计算物理

第一章：绪论

计算物理

定义：以计算机及计算机技术为工具和手段，运用计算数学的方法，解决复杂的物理问题的一门应用科学。

计算物理vs理论物理：

理论物理：将问题简化得到问题的显式解析解。（考虑主要因素、静态情况、线性化、常系数处理、规则边界）可能会抛弃掉本质特征。

计算物理：可以多考虑一些因素，模拟动态过程，保持非线性，保留变系数特点，可以考虑复杂的边界条件；省钱省时；更大自由度和灵敏度、安全，不存在物理实验的测量误差和系统误差（没有测试探头的干扰...）；可以自由选择参数；可以进行物理实验难以进行的计算机实验。

优势：仿真物理过程，发现物理现象，提供新消息；对物理问题数值分析，提供物理规律的数据；计算机为手段的实验物理，又是计算机武装起来的理论物理。

局限：计算所用的基本方程在具体问题中采用不同的简化与近似；稳定性，解存在、唯一性，收敛性，误差分析没有严格的论证，难以满足各种复杂实际问题的需要。

作用：解决传统解析研究无法解决的问题；建立连接理论与实验的桥梁；替代实验或减少实验成本的必要手段；揭示新的物理效应和规律。

第二章：蒙卡方法

蒙特卡罗方法

定义：以概率统计理论为指导的一类数值计算方法，是针对随机过程而提出的计算机模拟方法。

形式：直接蒙特卡洛模拟；蒙特卡洛积分；Metropolis模拟

基本思想：建立概率模型，使其参数为问题的解；对模型观察抽样实验来计算所求的统计特征，给出解的近似值；解的精度用估计值的标准误差描述。

解题步骤：构造概率模型；实现从已知概率抽样；建立估计量

核心问题：以概率统计为其主要的**理论基础**（大数定理、中心极限定理）；以随机抽样为主要手段。

大数定理：当样本量足够大时，随机变量的均值会收敛于其期望值。

中心极限定理：当样本量足够大时，随机变量均值会服从正态分布。

特点（优缺点）：

优点：MC方法的误差与问题的维数无关，特别适用于求解高维问题，计算时间于维数乘正比；MC方法受问题条件限制的影响小，可以求解复杂问题；可以构建不同的随机模型解决问题；

弱点：收敛速度慢，误差下降速度慢

随机数

特点：均匀性（误差）、独立性（相关）

分类：真随机数；准随机数；伪随机数

产生方法：取中法、移位法、同余法

平方取中法：$x_{n+1}=[x_n^2/10^s](mod\quad 10^{2s})\qquad \xi_n=x_n/10^{2s}$

线性同余法：$x_{n+1}=(ax_n+c)(mod\quad M)\qquad \xi_n=x_n/M$ 乘同余法：c=0

抽样方法：

离散：直接抽样法

连续：**直接抽样法**（反函数法）；**变换抽样法**；**舍选抽样法**；复合抽样法...

舍选法步骤：找出区间[a,b]内函数的最大值$\frac{1}{\lambda}$；选用[0,1]上均匀分布的随机数ξ_1，构造出[a,b]上的均匀分布随机数$\delta=a+(b-a)\xi_1$；再选取在[0,1]上均匀分布的随机数ξ_2，判断$\xi_2\le\lambda f(\delta)$是否满足，满足则选取$\eta=\delta$作为抽样值。

减少方差的技巧：增加随机点数n；减少计算结果的方差；尽量使用理论分析得到的期望值来代替模拟估计值

分层抽样法：积分区间划分为不相交的子区间，对区间进行合理划分
$I=\int_0^1 f(x)dx=\int_0^a f(x)dx+\int_a^1 f(x)dx)$

重要抽样法：引入偏倚函数g(x)使得被积函数在区间内平缓，$I=\int_0^1 f(x)dx=\int_0^1\frac{f(x)}{g(x)}g(x)dx$由$\frac{f(x)}{g(x)}$产生子样，附带权重g(x))

控制变量法、对偶变量法

随机游走:运用[0,1]区间的均匀分布随机数序列进行的运算。（查点法、**MC方法**）

Metropolis

Metropolis方法（重要抽样法）：按照一定的规则（选择过渡几率$w(x\to x')=min[1,\frac{f(x')}{f(x)}]$）来进行随机游走，经过大量游走，平衡后，产生点的分布满足所要求的分布(对不可归一化分布函数同样适用)。

步骤：1.取试探位置，$x_{try}=x_n+\eta_n,\quad \eta=random(-\delta,\delta),\delta$为试探步长。2.计算$r=\frac{f(x_{try})}{f(x_n)}$ 3.若$r\ge 1$则接受试探；若$r<1$生成随机数$\xi=random(0,1)$，若$\xi<r$那么接受这步游走；反之拒绝这步游走，留在原位置。

选择定则：步长选择应让1/3到1/2的试探步被接受；初始位置应在区域的概率分布密度最大的位置。

第三章：蒙卡运用

积分计算：

平均值法：$\bar{E}=(b-a)\frac{1}{N}\sum_{i=1}^N f(\xi_i)$

投点法:引入随机变量：$\eta(\xi_1,\xi_2)=\begin{cases}1,\xi_1\le f(\xi_2)\\0,\xi_1>f(\xi_2)\end{cases}$则
$I=E\{\eta(\xi_1,\xi_2)\}=\frac{1}{N}\sum_{i=1}^N\eta(\xi_{2i-1},\xi_{2i})=\frac{m}{N},\quad V\{\eta\}=E\{\eta^2\}-[E\{\eta\}]^2=I-I^2$

投点法比平均值法方差大：
$V\{\eta\}-V\{\eta_1\}=I-I^2-\int_0^1[f(x)-I]^2dx=\int_0^1 f(x)[1-f(x)]dx\ge 0$

尽量使用理论分析的期望值代替模拟估计值来减少方差，加快收敛。

统计力学的MC方法

Ising模型：利用metropolis方法：随机选取格点，反转该点自旋；计算能量差，若体系能量减小，改变有效；若体系能量增大，则再生成随机数r_i，若$r_i < e^{-\beta\Delta E}$则接受改变。

第四章：有限差分法

有限差分法

数值求解常微分方程或偏微分方程

具体操作：用差分代替微分（离散化），得到差分方程组，求解

五点（菱形）格式：
$\nabla\phi = \frac{\phi_{i+1,j}+\phi_{i,j+1}+\phi_{i-1,j}+\phi_{i,j-1}-4\phi_{i,j}}{h^2}$ $\phi_{i+1,j} + \phi_{i,j+1} + \phi_{i-1,j} + \phi_{i,j-1} + (h^2 f_{i,,j} - 4)\phi_{i,j} = h^2 q_{i,j}$

误差来源：截断误差（误差阶）；计算误差（稳定性）

边界条件：$\phi|G + g_1(s)\frac{\partial\phi}{\partial n}|_G = g_2(s)$

1.Dirichlet问题(g1=0) 2.Neumann问题(g2=0) 3.混合问题

边界离散化处理：直接转移法；线性插值法

求解差分方程方法：直接法（矩阵求逆）；随机游走；迭代法（广泛）

迭代法：直接迭代法；Guass-Seidal迭代法；超松弛迭代法

第五章：有限元素法

基本思想：基于变分原理，求解泛函取极小值变分问题（计算格式复杂，适用性较差）。

一般步骤：推导出给定边界条件的偏微分方程的等价的泛函表示；用三角形元素分割求解区域，编号；计算系数矩阵K、P；利用边界条件构造有限元的方程组求解方程组。

有限元素法vs有限差分法

有限元素法：变分原理，作用量求极值；三角形划分；计算精度总体协调，计算矩阵为对称正定的大型稀疏矩阵；输入数据量太大。

有限差分法：用差分代替微分，将微分方程转换为差分方程组；矩形网络划分，要求边界形状规则；各节点精度不一致；应用范围广。

第六章：分子动力学方法

随机模拟方法；确定性模拟方法（分子动力学方法）。（是否考虑内禀动力学规律）

分子动力学方法

限制：有限观测时间；有限系统大小

Verlet

Verlet算法：$r_i^{(n+1)} = 2r_i^{(n)} - r_i^{(n-1)} + F_i^{(n)}h^2/m \quad v_i^{(n)} = (r_i^{(n+1)} - r_i^{(n-1)})/(2h)$

位置Verlet算法：
$r_i^{(n+\frac{1}{2})} = r_i^{(n)} + \frac{1}{2}hv_i^{(n)} \quad v_i^{(n+1)} = v_i^{(n)} + hF_i^{(n+\frac{1}{2})}/m \quad r_i^{(n+1)} = r_i^{(n+\frac{1}{2})} + \frac{1}{2}hv_i^{(n+1)}$

速度Verlet算法：
$r_i^{(n+1)} = r_i^{(n)} + hv_i^{(n)} + F_i^{(n)}h^2/2m \quad v_i^{(n+1)} = v_i^{(n)} + h(F_i^{(n+1)} + F_i^{(n)})/(2m)$

蛙跳算法：
$v_i^{(n+\frac{1}{2})} = v_i^{(n-\frac{1}{2})} + hF_i^{(n)}/m \quad r_i^{(n+1)} = r_i^{(n)} + \frac{1}{2}hv_i^{(n+\frac{1}{2})} \quad v_i^{(n)} = \frac{1}{2}[v_i^{(n+\frac{1}{2})} + v_i^{(n-\frac{1}{2})}]$

势能函数

成键相互作用能=键伸缩能+键弯曲能+二面角扭转能

非成键相互作用能=范德华作用能+静电作用能

L-J势：$U(r) = 4\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6]$，其中$\epsilon$表示能量单位（$-\epsilon$是体系能到达的最小能量），$\sigma$是距离单位，当$r = \sigma$是位势取为0。

分子收到L-J势的相互作用力为：$\vec{F}_i = 48(\frac{\epsilon}{\sigma^2})\sum_{j=1,j\neq i}^{N}(\vec{r}_i - \vec{r}_j)[(\frac{\sigma}{r_{ij}})^{14} - \frac{1}{2}(\frac{\sigma}{r_{ij}})^8]$

非键邻居列表：最短模拟时间内，一个原子的邻居没有太大改变，定期更新，期间不变。

邻居搜索：简单方法：仍计算两两粒子之间的距离$O(n^2)$

格点方法：将元胞分成M^3个格子，邻居搜索在三维空间的周围3x3x3格子中搜索

截断距离：元胞尺寸设立$L/2 > r_c$

周期性边界条件：$A(\vec{r}) = A(\vec{r} + n\vec{L})$　　**最小像力约定**：$r_{ij} = min|\vec{r}_i - \vec{r}_j + n\vec{L}|$

双重截断：lower cutoff每个时间步长正常计算，lower cutoff~upper cutoff更新邻居列表时重新计算。

截断问题解决：shifted potential & switch function（处理力不连续、能量不连续问题）

采样定理：采样频率大于信号中的最高频率2倍（一般采用5~10倍）

时间步长选择：系统最短运动周期的1/10

约束动力学：分子化学键的振动频率高，可以固定分子键长，增大时间步长

shake算法：在势能项中添加拉格朗日因子和约束条件的乘积项。

shake failure原因：不合理初始条件、时间步长太大、平面环状基团较难约束。

微正则系综与正则系综的模拟流程

微正则系综：体系具有相同的粒子数、体积、能量

正则系综：体系具有相同的体积、粒子数、温度

MC & MD

MC：不需要计算能量梯度；适合离散状态的空间采样；设计如何move

MD：适合于运动方程可推导，可积分的情况。

MD局限：模型准确性；计算量大；采样效率

第七章：高性能计算和并行算法

任务：并行程序能处理的并发性最小的单元，一个任务只能由一个处理器执行，并发性只能在任务之间开发

进程（线程）：完成任务的实体（处理器）

进程间通信：通信（读写系统、网络）、同步（等待）、聚集（综合结果）

并行的效率：计算的时间是分钟级的，而数据传输时间是秒级的，那么这部分是可以并行的。

MPI并行编程模式：单/多程多数据流模式（**SPMD**/MPMD）

单程多数据流模式：一个程序同时启动多份，形成多个独立的进程，在不同的处理机上运行。

多程多数据流模式：使用多个程序处理多个数据集，合作求解一个问题。

MPI函数：MPI_Init:指示系统完成初始化。MPI_Finalize:让系统释放分配给MPI的资源，MPI程序最后一条可执行语句。

MD并行算法：粒子分解、域分解

负载平衡

CPU：核心少，控制能力、计算能力不弱，作为主机；多个线程，可以调动多个GPU并行。

GPU：计算核心多

第八、九章：计算机代数&Mathematica

数值计算：基于**有限精度数字**进行的计算

符号计算：基于**数学符号**进行数学公式计算

第十章：机器学习

机器学习

定义：利用已有数据进行学习，获取数据中的特征，构造或完善具有预言能力的模型，提高未来行动的效率、效果，以及准确性。 深度学习是实现机器学习的一种

分类：

监督学习：回归（连续值）/分类（离散值）算法

无监督学习：归类/聚合算法，降维

强化学习：游戏算法

方法：

线性回归、逻辑回归、集成学习、决策树与随机森林、神经网络

线性回归

损失函数：度量单样本预测的错误程度。

代价函数：度量全部样本集的平均误差。（损失函数求和，取平均）

目标函数：代价函数+正则化函数（最终优化的函数）

线性回归方法：

最小二乘法：一次计算得出结果；运算代价大；仅适用于线性模型。

梯度下降法：需要选择学习率，多次迭代；适用于各种模型。（批量（小数据集）、随机（大数据集）、小批量梯度下降法）

数据归范化：提升模型精度；加快模型收敛（归一化、标准化：是否改变特征数据分布）

训练误差：模型在训练集上的误差（经验误差） ~**泛化误差：**模型在新的数据集（测试集）的误差

欠拟合：添加新特征；增加模型复杂度；减少正则化系数

过拟合：更多训练数据；降维；正则化；集成学习

正则化：L1正则化（Lasso回归）：适用于稀疏模型　L2正则化（Ridge回归）：使权重平滑，防止过拟合

回归评价指标：MSE(均方误差)、RMSE、MAE、R方（拟合优度：$R^2 = \frac{SSR}{SST}$）

逻辑回归

线性回归+Sigmoid映射（处理分类问题）

存在一个决策边界 $w^T x$，可以将数据完成划分：$w^T x_i > 0$，$y_i = 1; else \quad y_i = 0$

Sigmoid函数作用：$g(z) = \frac{1}{1+e^{-z}}$ 使模型更加关注分类边界（鲁棒性）

逻辑回归：似然函数：$p(y|x;w) = [h(x)]^y[1-h(x)]^{1-y}$ 损失函数（交叉熵）：$L(\hat{y}, y) = -yIn\hat{y} - (1-y)In(1-\hat{y})$ 代价函数：$J(w) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$

决策树

原理：多级分类器，可以把一个复杂的多类别分类问题转化成若干个简单的分类问题。

“二叉树”最为常见；自上而下；贪心算法；

类型：ID3、C4.5、CART

信息增益：信息增益=信息熵-条件熵：　$g(D, A) = H(D) - H(D|A)$

信息熵$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|}$ 条件熵：$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i)$

信息增益率：$g_R(D, A) = \frac{g(D,A)}{H_A(D)}$

Gini-index：$Gini(D, A) = \frac{|D1|}{|D|} Gini(D1) + \frac{|D2|}{|D|} Gini(D2) \quad Gini(p) = \sum_{k=1}^{K} p_k(1-p_k)$

剪枝：去掉一些分支降低过拟合的风险；减少训练时间（存在欠拟合的风险）

预剪枝：估计节点能否给决策树带来泛化能力提升

后剪枝：剪枝后错误率下降，则替换。比预剪枝比，保留了更多的分支。

集成学习

定义：将多个较弱的统计模型集成起来预测

集成学习框架（方法）

袋装算法：将训练样本随机分配（重抽样），分别训练，再集成。　（要求数据集足够大）

随机森林：基于决策树每棵决策树都会给出判断，最后投票法给出分类结果

增强算法：基模型按次序一一训练 、转化。

自适应算法：后一个模型永远在前一个模型基础上完成。样本分类正确，降低构造下一个训练集的该样品的权重训练分类器。加大分类错误率低的弱分类器。

梯度增强树（GBDT）：迭代决策树算法。核心累加所有树的结果（回归树）（向前分步算法，梯度下降，缩减）

XGBoost：L2正则项；牛顿法；贪心方法；并行，多线程优化
LightGBM=XGBoost+GOSS+EFB+Histogram+Leaf-wise

神经网络

定义：基于人脑神经元网络对信息处理模式 （分类：DNN+CNN+RNN）

基本结构：输出层、隐藏层、输出层

超参数：隐藏层层数，神经元数目、梯度下降、正则化参数

训练过程：构造代价函数，求极值，给出w、b的最优值

感知机：二分类问题的线性分类的模型，目标函数：$f(x)=sign(w^T+b)$

反向传播算法（BP）：利用上一层神经元的阈值梯度，计算当前层的神经元阈值梯度和连接权值梯度。信号前传，误差后传，循环进行。

优点：自适应；非线性；推导过程严谨。

缺点：网络参数众多，收敛速度慢；隐层节点数没有明确准则；梯度下降容易陷入局部极小问题

无监督学习

聚类

定义：对研究对象分类，组内样本相似，组间样本差异。（硬聚类、软聚类）

距离尺度函数：几何距离、线性相关系数（时间点数据）、非线性相关系数（单调性）、互信息（时间上没有单调升降关系）

聚类算法：分层算法（聚合，分裂）、分割算法（K（簇数）均值）、密度聚类（eps、minPts；核心带你、边界点、噪声点）

聚类评价指标：均一性；完整性；轮廓系数；调整兰德系数

降维

定义：将训练的样本从高维空间转换的低维空间（线性变换） 不存在完全无损的降维

作用：减少冗余特征（具有线性关系），降低数据维度；数据可视化

主成分分析（PCA）：将一个大特征集转换成一个较小的特征集。（牺牲准确性）

得到主成分方法：计算数据矩阵的协方差矩阵（数据矩阵均值归一化，$\Sigma=\frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$）；得到特征值，特征向量；选择特征值最大的k个特征向量；转化数据空间。