1. **Explain why sometime a marginal distribution has to be computed in a graphical model.**
2. **Why class labels might be the key factor to determine if presumptively different data distributions can be indeed discriminated?**

* 从贝叶斯的角度来看，标签决定了先验分布的正确性。后续所做的一切都是计算不同标签对应的条件概率，一旦先验错误则模型的准确性从根本上无从谈起。

1. **Why knowledge-based ontology (representation) be a possible solution for many prior-based inference problems?**

* 根据人类现有的一些现有的相关领域知识对图的结构进行约束，使得数据尽可能的收敛到预设的结构中，这样学习到的模型就具有较强的可解释性。

1. **Why a graphical model with latent variables can be a much harder problem?**

**假设目标函数为：**

当同时还受到隐变量控制时（这里假设z为离散变量且只有K种不同的取值），因此目标函数为：

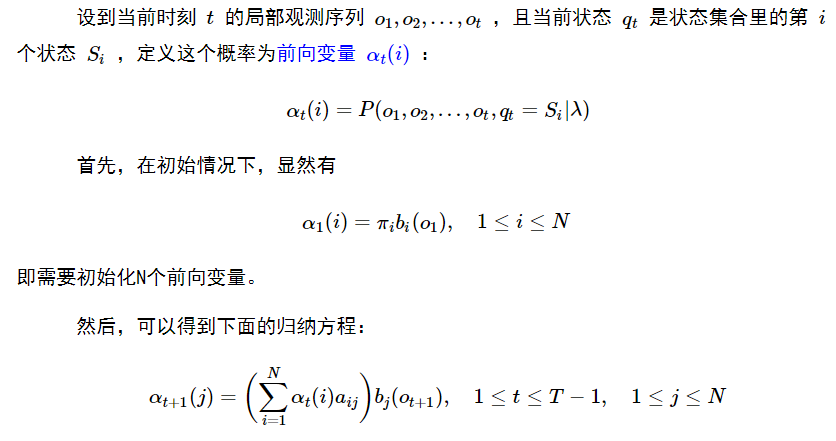
因此结果是一个非常复杂的项，从数值的角度很有可能得到局部解。

在图中对应着最大子图问题（我猜是指极大团），这是一个NP complete问题，除非穷举否则得不到全局解。

1. **What is the key assumption for graphical model? Using HMM as an example, how much computational complexity has been reduced because of this assumption?**

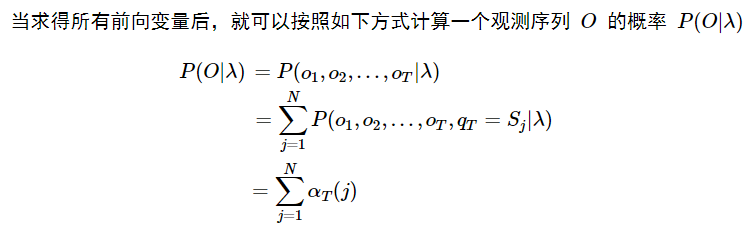
* key assumption 1:马尔可夫条件(下一时刻的状态仅与当前时刻的状态有关，不依赖于过去的状态)即局部感知。（YB认为1是最重要的）
* key assumption 2：观测独立假设（当前时刻观测值仅与当前时刻的隐状态有关）。
* 假设隐含状态共有N种可能的取值，链的长度为T，则计算复杂度从穷举的降低到了前项/后项算法的。

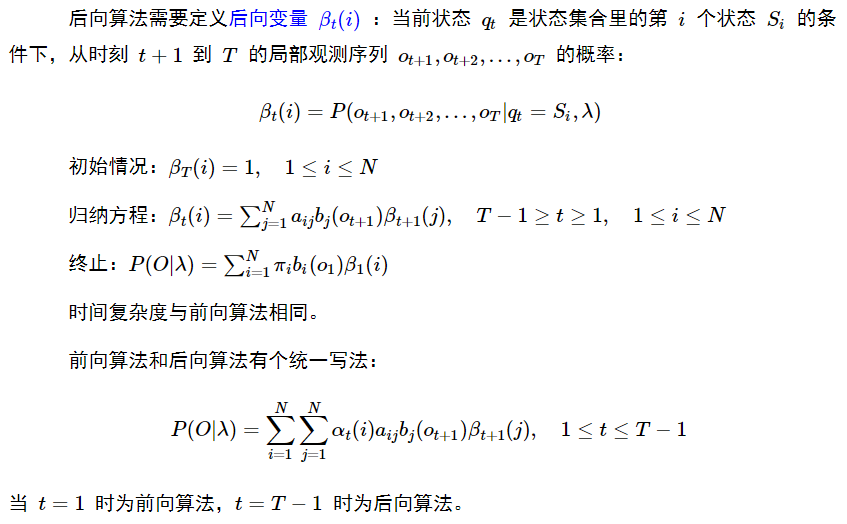
1. **How to interpret the HMM-based inference problem from a Bayesian perspective, using the forward/backward algorithm?**
2. **5**
3. **How to interpret the HMM-based inference problem from a Bayesian perspective, using the forward/backward algorithm?**

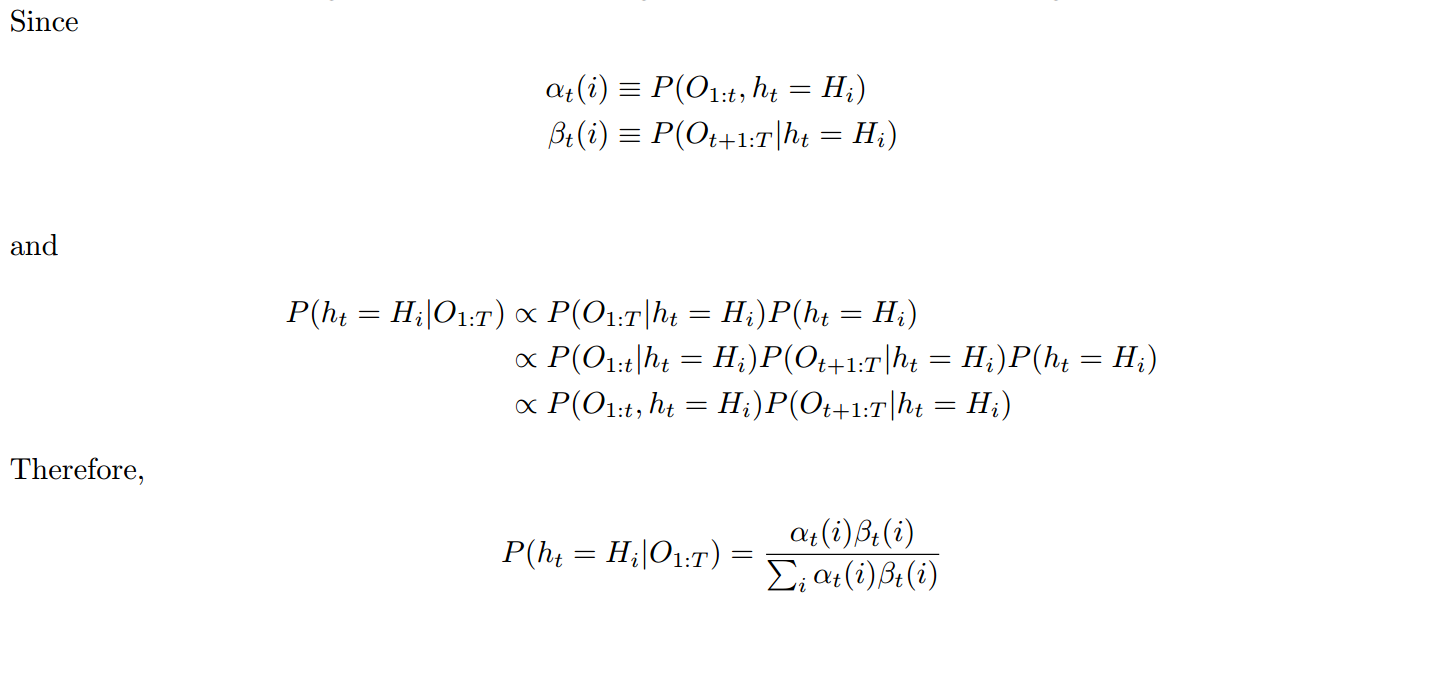
* Forward:

backward

forward



* Backward：
* 前项算法是基于先验的算法，将上游的先验逐层的递推下去，得到的后验概率不断变为下一步的先验，与人类学习的过程相似；后向算法是基于数据的回归算法，要看完所有的数据之后，从数据中寻找最好的模型，给出一个最好的解释。前向后向算法则是将人和机器的角度结合的算法。

1. **Show how to estimate a given hidden state for a given series of observations using the alpha and beta factors.**
2. **How a faster inference process would be constructed, given a converging network?**

* 没听懂，估计也不考

1. **How can an important node (where inference can be significantly and sensitively affected) be detected using an alpha and a beta process?**

画每一个k的曲线，看哪一个k先收敛，其中一个会占据支配地位。

1. **Why often an alpha process (forward) is more important than beta (backward)?**

* 若在forward的过程中已经收敛，K种隐藏状态中的一个状态已经占据了主要地位，则后续的数据不会对结果产生显著的影响**。**

1. **What are the key differences between an alpha and a beta process from human and machine intelligence point of views?**

* 前项算法是基于先验的算法，将上游的先验逐层的递推下去，得到的后验概率不断变为下一步的先验，与人类学习的过程相似；后向算法是基于数据的回归算法，要看完所有的数据之后，从数据中寻找最好的模型，给出一个最好的解释。前向后向算法则是将人和机器的角度结合的算法。

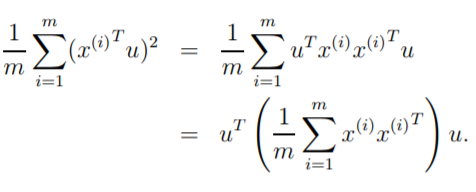
1. **How data would contribute to the resolution of an inference process from a structural point of view?**

* 如果有一个比较好的先验/比较准确的结构，则数据都是加强先验的条件，结构大大约束了可能搜索的范围。
* 如果没有好的先验，再多的数据也不一定能得到好的结果，因为数据的增加和模型可搜索范围的unbounded完全不在一个数量级。

**Representation and Deep Learning**

1. **PCA is an example of dimensional reduction method; give a full derivation of PCA with respect to its eigenvectors; explain SVD and how it is used to solve PCA.**

* PCA推导（参考吴恩达cs229笔记）

假设已经被标准化使得均值为0，方差为1，若要将数据投影到单位向量u上使得方差最大，则：

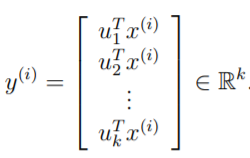
问题进一步被描述为：

用拉格朗日乘子法解决：

对求导数得：

令偏导数为0可得即u为的特征向量，且。因此u为特征值最大的特征向量，即主特征向量。

若要将其投影到k维空间，则

**

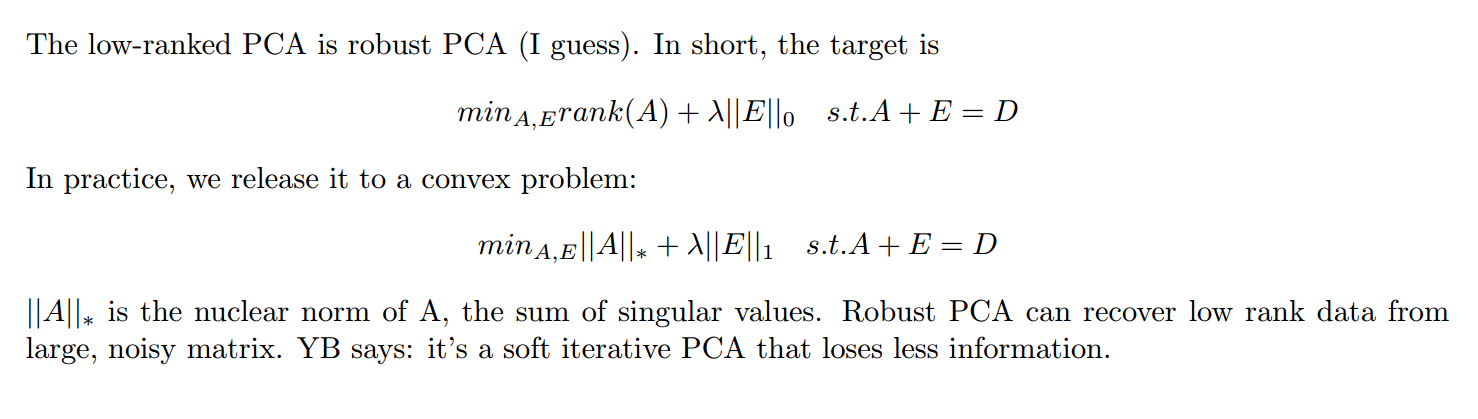
为使得最大，则需要每一维最大，因此选取的是对应特征值最大的k个特征向量，由于特征向量的性质，所选出的k个向量是正交的。

* 奇异值分解（SVD）

任意的矩阵M可被分解为：

其中的前k列即为的top k特征向量，即PCA中所需要的投影向量

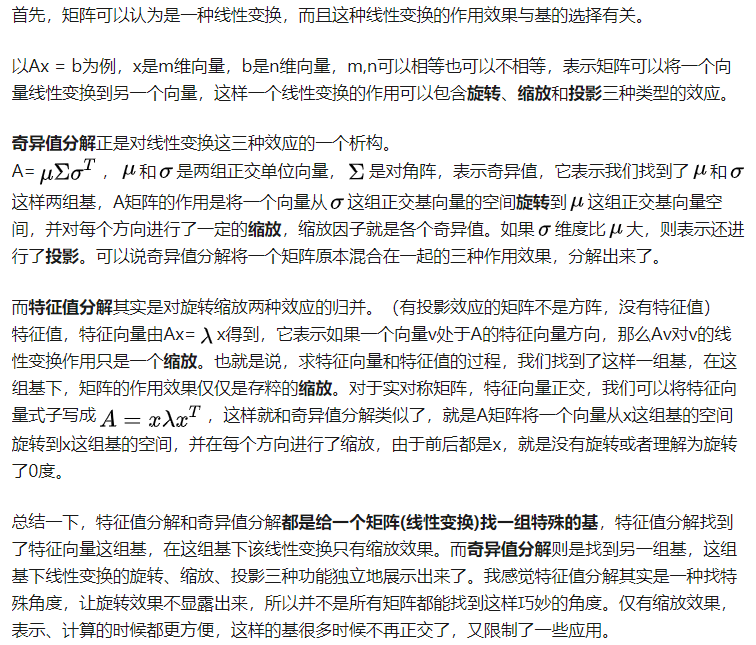
1. **Compare regular PCA with the low-ranked PCA, what would be advantage using the low-ranked PCA and how it is formulated?**

****

**PCA Robust PCA**

能更好的应对稀疏的噪声，不会被个别极端数据影响

1. **What is the difference between a singular value and its Eigen value? Explain the resulting singular values of an SVD for how the features were originally distributed;**



奇异值表示了数据对原始数据的贡献程度，如下图所示：

不冗余 冗余

1. **What is the key motivation (and contribution) behind deep learning, in terms of data representation?**

* 学习问题本质的结构和特征，而不是局限于对特征进行简单的变换
* 不同于传统方法将从“背景”中分离“信号”的思想，deep learning学习信号和环境的关系，借助context更好的理解object，强调pattern的学习。
* 分解：复杂的问题是由一些简单的，相似的块组合构成的，块可被重复利用。

1. **Compare the advantage and disadvantage of using either sigmoid or ReLu as an activation function?**

* Sigmoid:

优点：输出是归一化到0-1的值，可以被看做某种概率，相当于复杂的数学问题转化为0-1的逻辑决策；

缺点：

1.引入参数较多？？？（可是sigmoid函数没有参数啊）

2.会发生梯度消失，只有0附近的值有较为大的梯度，若网络较深则影响无法有效的向前传递。

3.sigmoid的组合会出现大量的局部解。

4.微分存在初值敏感问题。

* ReLU：

优点：解决梯度爆炸；模型较凸，减少局部解。

缺点：输出不归一，同层不同结点的输出无法相互比较。

1. **Discuss matrix decomposition as a strategy to solve a complex high dimensional problem into a hierarchy of lower dimensional combinations?**

* Deep Learning的重要思想：把复杂问题想象成高维矩阵的分解，不断对高维矩阵进行分解和重构，学习重构的模式。

1. **Discuss convexity of composite functions, what mathematical strategies might be used to at least reduce local solutions?**

* 马尔可夫假设：一层只与上一层有关，减少依赖关系。
* Hinton胶囊理论(分区)：与全连接不同，每个胶囊由一些小的网络组成，胶囊独立处理任务，胶囊采用路由算法确定连接关系。(胶囊对其输入执行一些相当复杂的内部计算，然后将这些计算的结果封装成一个包含信息丰富的输出的小向量。每个胶囊学习辨识一个有限的观察条件和变形范围内隐式定义的视觉实体，并输出实体在有限范围内存在的概率及一组“实例参数”。)
* 改变网络的拓扑结构

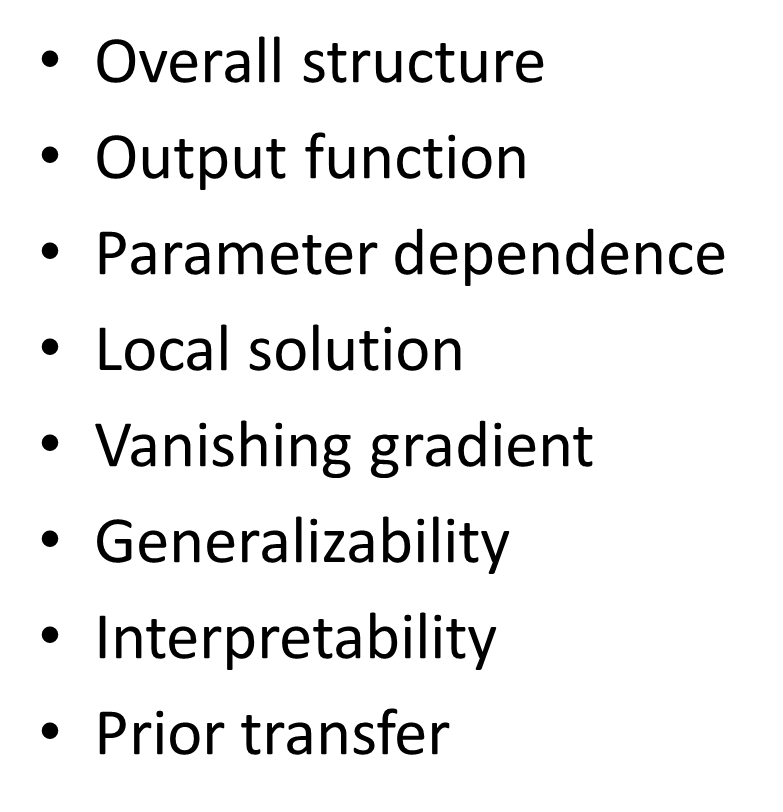
1. **Why normally we use L2 for the input layers and L1 for the actual modeling? Explain why still sigmoid activations are still used for the output layers?**

* L2 for input:数据不在欧式空间，输入数据的各维度表达的含义不同，因大小的不同会产生各种bias，使用L2使得输入更平滑，使得各维度差别不至于过大。
* L1 for the actual modeling：使参数矩阵稀疏，降低模型的复杂度，增加可泛化能力。
* Sigmoid:归一化到0-1输出，表示某种概率。

1. **What would be the true features of an object modeling problem? Give two examples to highlight the importance of selecting appropriate dimensions for feature representations;**

* 人脸朝向的识别：相比于有上万个像素的图像，只需要一个3维feature就可以解决。
* 利用DL模拟三维真值表，只需要3个sigmoid函数（不懂在说啥）

1. **Why does the feature decomposition in deep learning then a topological recombination could make a better sampling? What would be the potential problems making deep learning not a viable approach?**

* 模型中学习到的参数并不都是独特的，通过学习会得到一些**可重复**利用的**模块**，因此对数据的要求并没有表面上那么高。
* 

1. **Explain the importance of appropriate feature selection being compatible with model selection in the context of model complexity.**

* 数据比模型复杂——underfitting，模型比数据复杂——overfitting（我自己理解的）；
* 其实没讲，就光说重要了

1. **What would be the ultimate and best representation for a high dimensional and complex problem? How this might be possibly achieved?**

* 模型应该是可解释的，而不是黑箱。比如一开始先尝试decision tree，如果decision tree能够产生比较好的结果，则不必采用深度神经网络。

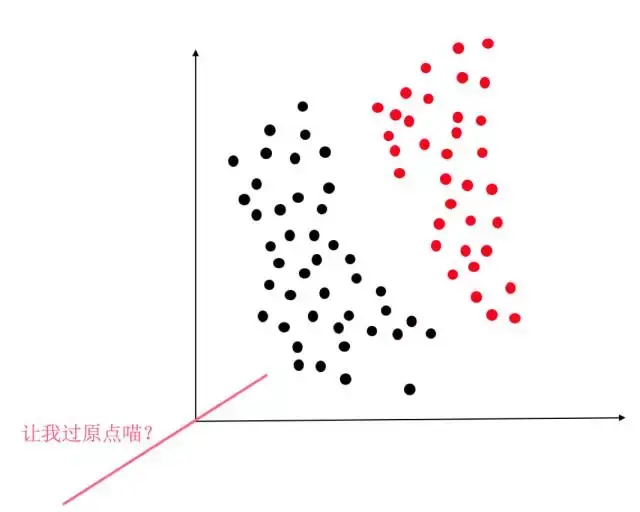
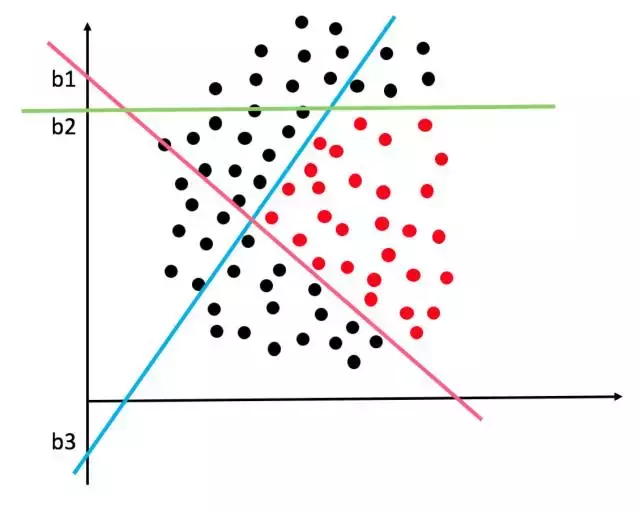
1. **How RNN can be expanded our learning to fully taking advantage of Turing machine? What RNN can whereas CNN cannot do?**

* RNN可以进行循环，解决反馈、长短程相互作用的问题，被证明是图灵完备的。相比于CNN，RNN可以学到运算的逻辑过程，而不仅仅是输入输出的结果。
* RNN可以做的：时序相关的问题，输入之间的相互作用……

1. **What is the central additional difficulty of RNN compared to CNN?**

* RNN的每一次训练可以被看做一个小图灵机，它的收敛与逻辑无法与其他的训练结果归一，实际上求解的是每一个局部的loss。如何将这些局部的loss组合起来形成真正的global loss实际上是非常复杂的。

1. **In the activation function, there is a constant term “b” to learn, why it is important?**

* 我理解这题应该问的是神经网络中的偏置项，真不知道和激活函数有啥关系……
* 不加偏置项所得到的0-1超平面就必须过原点，使得绝大部分问题不可分。
* 机械的进行各维度之间的smooth

1. **LSTM integrate short and long term processes, what is the central issue to address to achieve at least some success?**

* Long short-term memory (LSTM)这里想要纠正一下，是长期的短时记忆的意思，不是什么long-term short-term memory.
* LSTM在原有RNN的基础上添加了3个控制门：输入门、输出门、遗忘门，使得单元的状态可以选择性遗忘和记忆，它的更新是缓慢的，因此可以记住更长时间的内容。
* LSTM不仅仅进行了标量的计算，把一些问题想象成矢量计算，多出了空间、大小等物理意义，因此其大小就有可比性。把长程的关系想象成矢量运算，短程问题想象成标量运算，长短是两种不同的测度。（完全不懂，纯转述）