概率论与随机过程 (2)，project_1© 清华大学电子工程系

第一次 project

一. 主题介绍

本次大作业的主题是——马氏链蒙特卡洛方法

马氏链蒙特卡洛（Markov Chain Monte Carlo, MCMC）方法，是马氏链理论的一个重要应用。从 1950 年萌芽，马氏链蒙特卡洛方法在实践中不断发展，逐渐成长为一个颇具分量的理论分支，广泛应用于各种学科领域（如信息科学、物理、化学、生物学、金融、材料等）的科学计算，展示出越来越强大的威力。

MCMC 的一个简短介绍见文献 [7]——林元烈编著的《应用随机过程》第 3.5 节 P^n 的极限性态与平稳分布。英文的简短介绍可见文献 [1]——《Pattern Recognition and Machine Learning》Chapter 11 Sampling Methods。更系统的介绍见文献 [4]。

二. 作业题目

(a) 二维高斯的相关系数的估计。

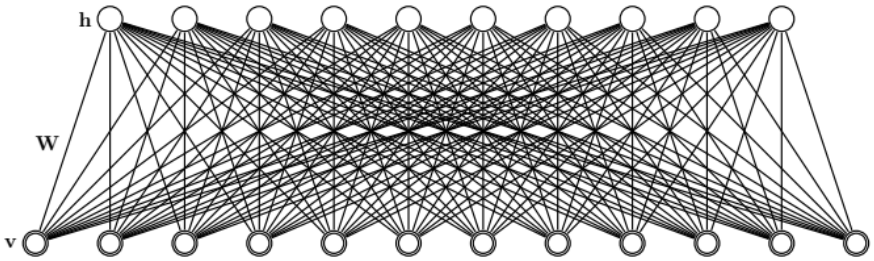
用 Metropolis-Hastings 算法，对下述二维高斯分布进行随机采样

$$\mathcal{N}\left\{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}\right\}$$

使用生成的随机样本来估计二维高斯的相关系数 ρ，并与真值比较。具体设计出你用的 Metropolis-Hastings 算法，结合实验结果，分析算法性能，特别是估计的准确性和计算效率。

(b) RBM(Restricted Boltzmann Machine) 模型的归一化常数的估计。

受限波尔兹曼机（RNM）是深度学习中最重要最基础的模型之一，它是一种常用的无向图模型，具体结构如下：



RBM 由一层观测变量和一层隐变量构成，变量均为 0, 1 取值，模型的的参数为 $\theta = \{W, b, a\}$。模型的能量为:

$$\begin{aligned} E(v, h; \theta) &= -v^{\mathrm{T}} W h - b^{\mathrm{T}} v - a^{\mathrm{T}} h \\ &= -\sum_i \sum_j W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j \end{aligned}$$

RBM 模型的概率分布即变量的联合分布为：

$$p(v, h; \theta) = \frac{1}{Z(\theta)} \exp\left(-E(v, h; \theta)\right)$$

其中

$$Z(\theta) = \sum_v \sum_h \exp\left(-E(v, h; \theta)\right)$$

即 RBM 的归一化常数。具体 RBM 的介绍可参见文献 [5] 的第 2 章。 **5种算法**

一般常用的估计 RBM 模型归一化常数的方法有 AIS（见文献 [5] 的第 4 章），SAMS[6]，Wang-Landau 算法，TAP 方法 [3]，RTS[2] 等等，请同学们在充分调研和阅读相关文献 (包括但不限于附件中的文献) 的基础上，设计出有效的采样算法，对四种不同的 RBM 模型，进行归一化常数的估计，结合实验结果，分析算法性能，特别是比较采用不同的采样算法下估计的准确性和计算效率。在得到归一化常数估计的结果基础之上，请同学们在真实的测试数据上算出不同模型的的似然值并进行比较。

选做：同学们也可以尝试自己重新训练 RBM，以在测试数据上得到更好的似然值。

附件文件说明：

此次四种不同的 RBM 模型均是在 MINIST 手写数字数据集上训练出来的，观测变量均为 28*28，隐变量分别是 10, 20, 100, 500。各个模型的参数已存为 h10.mat、h20.mat、h100.mat、h500.mat 四个文件中。训练数据为 60000 个，存入 train.mat 中，测试数据为 10000 个，存入 test.mat 中。所有数据存在 data 文件中

三. 具体要求

(a) 希望同学充分调研和阅读相关文献，积极动脑 + 动手，取得有**自己见解**的结果，整理成最终报告。

(b) 最终提交包括:

i. 报告
报告的书写要求参见《Project 报告撰写建议》。

ii. 源程序 其中务必包括一个命名为run.m的文件，不带任何参数可直接运行得到结果，输出四个不同的 RBM 模型的归一化常数估计（你的最好估计）和在测试数据上的总似然值，将 8 个数记录在列矢量 z 中，并存成一个 z.mat 文件。确保设置好相对路径，如果程序不能正常运行输出结果，将视情况扣分。

将以上两项一起压缩打包，命名为“学号 _ 姓名.rar”进行提交。

(c) 评分标准：报告书写清晰和规范，工作新意及深入程度，工作量及完整程度，归一化常数的估计效果。

(d) 一旦发现抄袭，计零分。

(e) 请大家在规定截止时间前提交。晚交的处理方法如下：按晚交天数，以 90% 的几何级数进行折扣。晚交时间在 (0, 24 小时]，按 90% 折扣。晚交时间在 (24 小时, 48 小时]，按 90%*90% 折扣。以此类推。

参考文献

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] David Carlson, Patrick Stinson, Ari Pakman, and Liam Paninski. Partition functions from rao-blackwellized tempered sampling. *arXiv preprint arXiv:1603.01912*, 2016.

[3] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In *Advances in Neural Information Processing Systems*, pages 640–648, 2015.

[4] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[5] Ruslan Salakhutdinov. *Learning deep generative models*. PhD thesis, University of Toronto, 2009.

[6] Zhiqiang Tan. Optimally adjusted mixture sampling and locally weighted histogram analysis. *Journal of Computational and Graphical Statistics*, (just-accepted), 2015.

[7] 林元烈. 应用随机过程. 清华大学出版社, 2002.