Math 基础：矩阵求导，y: scalar/m-vec
x: scalar/n-vec

$\frac{\partial \vec{y}}{\partial x}=\begin{pmatrix}\frac{\partial y_1}{\partial x}\\ \vdots\\ \frac{\partial y_m}{\partial x}\end{pmatrix}$ $\frac{\partial y}{\partial \vec{x}}=\begin{pmatrix}\frac{\partial y}{\partial x_1}\\ \vdots\\ \frac{\partial y}{\partial x_n}\end{pmatrix}$ $\frac{\partial \vec{y}}{\partial \vec{x}}=\begin{bmatrix}\frac{\partial y_1}{\partial x_1}&\cdots&\frac{\partial y_1}{\partial x_n}\\ \vdots&&\\ \frac{\partial y_m}{\partial x_1}&\cdots&\frac{\partial y_m}{\partial x_n}\end{bmatrix}$

$\left(\frac{\partial Y}{\partial x}\right)_{i,j}=\left[\frac{\partial Y_{i,j}}{\partial x}\right]$ $\left(\frac{\partial y}{\partial \vec{x}}\right)_{i,j}=\frac{\partial y}{\partial x_{i,j}}$

① $\frac{\partial a^T\vec{x}}{\partial \vec{x}}=\frac{\partial \vec{x}^T\cdot a}{\partial \vec{x}}=a$ ☆ 重要求导！

② $\frac{\partial \vec{x}^TA\vec{x}}{\partial \vec{x}}=(A+A^T)\vec{x}$

Lagrange：对于优化问题：$\min_x f_0(x)$, s.t.: $f_i(x)\le 0, i=1,\cdots,m$; $h_i(x)=0, i=1,\cdots,n$

则：$L(x,\lambda,\nu)=f_0(x)+\sum_{i=1}^m \lambda_i f_i(x)+\sum_{j=1}^n \nu_j h_j(x)$

其中λ和ν是拉格朗日乘子，$\lambda_i\ge 0$，ν_i无约束

KKT条件：$\begin{cases} f_i(x)\le 0, i=1,\cdots,m \\ h_i(x)=0, i=1,\cdots,n\end{cases}$，$\lambda_i\ge 0$

且 $\lambda_i f_i(x)=0, i=1,\cdots,m$，$\nabla_x L(x,\lambda,\nu)=0$

奇异值分解：$A=U\Sigma V^T$，$U\in\mathbb{R}^{m\times m}$，$U^TU=I$；$V\in\mathbb{R}^{n\times n}$，$V^TV=I$；$\Sigma\in\mathbb{R}^{m\times n}$，$(\Sigma)_{ii}=\sigma_i$，且$\sigma_i$为非负数且满足：$\sigma_1\ge\sigma_2\cdots\ge 0$

σ_i由来：$A^TAx=\lambda x$，$|A^TA-\lambda I|=0$

$\sigma_i=\sqrt{\lambda_i}$，而$U$中每个列向量为$AA^T$中eigen vector
V中每个列向量为A^TA中eigenvector

MAP：$\hat{\theta}=\arg\max_\theta p(\theta|D)\propto p(\theta)\cdot p(D|\theta)$

MLE：$\hat{\theta}=\arg\max_\theta p(D|\theta)$

GD：$x^{k+1}\leftarrow x^k-\alpha_k\nabla f(x^k)$

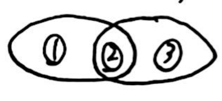
ELBO ≥ 0

$\log p(X)=\int q(z)\log\frac{p(X,z)}{q(z)}dz+KL(q(z)\|p(z|X))$

$H(X)=-\sum_{x\in X}p(x)\log p(x)=\mathbb{E}[\log\frac{1}{p(x)}]$

$H(Y|X)=-\sum_{x\in X}p(x)\sum_{y\in Y}p(y|x)\log p(y|x)$
$=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{1}{p(y|x)}$

$I(X;Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=H(X)-H(X|Y)$ or: $H(Y)-H(Y|X)$



$H(X)$: ①+② $H(Y)$: ②+③
$H(X|Y)$: ① $H(Y|X)$: ③ $I(X;Y)$=②

决策树：递归地选最优属性 (metric: I; Gini; error)
ID3 CART
它有可解释性，易overfit (可用pruning缓解)，对缺失属性值敏感，inductive bias 少。
dim 样本数

KNN：Lazy learning $O(1)$; Predict: $O(MN)$
它对noise、K值、度量metric敏感，对数据的量与质量要求高；易overfit；可用于分类&回归；表现有理论保障

Perceptron：$proj_b a=\frac{|a^Tb|}{\|b\|_2^2}b$

初：$w=\vec{0}$ $b=0$，$\hat{y}=sign(w^Tx+b)=\begin{cases}+1, & w^Tx+b\ge 0\\ -1, & else\end{cases}$

$w\leftarrow w+y^{(t)}x^{(t)}$ $b\leftarrow b+y^{(t)}$ if wrong $\hat{y}$；$\vec{x}=\begin{bmatrix}\vec{x}\\1\end{bmatrix}$

$\Rightarrow \theta=\begin{bmatrix}\vec{w}\\b\end{bmatrix}$，$\theta\leftarrow\theta+y^{(t)}x^{(t)}$ if $\hat{y}\ne y$

在上述中，w可视为x的线性组合；有overfitting

mistake $\le(R/\gamma)^2$，γ为margin*，$\|x^{(i)}\|\le R$

*θ^*为完美超平面参数且$\|\theta^*\|=1$，$|\theta^*x^{(i)}|\ge\gamma$

若Data Linearly Unseperable，则不work；且有overfit

SVM：Formulation，input：$S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$

Find $\arg\min_{w,\xi_1,\cdots,\xi_m}\|w\|^2+C\sum_i\xi_i$ s.t.：
$\forall i$，$y_i(wx_i)\ge 1-\xi_i$，$\xi_i\ge 0$

Dual $\Longleftrightarrow$ $\arg\min_\alpha\frac{1}{2}\sum_i\sum_j y_iy_j\alpha_i\alpha_j x_i\cdot x_j-\sum_i\alpha_i$

s.t.，$0\le\alpha_i\le C_i$，$\sum_i y_i\alpha_i=0$

points (x_i,y_i) whose $\alpha_i\ne 0$: support vector

or: $y_i(wx_i)=1$; final: $w=\sum_i\alpha_i y_i x_i$

Kernel：将样本映射至高维以solve 线性不可分问题。但也希望在高维中算内积，可用原始数据在低维中的操作得到！↳计算大！；即：

$K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\phi(x_j)$

怎样的K可作？① 对称函数 ②核矩阵半正定$^\Delta$

*：$K_{ij}=K(x_i,x_j)$ Δ：$\forall x\ne\vec{0}$，$x^TKx\ge 0$

常见：$(x_i^Tx_j)^d$；$\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$；$\exp(-\frac{\|x_i-x_j\|}{\sigma})$；
gauss laplace
$\tanh(\beta x_i^Tx_j+\theta)\to$ sigmoid

若k_1、k_2核func，则$\gamma_1k_1+\gamma_2k_2$也是，$g(x_1)K(x_1,x_2)g(x_2)$也是

Linear Regression：回归！$\arg\min_{w,b}\sum_{i=1}^m(y_i-wx_i-b)^2$

闭式解：$w=\frac{\sum_{i=1}^m y_i(x_i-\bar{x})}{\sum_{i=1}^m x_i^2-\frac{1}{m}(\sum_{i=1}^m x_i)^2}$，$b=\frac{1}{m}\sum_{i=1}^m(y_i-wx_i)$

更一般地：$w^*=\arg\min_w[(\vec{y}-Xw)^T(\vec{y}-Xw)]$，$w^*=(X^TX)^{-1}X^Ty$

Logistic Regression：虽回归，但依然分类任务！

$y=\frac{1}{1+e^{-(w^Tx+b)}}$，$\ln\frac{P(y=1|x)}{P(y=0|x)}=w^Tx+b$

↳视作正例可能性！$(w^Tx+b\to\beta x)$.

$p(y_i|x_i;w,b)=y_i\,p_1(\hat{x}_i;\beta)+(1-y_i)\,p_0(\hat{x}_i;\beta)$.

$p_1(\hat{x}_i;\beta)=p(y=1|\hat{x}_i,\beta)=\frac{e^z}{1+e^z}$，$p_0=\frac{1}{1+e^z}$

$l(\beta)=\sum_{i=1}^m(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i}))$

$\nabla_\beta l(\beta)=\sum_{i=1}^m\hat{x}_i(P(Y=1|\hat{x}_i,\beta)-y_i)$ (有时x前有- (PPT中))

☆：$\theta_k\leftarrow\theta_k+\lambda(h_\theta(x^{(i)})-y^{(i)})x_k^{(i)}$

GD&BP：GD：$\theta\leftarrow\theta-\gamma\nabla_\theta J(\theta)$，用所有样本

SGD：$\theta\leftarrow\theta-\gamma\nabla_\theta J^{(i)}(\theta)$ (i are shuffled) per

GD γ个波动↑(曲折)；SGD比GD更易曲折，每步 chan

Minibatch：取k个样本子集作GD。$\Sigma O(\log\frac{1}{\epsilon})$. $T:O(\frac{1}{\epsilon})$

BP：Basic：$f_L(w_L,f_{L-1}(\cdots f_1(w_1,x))\cdots)$，$\varphi(z)=\frac{1}{1+e^{-z}}$

$\varphi'(z)=\varphi(z)(1-\varphi(z))$，好算导数；GD、SGD：Step：$O(\frac{1}{\epsilon})$ time $O(n)$

W^L target

$\to x^{L-1} \to S^L \rightleftarrows x^L \to \text{Loss}$

闭式解:$(X^TX)^{-1}$ 求逆 $O(ND^2)$, $S: O(D^3)$

$\frac{\partial L}{\partial W^L} = \frac{\partial L}{\partial x^L} \cdot \frac{\partial x^L}{\partial S^L} \cdot \frac{\partial S^L}{\partial W^L}$, 链式法则

Recom Sys & Matrix Factorization

Unconstrained: $R \in \mathbb{R}^{m\times n}$, R's rand $k << \min(m,n)$, $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{n\times k}$, s.t., $R = UV^T$

$J(U,V) = \frac{1}{2}\|R - UV^T\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(R_{ij} - U_{j,\cdot} \cdot V_{\cdot,i}^T)^2$

where R_{ij} is known. 可加 $\frac{\lambda}{2}(\|U\|_2^2 + \|V\|_2^2)$ 正则化

$\nabla U_{j,\cdot} J_{ij}(U,V) = -E_{ij} V_{\cdot,i}^T + \lambda U_{j,\cdot}$

$\nabla V_{\cdot,i}^T J_{ij}(U,V) = -E_{ij} U_{j,\cdot} + \lambda V_{\cdot,i}$ (SGD)

$U_{j,\cdot} \leftarrow U_{j,\cdot} - \eta \nabla U_{j,\cdot} J$; $V_{\cdot,i}^T \leftarrow V_{\cdot,i}^T - \eta \nabla V_{\cdot,i}^T J$

Or: Fix V^T, 解 U; Fix U, 解 V^T; 循环

若已知 SVD $R = Q\Sigma P^T$, 则可 truncate $\to k$.

$U \triangleq Q_k \Sigma_k$, $V \triangleq P_k$, 开始优化

Constrained: $U, V = \underset{U,V}{\operatorname{argmin}} \frac{1}{2}\|R - UV^T\|_2^2$ s.t.

$U_{ij} \geq 0$, $V_{ij} \geq 0$

集成: Boosting: Goal: $\hat{h}(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$

Adaboost $\in$ Boosting: $h_t(x)$ 间有依赖关系

$D_1(x) = 1/m$

for $t = 1, \cdots, T$ do ← T: 基学习器个数

choose h_t (一般: weak) (and train)

算 error rate ϵ $(x \sim D_t)$

if $\epsilon > 0.5$, break (太差!)

$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

$D_{t+1}(x) = \frac{D_t(x)}{Z_t *} \times \begin{cases} \exp(-\alpha_t), \text{if 预测对} \\ \exp(\alpha_t), \text{if 错} \end{cases}$

$= \frac{D_t(x)}{Z_t} \exp(-\alpha_t h_t(x) f(x))$

end for

$h(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$ ⇒ 只能作二分类

*: $Z_t = e^{\alpha_t}\epsilon_t + e^{-\alpha_t}(1-\epsilon_t)$, $\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$

$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$

△: 对于无法接受带权样本的基算法, 可重采样

Bagging: for $t = 1, \cdots, T$ do

for $s = 1, \cdots, S$ do ← 每个 h_t 用自己的 S_t

$i_s \sim \text{Uniform}(1, \cdots, N)$

$S_t = \{(x^{(i_s)}, y^{(i_s)})\}_{s=1}^{S}$

$h_t = \text{train}(S_t)$

能作分类 & 回归

return $\hat{h}(x) = \text{aggregate}^*(h_1, \cdots, h_T)$

*: 如分类中, 为 majority vote; 回归中, average

Feature Bagging: 随机选 S 个 feature. 每个样本只关注这 S 个, 训; 每个 h_t, D_t 重新做一个新的

Random Forest: 抽个 sample 集 ⇒ 基决策树: 每个结点从属性集合选子集, 在子集上选最优属性。

*: $k = \log_2 d$; $k=1$: 随机; $k=d$: 原决策树 learn

△: Boottrapping: 抽 N 个样本 (有放回); 36.8% 不在 sample 集 $\lim_{m\to\infty}(1-\frac{1}{m})^m = 1/e$

表现不错, 且收敛性与 Bagging 相似

K-means: Initialize $c = \{c_1, \cdots, c_k\}$ *

Repeat till Converge:

for i in $\{1, \cdots, N\}$

$z^{(i)} \leftarrow \underset{j}{\operatorname{argmin}} \|x^{(i)} - c_j\|_2^2$

for j in $\{1, \cdots, k\}$

$c_j \leftarrow \operatorname{argmin} \sum_{i: z^{(i)} = j} \|x^{(i)} - c_j\|_2^2$

$\Rightarrow c_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$ (μ)

*: 可随机, 也可 Furthest Point Heuristic, 也可:

++: 先任选 c_1, pick c_j from 分布:

$P(c_j = x^{(i)}) \propto \min_{j'<j} \|x^{(i)} - c_{j'}\|^2$, $x^{(i)}$ are not picked

EM: E: $Q(\theta|\theta^t) = \sum_z \underbrace{P(z|X,\theta^t)}_{\text{posterior}} \log \underbrace{P(X, z|\theta)}_{\text{prior}}$

M: $\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} [Q(\theta|\theta^t) + \log p(\theta)]$ 加: MAP 不加: MLE

GMM: 假设点生成: K 个高斯以 π_k 概率随机选一个. 然后点生成服从分布 $N(x_n|\mu_k, \Sigma_k)$

则 $\log p(X, z|\theta) = \sum_{i=1}^{N} \log \pi_{z_i} + \log N(x_i|\mu_{z_i}, \Sigma_{z_i})$

$r_{ik} = p(z_i = k | x_i, \theta^{old})$

$= \frac{p(x_i|z_i=k, \theta^{old})\, p(z_i = k|\theta^{old})}{p(x_i|\theta^{old})}$

$= \frac{\pi_k^{old} N(x_i|\mu_k^{old}, \Sigma_k^{old})}{\sum_{j=1}^{k} \pi_j^{old} N(x_i|\mu_j^{old}, \Sigma_j^{old})}$ E

M: $Q(\theta, \theta^{old}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}[\log \pi_k + \log N(x_i|\mu_k, \Sigma_k)]$

且: $\sum_k \pi_k = 1$, $\mathcal{L} = Q(\theta, \theta^{old}) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$

$\Rightarrow \begin{cases} \pi_k^{new} = \frac{1}{N}\sum_{i=1}^{N} r_{ik} \\ \mu_k^{new} = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}} \\ \Sigma_k^{new} = \frac{\sum_{i=1}^{N} r_{ik}(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T}{\sum_{i=1}^{N} r_{ik}} \end{cases}$

GMM ①高斯 sperical, have 相同 weights & covariance ②hard-assign points to 高斯 ⇓ kMeans

PAC: 主成分分析:

① $x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$: X ② XX^T (协方差)

③ 特征值分解 ④ 取最大 d' 个 eigenvalue 对应的 eigenvector, 组成 $w^* \in \mathbb{R}^{d \times d'}$

则 $w^{*T} x$ 后进入低维

CNN: 注重于 local connectivity; parameter sharing; pooling/subsampling hidden units

卷积中: kernel 数之和为 1; 卷积中还有 bias!!

Input: (H_{in}, W_{in}, C_{in}), 卷积: (K, K, C_{in}). 填充 P, stride S, 则输出 $(H_{out}, W_{out}, C_{out})$:

$\begin{cases} H_{out} = \left\lfloor \frac{H_{in} + 2P - K}{S} \right\rfloor + 1 \\ C_{out} = N \\ W_{out} = \left\lfloor \frac{W_{in} + 2P - K}{S} \right\rfloor + 1 \end{cases}$

一个卷积一个 (不是一个通道)

(Para of bias = C_{out})

每个 channel 的卷积结果是相加而非平均

Pooling layers: e.g., Max pooling (no learnable 参数)

CNN 两个关键性质: Equivariance: 输入经过变换后, 输出也经历某种变换; Invariance: 输出不变! (平移不变)

↳ 卷积 (e.g. Conv w.r.t 平移) ↳ 池化 (e.g.: conv+pool+FC ⇒ 不变)

SVM推导：约束 $w^T x_*^- + b = -1, w^T x_*^+ + b = 1$

$\gamma = \frac{1}{\|w\|}$，$d = \frac{2}{\|w\|}$，则构造问题

$\min_{w,b} \frac{1}{2}\|w\|^2$, s.t. $y_i(w^T x_i + b) \geq 1$

$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(w^T x_i + b))$

$\alpha = (\alpha_1; \alpha_2; \cdots; \alpha_m)$. 则KKT条件：

① primal：$y_i(w^T x_i + b) \geq 1$ ② dual：$\alpha_i \geq 0$

③ Complementary：$\alpha_i (y_i(w^T x_i + b) - 1) = 0$

④ Stationary：$\nabla_w L = 0$

对w偏导：$w = \sum_{i=1}^{m} \alpha_i y_i x_i$，代回：

b：$\sum_{i=1}^{m} \alpha_i y_i = 0$

$\max_\alpha \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$, $\alpha_i \geq 0$，解出α后：

$f(x) = w^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$

Why biased? $\mathbb{E}(\hat{\sigma}^2_{MLE}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2\right]$

则 $E[\hat{\sigma}^2_{MLE}] = \frac{n-1}{n}\sigma^2$, bias $= -\frac{\sigma^2}{n}$

Probably Approximate Correct learning：

$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \leq \delta$, $n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$

K-mean++：achieves a $O(\log k)$ approximation to optimal clustering；这种思路对初始化法均有益

ELBO：$\log p(X) = \log \frac{p(X,Z)}{p(X,Z)} \cdot p(X)$

$= \left[\log \frac{p(X,Z)}{q_1(Z)} \frac{q_1(Z)}{p(Z|X)}\right] \cdot \int q_1(Z)dZ$

$= \int q_1(Z) \log \frac{p(X,Z)}{q_1(Z)} dZ + \int q_1(Z) \log \frac{q_1(Z)}{p(Z|X)} dZ$.

$= \underbrace{\int q_1(Z) \log \frac{p(X,Z)}{q_1(Z)} dZ}_{\text{后验 } \hookrightarrow \text{ELBO}} + \underbrace{KL(q_1(Z)\|p(Z|X))}_{\text{先验 } \geq 0}$

$\text{ELBO} = \int q_1(Z)\log p(X,Z)dZ - \int q_1(Z)\log q_1(Z)dZ$

$= E_{q_1(Z)}[\log p(X,Z)] + H(q_1(Z))$

Perceptron mistake bound：$(R/\gamma)^2$

Proof：设θ^*为完美类超平面参数，且$\|\theta^*\| = 1$

且 $\forall i$ $y^{(i)} \cdot (\theta^* \cdot x^{(i)}) \geq \gamma$ $\quad \|\theta^{(k+1)}\| \geq k\gamma$

则 $\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$ $\quad \Uparrow$

$\theta^{(k+1)} \cdot \theta^* \geq \theta^{(k)} \cdot \theta^* + \gamma$，又$\theta^{(0)}$空初始化

同时 $\|\theta^{(k+1)}\|_2^2 = \|\theta^{(k)} + y^{(i)} x^{(i)}\|_2^2$ $\quad$ <0，分错的！

$= \|\theta^{(k)}\|_2^2 + y^{(i)2}\|x^{(i)}\|_2^2 + 2y^{(i)}(\theta^{(k)} \cdot x^{(i)})$

$\leq \|\theta^{(k)}\|_2^2 + \|x^{(i)}\|_2^2$

$\therefore k^2\gamma^2 \leq \|\theta^{(k+1)}\|_2^2 \leq kR^2$, $k \leq \frac{R^2}{\gamma^2}$

Logistics：$p(y_i|x_i; w, b) = y_i p_1(\hat{x}_i; \beta) + (1-y_i)p_0(\hat{x}_i;\beta)$

$l(\beta) = \sum_{i=1}^{m}(-y_i \beta^T \hat{x}_i + \ln(1 + e^{\beta^T \hat{x}_i}))$

$\frac{\partial l(\beta)}{\partial \beta} = -\sum_{i=1}^{m} \hat{x}_i (y_i - p_1(\hat{x}_i; \beta))$

Adaboost：$\epsilon = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \neq H_{final}(x_i))$

$\leq \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i(\sum_{t=1}^{T} \alpha_t h_t))$

$D_{T+1}(i) = \frac{D_T(i)}{Z_T} \exp(-\alpha_T y_i h_T(x_i))$

$D_T(i) = \frac{D_{T-1}(i)}{Z_{T-1}} \exp(-\alpha_{T-1} y_i h_{T-1}(x_i))$

$\vdots$

$D_1(i) = 1/n$

$\therefore \epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)) = \sum_{i=1}^{n} D_{T+1}(i) \prod_{t=1}^{T} Z_t$ $\quad$ 1

$\therefore \epsilon \leq \prod_{t=1}^{T} Z_t$，又 $Z_t = e^{\alpha_t}\epsilon_t + e^{-\alpha_t}(1-\epsilon_t)$

$\frac{\partial Z_t}{\partial \alpha_t} = \epsilon_t e^{\alpha_t} - (1-\epsilon_t)e^{-\alpha_t} \Rightarrow \alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$

$\frac{\partial^2 Z_t}{\partial \alpha_t^2} = \epsilon_t e^{\alpha_t} + (1-\epsilon_t)e^{-\alpha_t} > 0$ $\quad Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$

$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

或：$\ell_{exp}(H|D) = \mathbb{E}_{x\sim D}[e^{-f(x)H(x)}]$

$\frac{\partial \ell}{\partial H(x)} = e^{-H(x)}p(f(x)=1|x) + e^{H(x)}p(f(x)=-1|x) = 0$

$H(x) = \frac{1}{2}\ln\frac{p(f(x)=1|x)}{p(f(x)=-1|x)}$