Computational & Data-driven Physics

Yilin YE

Update: January 2, 2023

Abstract

Master 2 ICFP, September - December 2022, Friday 14h00 - 18h00.

Computational Physics: **Alberto ROSSO** (LPTMS);

Data-driven Physics: **Rémi MONASSON** (ENS) & **Simona COCCO** (ENS)

See more information on: *http://lptms.u-psud.fr/wiki-cours/index.php/CoDaDri2*

Contents

Part I

COMPUTATIONAL

1 02/09/22: History and Motivation?

Not enough material for test, problem following

neuron ⟶ $p^+ + e^-$

neutron + photon ⟶ lots of neutrons.

If before the decay, more than 1 neutron, then would exceed the bound towards explosion.

Ulam

probability to win = nb of games / 54!

ENIAC huge computer bigger than a room. ≃ to nb. of winning / 10000 cards series To realize Monte Carlo, we have to take extremely large number tests.

Random number generator between 0 and 1.

MAMIAC, No metropolis

$$H = \sum_{i=1}^{N} \frac{p_x^2 + p_y^2 + p_z^2}{2m} + \sum_{i<j} V(r_i - r_j) \qquad \text{probability}[r1, r2, ...,] = \exp(-\beta H)/Z$$

$$\langle O \rangle_\beta = \frac{\int dp_i dx_j \exp(-\beta H) O}{\int dp_i dx_j \exp(-\beta H)}$$

Suppose that we would compute π, we drop small particles with random x, y coordinates inside the square with edge length equal to 2. Then count the number of particles inside the circle, and the total number. π is just equal to 4 times the ratio between these two numbers, if we drop so many particles.

$$\frac{A_\circ}{A_\square} = \frac{\int dxdy\pi(x,y)O(x,y)}{\int dxdy\pi(x,y)}$$

where $\pi(x, y)$ =1 if (x,y) in the square; O(x,y) = 1 if (x,y) in the circle.

$$\frac{A_\circ}{A_\square} = \frac{\pi r^2}{(wr)^2} = \frac{\pi}{4} \simeq \frac{\sum_{i=1}^{N} O(x,y)}{N} \pm \frac{std(O)}{\sqrt{N}}$$

where std is in order of 1.

$$std(O) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} O^2(x,y) - \left(\frac{\sum O(x,y)}{N}\right)^2}$$

2 09/09/22: Markov chain Monte Carlo (MCMC)

$p_i(x, y) = 1$ if allowed

1. start from an allowd position (x_0, y_0)

2. (x_t, y_t), $x_{\text{new}} = x_t + \delta x$, $y_{\text{new}} = y_0 + \delta y$

If the new coordinate is allowed, then we select as $x(t+1) = x_{\text{new}}$ same for y. If not allowed, we would follow $x(t+1) = x(t)$ and $y(t+1) = y(t)$, namely rejecting out-of-value and counting again the last position.

Acceptance ratio = nb time you acceptance / time

$$\frac{A_{\circ}}{A_{\square}} = \ldots \pm \frac{\text{std}(O)}{\sqrt{N_{\text{eff}}}}$$

where N_{eff} independant $\ll N_{\text{total}}$

If δ is too small, P_{acc} is towards 1; if δ is too large, P_{acc} turns to 0.

halt THUMB RULE, means that we should keep P_{acc} near 1/2.

$$T = \begin{pmatrix} P_{1\to1} & P_{2\to1} & P_{3\to1} \\ P_{1\to2} & P_{2\to2} & P_{3\to2} \\ P_{1\to3} & P_{2\to3} & P_{3\to3} \end{pmatrix} \qquad \sum_j P_{i\to j} = 1 \qquad P_{i\to j} \geq 0$$

In general is not symmetric matrix, so eigenvalues can be complex..

3 16/09/22: IMPORTANCE SAMPLING

- Direct Sampling $\{X_1, \cdots, X_N\}$. X is drawn from $\Pi(X_i)$. Rejection is to draw random parts
- Markov Chain $\{X_1, \cdots, X_{N-1}; X_N\}$

IMPORTANCE SAMPLING Consider normalized probability: $P(x)$, and typically $N_{\text{total}} \sim 10^4$

$$\langle O \rangle = \int Q(x)P(x)dx \simeq \frac{1}{N_{\text{total}}} \sum_{i=1}^{N_{\text{total}}} O(x_i)$$

then with $c = 20$, and $O(x) = \theta(-x) = 1$ only if $x < 0$, we have the integral:

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-c)^2\right] dx \simeq 2.75 \times 10^{-89} = p$$

$$X_i = \sqrt{-2\ln Y_1} \cos(2\pi Y_2) + c$$

Suppose $O = 1$ with probability: $(1-p)$; $O = 0$ with probability: p. We have the standard derivation:

$$\text{std}(O) = \sqrt{\underbrace{(1^2(1-p) + 0^2 p)}_{\langle \text{O}^2 \rangle} - \underbrace{p^2}_{\langle \text{O} \rangle^2}} = \sqrt{p(1-p)}$$

$$\text{error} = \frac{\text{std}(O)}{\sqrt{N_{\text{total}}}} = \frac{\sqrt{p(1-p)}}{\sqrt{N_{\text{total}}}}$$

$P(X)$ the probability we want to sample; $Q(X)$ PDF defined in the same domain. Thus we have

$$\langle O\rangle = \int dXP(X)O(X) = \int dXO(X)Q(X)\underbrace{\frac{P(X)}{Q(X)}}_{\text{w(x)}} = \int dX\,\underbrace{O(X)w(X)}_{\widetilde{\text{O}}\text{(X)}}Q(X) = \frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}} O(X_i)w(X_i)$$

where X_i is drawn from $Q(X)$.

Exponential ..? t is a p

$$Q(X) = \frac{e^{tx}P(X)}{\int e^{tX'}P(X')dX'}$$

$$W(X) = \frac{P(X)}{Q(X)} = e^{-tX}\int e^{tX'}P(X')dX'$$

$$\int \frac{1}{\sqrt{2\pi}}\exp\left(tX - \frac{1}{2}c^2 - \frac{1}{2}X^2 + cX\right) = \underbrace{\left\{\int dX\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-(c-t))^2\right]\right\}}_{=1}\cdot\exp\left(\frac{1}{2}t^2 + ct\right) = \exp\left(\frac{1}{2}t^2 + ct\right)$$

Part II

Data-Driven

Contact: monasson@phys.ens.fr

Reference: « ***From Statistical Physics to Data-Driven Modelling*** *with Applications to Quantitative Biology* », by Simona Cocco, Rémi Monasson, and Francesco Zamponi, 2022, Oxford University Press. ISBN 978-0-19-886474-5, DOI: 10.1093/oso/9780198864745.001.0001

4 14/10/22: Bayes'

This first chapter presents basic notions of Bayesian inference, starting with the definitions of elementary objects in probability, and Bayes' rule.

data ⟶ get mechanism, model? predictions, generation..

Bayes' inference: P(data, model), joint probability with data & model.

Suppose y is a random variable $\in \{a_1, a_2, \cdots, a_L\} = A$. The probability $P(y = a_i)$ writes $P(a_i)$ or just P_i. We have property like normalization: $\sum_i P_i = 1$ with $P_i \geq 0$.

As for joint probability, $y = (a, b) \in A \times B$, $P(a_i, b_j) = P_{ij}$.

- marginal $P(a_i) = \sum_j P(a_i, b_j)$.
- independence, $P(a_i, b_j) = P(a_i) \times P(b_j)$
- conditional probability, $P(a_i|b_j) = \frac{P(a_i, b_j)}{P(b_j)}$
- normalization, $\sum_i P(a_i|b_j) = 1$.

Bayes formula

$$P(a|b) = \frac{P(a,b)}{P(b)} = \frac{P(b|a) \times P(a)}{P(b)}$$

with a = model, b = data, $P(b|a) = P(\text{data}|\text{model})$ = likelihood of model, $P(a)$ as prior, $P(b)$ called evidence, $P(a|b)$ refers to posterior distribution of models given the data.

The German Tank Problem Suppose total numbers of tanks as N, after one battle some were destroyed. Number of tanks destroyed in the i-th observation writes y_i, we show $1 \leq y_1 < y_2 < y_3 < \cdots < y_k \leq N$, $Y = \{y_1, \cdots, y_K\}$ with $K \leq y_K$. Likelihood, the probability of data given the data $P(Y|N)$, where N is number of models we infer.

$$C_N^K = \begin{pmatrix} N \\ K \end{pmatrix} = \frac{N!}{K!(N-K)!} \qquad P(Y|N) = \frac{1}{C_N^K} \qquad P(N|Y) = \frac{P(Y|N) \times P(N)}{P(Y)}$$

where we have uniform $P(N)$ over N. For evidence, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$, $\Gamma(n) = (n-1)!$, so

$$P(Y) = \sum_{N\geq y_K} \frac{1}{C_N^K} = \sum_{N=y_K}^{\infty} \frac{K!(N-K)!}{N!} = K! \sum_{N=y_K}^{\infty} \frac{(N-K)!}{N!} = K!\cdot\frac{1}{K-1}\frac{\Gamma(y_K-K+1)}{\Gamma(y_K)} = \frac{K!(y_K-K)!}{y_K!}\frac{y_K}{K-1}$$

Here, we pose the uniform $P(N)$, and thus

$$P(N|Y) = \frac{P(Y|N)}{P(Y)} = \frac{K-1}{y_K}\frac{C_{y_K}^K}{C_N^K}$$

Remark: it depends on Y through y_K only...

More likely value of $N = y_K$; the average value of $N > y_K$.

$$\langle N\rangle = \sum_{N\geq y_K} P(N|Y)\cdot N = \frac{K-1}{K-2}\times(y_K-1) = \frac{1}{1-\frac{1}{K-1}}(y_K-1) \simeq y_K + \frac{y_K-K}{K} > y_K$$

$$\left\langle N^2\right\rangle - \langle N\rangle^2 = \frac{K-1}{(K-2)^2(K-3)}(y_K-1)(y_K-K+1)$$

Laplace's birth rate problem Data Y from the city of Paris: Births between 1745 and 1770, girls: $y = 241945$, boys: $n-y = 251527$. Suppose θ = the birth rate of girls.

$$\int_{1/2}^{1} d\theta P(\theta|Y) = ?$$

Still we have Bayes' formula

$$P(\theta|Y) = \frac{P(Y|\theta)\times P(\theta)}{P(Y)}$$

$$P(y|\theta, n) = \begin{pmatrix} n \\ y \end{pmatrix}\theta^y(1-\theta)^{n-y}$$

$$P(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

It's an example of Beta distribution

$$\mathrm{Beta}(\theta,\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

with $\alpha = y+1 \geq 1, \beta = n+1-y \geq 1$.

$$\theta_{\max} = \frac{\alpha-1}{\alpha+\beta-2} \qquad \langle\theta\rangle = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

With data above, we obtain $\theta_{\max} = 0.490291$, $\langle\theta\rangle = 0.490291$, $\mathrm{var}(\theta) = 0.0007117$. $n = \alpha_\beta - 2$, $y = \alpha - 1$, $\theta_{\max} = y/n$.

$$P_{\text{post}}(\theta|y) = \frac{\alpha^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = B(\alpha,\beta)\exp\{n\underbrace{[\theta_{\max}\log\theta + (1-\theta_{\max})\log(1-\theta)]\}}_{\mathrm{f}(\theta)}$$

The binary variable $y = 0, 1$ with probability= p_0, p_1, respectively. We have the entropy $S = -p_0 \log p_0 - p_1 \log p_1$.

$$P_{\text{post}} = \text{const.} \exp\left[n f(\theta)\right] \underbrace{\simeq}_{\theta \to \theta_{\max}} \text{const.} \exp\left\{n\left[f(\theta_{\max}) - \frac{1}{2}(\theta - \theta_{\max})^2 f''(\theta_{\max})\right]\right\}$$

$$\begin{aligned} B &= \int_0^1 d\theta \exp\left[n f(\theta)\right] = \exp\left[n f(\theta_{\max})\right] \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2} n f''(\theta - \theta_{\max})^2\right] \\ &= \exp\left[n f(\theta)\right] \sqrt{\frac{2\pi}{n}\theta_{\max}(1 - \theta_{\max})} \end{aligned}$$

Near the max of Gaussian, we estimate something by variation. But 1/2 is far away from $\theta_{\max}$... Consider $u \to 0$ to extend $f(\theta)$ from integration interval to $\theta_{\max}$.

$$\int_{1/2}^{1} d\theta \frac{\exp\left[n f(\theta)\left(\theta - \frac{1}{2}u\right)\right]}{B} = \frac{1}{B}\int_0^{1/2} du \exp\left[n f\left(\frac{1}{2} + u\right)\right] = \frac{1}{B}\exp\left[n f\left(\frac{1}{2}\right)\right]\int_0^{\infty} du \exp\left[-n f'\left(\frac{1}{2}\right) u\right]$$

$$= \underbrace{\frac{1}{\sqrt{\frac{2\pi}{n}\theta_{\max}(1 - \theta_{\max})}} \frac{1}{n\left|f'\left(\frac{1}{2}\right)\right|}}_{\frac{\text{const.}}{\sqrt{\text{n}}}} \exp\left\{n\left[f\left(\frac{1}{2}\right) - f\left(\theta_{\max}\right)\right]\right\} \simeq 10^{-42}$$

5 21/10/22: Asymptotic inference

In this chapter, we will consider the case of asymptotic inference, in which a large number of data is available and a comparatively small number of parameters have to be inferred. In this regime, there exists a deep connection between inference and information theory, whose description will require us to introduce the notions of entropy, Fisher information, and Shannon information. Last of all, we will see how the maximum entropy principle is, in practice, related to Bayesian inference.

N data $y_{i=1,\cdots,N}$ drown from $P(y|\widetilde{\theta})$. $\widetilde{\theta}$ is the real distribution.

$$\widetilde{\theta} \to Y = \{y_i\} \to \theta \sim \widetilde{\theta}$$

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \propto \prod_{i=1}^{N} P(y_i|\theta) = \exp\left(\sum_{i=1}^{N} \underbrace{\log P(y_i|\theta)}_{\text{x}_\text{i}}\right) = \exp\left(N \int \mathrm{d}y P(y|\widetilde{\theta}) \log P(y|\theta) + \underbrace{\cdots}_{\sqrt{\text{N}}}\right)$$

where x_i is a Gaussian variable with mean and variance as:

$$\langle x \rangle = \int \mathrm{d}y P(y|\widetilde{\theta}) \log P(y|\theta) \qquad \int \mathrm{d}y P(y|\widetilde{\theta}) \log^2 P(y|\theta) - \langle x \rangle^2 < \infty$$

$$P(\theta|Y) \propto \exp\left[-N S(\theta)\right] \qquad S(\theta) = -\int \mathrm{d}y P(y|\widetilde{\theta}) \log P(y|\theta)$$

where we call $S(\theta)$ as cross-entropy.

Entropy of growed-truth distribution $\widetilde{S} = S(\widetilde{\theta})$. I want to show that $S(\theta) \geq \widetilde{S}$. Proof:

Let $P(y)$ and $Q(y)$ be two distributions over y. We define the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(P|Q) = \int \mathrm{d}y P(y) \log \left[\frac{P(y)}{Q(y)} \right]$$

which = cross-entropy between P and Q − entropy of P.

$$S(\theta) - \widetilde{S} = D_{\mathrm{KL}} \left[P(\cdot|\widetilde{\theta}) || P(\cdot|\theta) \right] \geq 0$$

Here · refers to y... Let me draw y from P, and $Z(y) = \frac{Q(y)}{P(y)}$ → random variable.

$$\langle \log Z \rangle = \int \mathrm{d}y P(y) \log \frac{Q(y)}{P(y)} = -D_{\mathrm{KL}}(P||Q)$$

$$\langle Z \rangle = \int \mathrm{d}y \cancel{P(y)} \frac{Q(y)}{\cancel{P(y)}} = 1 \qquad \Rightarrow \qquad \log \langle Z \rangle = 0$$

which we address Jenssen inequality

$$\langle \log Z \rangle \leq \underbrace{\log \langle Z \rangle}_{0} \qquad \leftrightarrow \qquad D_{\mathrm{KL}}(P||Q) \geq 0$$

Application below: (MLE = Maximum Likelihood Estimation)

$$P(\theta|Y) \propto \exp\left[-NS(\theta)\right] \rightarrow \theta_{\mathrm{MLE},N\rightarrow\infty} = \widetilde{\theta}$$

Last week, we found

$$P_{\mathrm{post}} \left(\theta_{\mathrm{hypo}} - \frac{\epsilon}{2} \leq \theta \leq \theta_{\mathrm{hypo}} + \frac{\epsilon}{2} \right) \simeq \exp\left[-N D_{\mathrm{KL}} \left(P_{\widetilde{\theta}} || P_{\theta,\mathrm{hypo}} \right) + \theta(N\epsilon, \log N) \right.$$

Consider D_{KL} and mean field

$$Q_\xi(y) = \prod_{i=1}^{L} \left(\frac{1 + y_x \xi}{2} \right) \qquad D_{\mathrm{KL}} = \sum_y Q_\xi(y) \log \frac{Q(y)}{P(y)}$$

$$D_{\mathrm{KL}} = \sum_{y_i = \pm 1} \prod_i \left(\frac{1 + y_i \xi}{2} \right) \log \left[\frac{\prod_i \left(\frac{1 + y_i \xi}{2} \right)}{\frac{\exp(-\beta E)}{Z}} \right]$$

since $E_{Ising} = -J \sum_{\langle i,j \rangle} y_i y_j$, we have

$$D_{\mathrm{KL}} = L \sum_{y = \pm 1} \left(\frac{1 + y_i \xi}{2} \right) \log \left(\frac{1 + y_i \xi}{2} \right) - J \sum_{y_i = \pm 1} \prod_i \left(\frac{1 + y_i \xi}{2} \right) \sum_{\langle a,b \rangle} y_a y_b + \log Z$$

For binary variable $y = \pm 1$ with $p_\pm = (1 \pm \xi)/2$, we get its entropy as $S = -p_+ \log p_+ - p_- \log p_-$. Note, D refers to the dimension of system.

$$D_{\text{KL}} = -LS_{Ising}(\xi) + \log Z = JLD\xi^2$$

$$\frac{\partial}{\partial \xi} D_{\text{KL}} = 0 = L \left\{ -JD2\xi + \left[\underbrace{+\frac{1}{2} \log \left(\frac{1+\xi}{1-\xi} \right)}_{\text{Arctanh}(\xi)} \right] \right\}$$

Note, $\xi = \tanh\left(J \cdot 2D\xi\right)$.

Fisher information

$$D_{\text{KL}}\left(P_{\widetilde{\theta}} || P_\theta\right) = \frac{1}{2}\left(\theta - \widetilde{\theta}\right)^2 I(\widetilde{\theta}) \qquad I(\widetilde{\theta}) = -\int \mathrm{d}y P(y|\widetilde{\theta}) \frac{\partial^2}{\partial \theta^2} \log P(y|\theta)\Big|_{\theta = \widetilde{\theta}}$$

Asymptotically, the Fisher information controls how wide is the posterior.

$$P(\theta|Y) \propto \exp\left[-ND_{\text{KL}} P_{\widetilde{\theta}}\right] \sim \exp\left[\frac{1}{2} NI(\widetilde{\theta}) \left(\theta - \widetilde{\theta}\right)^2\right]$$

where $\text{var}(\theta) = \frac{1}{NI(\widetilde{\theta})}$.

BEYOND asymptotics (finite N)

$$\widetilde{\theta} \longrightarrow Y \longrightarrow \theta = \theta^*(Y)$$

Maximum Likelihood Estimator (MLE): estimator $\theta^*(Y)$ =argmax $P(Y|\theta)$

Maxmimum A Posterior (MAP): $\theta^*(Y)$ =argmax $[P(Y|\theta)P(\theta)]$

Bayesian Estimator: $\theta^*(y) = \int d\theta \theta P_{post}(\theta|Y)$.

<u>Unbiased estimator</u>:

θ^* is unbiased of $\int \mathrm{d}Y \theta^*(Y) P(Y|\widetilde{\theta}) = \widetilde{\theta}$, where $Y = \{\cdots y_i\}$ variables with mean μ and var σ^2.

$$\mu^*(Y) = \frac{1}{N}\sum_i y_i \qquad \left(\sigma^2\right)^*(Y) = \frac{1}{N-1}\sum_i \left[y_i - \mu^*(y)\right]^2$$

For any unbiased estimator θ^*,

$$\text{var}(\theta^*) = \int \mathrm{d}y P(y|\widetilde{\theta}) \left[\theta^*(Y) - \widetilde{\theta}\right]^2 \geq \frac{1}{NI(\widetilde{\theta})}$$

called Cramer-Rao bound.

y = random variable from $P(y|\theta)$. SC(θ)="score" of $\theta = \frac{\partial}{\partial \theta} \log P(y|\theta)$.

$$\langle SC \rangle = \int \mathrm{d}y P(y|\theta) \frac{\partial}{\partial \theta} \log P(y|\theta) = \int \mathrm{d}y \cancel{P(y|\theta)} \frac{\partial_\theta P(y|\theta)}{\cancel{P(y|\theta)}} = 0$$

$$\left\langle SC^2 \right\rangle = \int \mathrm{d}y P(y|\theta) \left[\frac{\partial}{\partial \theta} \log P(y|\theta)\right]^2 = I(\theta)$$

Shannon information and the Maximum Entropy Principle

$$H(P_{xy} = P_x \times P_y) = H(P_x) + H(P_y)$$

leads to only one possible function log.

$$\langle H \rangle = C \sum_y P(y) \log \left[\frac{1}{P(y)} \right] = C \cdot - \sum_y P(y) \log P(y)$$

$$H(P_{xy}) = - \sum_{x,y} P(x,y) \log P(x,y)$$

$$H(P_x) = - \sum x, y P(x,y) \log P(x) \qquad H(P_y) = - \sum x, y P(x,y) \log P(y)$$

$$\underbrace{H(P_x) + H(P_y) - H(P_{x,y})}_{\text{MI(x,y)}} = \sum_{x,y} P(x,y) \log \left[\frac{P(x,y)}{P(x)P(y)} \right]$$

Principle of maximum entropy Random Shannon entropy..

Realization of random event y

$$\text{entropy} = \sum_y p(y) \log \left(\frac{1}{p(y)} \right)$$

Problem: Infer some distribution $p(y), y = y_1, \cdots, y_L$, with the knowledge that $\vec{p} \cdot \vec{f} = \sum_y p(y) f(y) = f_0$.

Question: What is $p(y)$?

Suppose I know nothing but $\sum_y p(y) = 1$. Natural choice is $p(y) = \frac{1}{L}$ (principle of indifference Laplace). P_{unif} is maximizing Shannon entropy $S_{unif} = \log L$. ... ... Calculations:

$$\max \left\{ - \sum_y p(y) \log p(y) + \lambda \left[\sum_y p(y) f(y) - f_0 \right] + \mu \left[\sum_y p(y) - 1 \right] \right\}$$

$$\frac{\partial}{\partial p(y)}(\cdot) = - \log p(y) - 1 + \lambda f(y) + \mu = 0$$

$$p(y) = e^{\mu - 1} \cdot e^{\lambda f(y)} = \frac{e^{\lambda f(y)}}{\sum_i e^{\lambda f(i)}}$$

Note the last = rewrites $e^{\mu-1}$ by normalisation as the summation of numerator. In addition, if we add another constrain $\sum_y p(y) g(y) = g_0$, we get

$$p(y) = \frac{e^{\lambda f(y) + \lambda' g(y)}}{\sum_i e^{\lambda f(i) + \lambda' g(i)}}$$

Consider y as microstatic = $\{\vec{x}_i, \vec{p}_i\}$, and its Hamiltonian $H(y) = \sum_i \frac{\vec{p}_i^2}{2m}$

$$\sum_y p(y) H(y) = E_0 \qquad \Rightarrow \qquad p(y) \propto \exp\left[\lambda H(y)\right] = \exp\left[-\beta H(y)\right]$$

Exercise: for a dice f=1,2,3,4,5,6, with $\langle f \rangle = 4$. What is $p(f)$?

6 28/10/22: High-dimensional inference and Principal Component Analysis

While we have focused so far on how to infer simple models (with one of few parameters) from few of many data, we are going to address now a more complex situation, where the dimensionality of the unknown parameter vector $\vec{\theta}$ is comparable to the number of available data. This situation, called high-dimensional inference, is relevant in many practical applications of statistical inference methods. We will concentrate on one of them, called principal component analysis (PCA).

Multivariate Gaussian distribution

$$P(y = \{y_1, \cdots, y_L\}) = \frac{\sqrt{\det \tau}}{(2\pi)^{L/2}} \exp\left(-\frac{1}{2}\sum_{i,j=1}^{L} y_i \tau_{ij} y_j\right)$$

where τ_{ij} is $L \times L$ def precision matrix. We know

$$\sum_y p(y) y_i = \langle y_i \rangle = 0 \qquad \sum_y p(y) y_i y_j = \left\langle y_i y_j \right\rangle = \left(\tau^{-1}\right)_{ij}$$

We consider that probability, and generates data $\{y^1, \cdots, y^M\} = Y = L \times M$

Principal Component Analysis

$$\tau = I - \frac{s}{1+s}\left|e\right\rangle \left\langle e\right| \qquad \tau_{ij} = \delta_{ij} - \frac{s}{1+s} e_i e_j$$

with $s > 0$ and e refers to L-dimension vector (normalized)

$$\tau^{-1} = I + s\left|e\right\rangle \left\langle e\right| = \text{Covariance.matrix} = \left\langle \mathrm{y_i y_j} \right\rangle$$

with eigenvalues of $C = \{1+S, 1, 1, \cdots\}$.

We want to infer $|e\rangle$ from $Y = \{y_{i=1,\cdots,L}^{m=1,\cdots,M}\}$. Note $\hat{C}_{ij} = \frac{1}{M}\sum_m y_i^m y_j^m$ is called correlated matrix

$$P(Y|\tau) = \left(\frac{\sqrt{\det \tau}}{(2\pi)^{L/2}}\right)^M \exp\left(\underbrace{-\frac{1}{2}\sum_{i,j}\sum_{m=1}^{M} y_i^m \tau_{ij} y_j^m}_{-\frac{1}{2}\sum_{\mathrm{ij}}\tau_{\mathrm{ij}}\left(\sum \mathrm{y_i^m y_j^m}\right)}\right) = \frac{(\det \tau)^{M/2}}{(2\pi)^{LM/2}} \exp\left(-\frac{M}{2}\sum_{ij}\hat{C}_{ij}\left(\delta_{ij} - \frac{s}{1+s}e_i e_j\right)\right)$$

$$\propto \exp\left(\frac{sM}{2(1+s)}\sum_{ij} e_i \hat{C}_{ij} e_j\right)$$

If the true value is C_{ij}, the exact value from data is written

$$\hat{C}_{ij} = C_{ij} + \frac{Z_{ij}}{\sqrt{M}}$$

where Z_{ij} is random number at order 1. Then we are interested in the elgenvectors below:

$$\sum_j \hat{C}_{ij} v_i = \lambda v_j = \sum_j C_{ij} v_i + \frac{1}{\sqrt{M}} \underbrace{\sum_j Z_{ij}}_{\sim\sqrt{\mathrm{L}}}$$

Back to $\tau = \mathbb{I}_{L\times L}$; WHAT IS $\rho(\lambda)$ eigenvalue of $\hat{C}$? and its distribution? As $M \to \infty$, the Dirac peak distribution turns to the one with width. We call that $\lambda_\pm(\frac{1}{\sqrt{M}})$ (which can ben computed lated by terms of M)

Spectrum of random covariance matrix (Marchenko, Pastur, 1967) L independent var: $y_i, i = 1, \cdots, L$. M realizations leads to

$$\hat{C}_{ij} = \frac{1}{M}\sum_{m=1}^{M} y_i^m y_j^m \qquad \text{eigenvalues} = \lambda_l$$

$$\rho(\lambda) = \frac{1}{L}\sum_e \delta(\lambda - \lambda_e)$$

If $L, M \to \infty$ at fixed ration $r = L/M \to \rho_{\mathrm{MP}}(\lambda)$

$$\rho_{\mathrm{MP}}(\lambda) = \begin{cases} \dfrac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r\lambda}, (r < 1) \\ \dfrac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi r\lambda} + \left(1 - \dfrac{1}{r}\right)\delta(\lambda), (r > 1) \end{cases}$$

with $\lambda_\pm = (1 \pm \sqrt{r})^2$.

Case of dependent variables with $C = 1 + s\,|e\rangle\langle e\,|$ top eigenvalue of empirical $\hat{C}$. (i) There is no change but still MP distribution is $s^2 < r = L/M$; (ii) $r < s^2$, there would be additional one eigen vector $|e\rangle$ with $\lambda_+ \sim S + (1 + r)$, and $\langle\hat{e}|\hat{e}\rangle^2 = 0$ for $s < \sqrt{r}$, but $\langle\hat{e}|\hat{e}\rangle^2 = \frac{1-r^2/s}{1+r/s}$ for $s > \sqrt{s}$.

7 25/11/22: Priors, regularization, sparsity

So far we have not discussed much the role of th prior distribution $p(\vec{\theta})$ in Bayes' rule. To be more precise, we have considered priors that did not depend on $\vec{\theta}$, the so-called uniform priors. However, this hypothesis is not optimal in many situations. Adequate priors are needed when the likelihood along does not define a well-conditioned or well-behaved posterior distribution...

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{..} \qquad p(\pm) = \frac{e^{\pm h}}{e^{+h} + e^{-h}}$$

Practical 1 : conjugated priors

Likelihood $p(y|\theta)$, after Laplace $p(y|\theta) \propto \theta^y(1-\theta)^{n-y}$, where y # positive events,

Conjugated prior: $p_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, which is Beta distribution

Posterior $\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$, we call $y' = y+\alpha-1$ the "positive event", and $n'-y' = n-y+\beta-1$, n' as effective # of data.

$$\theta_{\text{MAP}} = \frac{y'}{n'} = \frac{y+\alpha-1}{n+\alpha+\beta-2} = \frac{y}{n} \times \underbrace{\frac{n}{n+\alpha+\beta-2}}_{\kappa\in(0,1)} + \frac{\alpha-1}{\alpha+\beta-2} \times \underbrace{\left(1-\frac{n}{n+\alpha+\beta-2}\right)}_{1-\kappa}$$

so θ_{MAP} is a value $\in (\theta_{\text{prior}} = \frac{\alpha-1}{\alpha+\beta-2}, \frac{y}{n} = \theta_{\text{MLE}})$. No data refers to θ_{prior} while plenty of data refers to θ_{MLE}. We call θ_{MAP} "shrinkage":

$$\theta_{\text{MAP}} = (1-\kappa)\theta_{\text{prior}} + \kappa\theta_{\text{MLE}}$$

What is exponential family?? Consider D-dim data θ:

$$P(y|\theta) = \exp\left[\Psi(y)^T \cdot \theta - V(\theta)\right]$$

For instance, $y = (y_1, \cdots, y_n = \pm 1)$, $\theta = \left\{h_i, J_{ij}\right\}$ the couplings with dimension $N + N(N-1)/2$, for N single variables and $N(N-1)/2$ products in Ψ.

Suppose you have a Gaussian distribution

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

With fixed variation, what is prior over μ? **It is still a Gaussion.**

$$p(\mu) = \frac{1}{\sqrt{2\pi\eta^2}} \exp\left[-\frac{(\mu-\bar{\mu})^2}{2\eta^2}\right]$$

With fixed moment, what is prior over σ^2? Take $\tau = \sigma^{-2}$, we get **core of Gamma function**

$$p(\tau) \propto \tau^{\alpha-1}e^{-\beta\tau}$$

Centered Multivariate Gaussian (L-dim)

$$p(y_1, \cdots, y_L|\tau) = \frac{1}{(2\pi)^{L/2}\sqrt{\det(\tau^{-1})}} \exp\left(-\frac{1}{2}\sum_{i,j}^{L} y_i \tau_{ij} y_j\right)$$

We have $\tau_{ij}^{-1} = \left\langle y_i y_j \right\rangle_P$ and τ_{ij} a $L \times L$ positive-definitive matrix.

Conjugated prior: **Wishart distribution**

$$p(\tau) \propto (\det\tau)^{(\alpha-1)/2} \times \exp\left(-\frac{1}{2}\sum_{i,j} \tau_{ij} V_{ij}\right)$$

L_p-norm based priors Suppose that you have a model with data $\theta = (\theta_1, \cdots, \theta_n)$

$$p(\theta) \propto \exp\left(-\frac{\gamma}{p!}\sum_{i=1}^{D} |\theta_i|^P\right)$$

$p = 2$, Gaussian; $p = 1$, exponential distribution; $p = 0$, sparsity-enforcing prior.

$D = 1, M = 1$ data point, $y = \theta + z \sim \mathcal{N}(0,1)$, where z is Gaussian noise. "dumb" version of linear regression, consider higher dimension

$$y^m = \sum_i A^{m_i}\theta_i + z^m$$

In the case $D = M = 1$ case, we have likelihood:

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)$$

Prior L_2: $p(\theta) \propto \exp\left(-\frac{1}{2}\gamma\theta^2\right)$

Prior L_1: $p(\theta) \propto \exp\left(-\gamma\,|\theta|\right)$.

We'd like to consider in L_2: $\min_\theta \frac{1}{2}(y-\theta)^2 + \frac{1}{2}\gamma\theta^2$, getting $\theta = y/(1+\gamma)$, $\theta^{\text{MAP}}_{L_2} = \frac{y}{1+\gamma}$, $\theta^{\text{MLE}} = y$

We'd like to consider in L_1: $\min_\theta \frac{1}{2}(y-\theta)^2 + \gamma\,|\theta|$, getting $-(y-\theta) + \gamma \cdot \text{sign}(\theta)$. In the region $y \in (-\gamma, +\gamma)$, we always have $\theta = 0$!

$p \leq 1$: good for sparsity; $p \geq 1$: convexity.

$$P(\theta|Y) = \frac{P(Y|\theta) \times P_\gamma(\theta)}{P_\gamma(Y)}$$

since $P_\gamma(\theta) \sim \exp(-\gamma\,|\theta|)$ with model parameter θ, so $P(Y)$ also depends on γ (hyperparameter). We then see $\int d^D\theta P(Y|\theta)P_\gamma(\theta)$, the max to get γ.

Suppose 80% training set, 20% for test set... $\log P(Y|\theta)$ decrease as γ increase for training, or + then - for test (lower than training), with big gap at $\gamma = 0$.

Universal prior : (Jeffrey's prior)

$\theta \longrightarrow \frac{e^h}{e^h + e^{-h}}$, uniform over θ, getting $p(h) = \frac{1}{(e^{h/2}+e^{-h/2})^2}$. Note $p(\theta) = p(h) \times \left|\frac{\mathrm{d}h}{\mathrm{d}\theta}\right|$... Non-linear transformation, so uniform distribution θ results non-uniform distribution of h.

$\theta_{k+1} - \theta_k \propto \frac{1}{Mp(\theta)}$

$$D_{KL}\left[p(y|\theta_{k+1})||p(y|\theta_k)\right] = 0 + 0 \times \epsilon + \frac{1}{2}\epsilon^2 I(\theta_k) + \cdots \propto \frac{I(\theta_k)}{p^2(\theta_k)}$$

where $\theta_{k+1} = \theta_k + \epsilon$, $I(\theta_k)$ is Fisher information (see "Asymptotic inference"). In the end, we obtain $p_{\text{Jeffrey}}(\theta) \propto \sqrt{I(\theta)}$

8 02/12/22: State space models (Hidden Markov Models)

So far, we have considered inference problems in which time played no role. In many applications, however, data are time series produced by a dynamical process. How can we infer the underlying rules defining this process? We will address this question in two frameworks: a simple one, in which measurements give directly access to the dynamical sequence of the states visited by the system, and a richer one, in which we only have indirect knowledge about the states.

Very simple example: $x = \{s_1 = \pm 1, \cdots, s_N\} \in X(2^N)$, $\Rightarrow$ data = time series $\{X_0, X_1, \cdots, X_T\}$.

Assumption Dynamics is Markovian.

$$P(X_{t+1}|X_t, X_{t-1}, \cdots, X_1, X_0) = P(X_{t+1}|X_t) = \underbrace{\Omega}_{\text{transition.probability}} (X_t \to X_{t+1})$$

Weaker: (finite memeory) $P(X_{t+1}|X_t, X_{t-1}, \cdots, X_1, X_0) = P(X_{t+1}|X_t, X_{t-1}, \cdots, X_{t-L+1})$. Here, we could combine $X_t, X_{t-1} \cdots X_{t-L+2}$ as $\widetilde{X}_t$.

$$P(\text{data}) = P(X_0, \cdots, X_T) = \prod_{t=0}^{T-1} \Omega(X_t \to X_{t+1}) = \prod_{x,y} \Omega(x \to y)^{N(x \to y)} = \exp\left\{ \sum_{x,y} N(x \to y) \log \Omega(x \to y) \right\}$$

where $N(x \to y)$ is s.t. $x = x_t, y = x_{t+1}$.

Consider MLE of Ω: $\Omega_{\text{MLE}}(x \to y) = \frac{N(x \to y)}{\sum_z N(x \to z)}$

$$\max_{\Omega} \left\{ \log P(\text{data}|\Omega) + \sum_x \lambda_x \left[\sum_y \Omega(x \to y) - 1 \right] \right\} \qquad \sum_{x,y} N(x \to y) \log \Omega(x \to y)$$

$$\frac{\partial(.)}{\partial \Omega(x \to y)} = 0 = \frac{N(x \to y)}{\Omega(x \to y)} + \lambda_x \qquad \to \qquad \sum_y N(x \to y) + \lambda_x \sum_y \Omega(x \to y) = 0$$

Assumption Stationary distribution. Wait long enough: $P(x_\tau = x) = \pi(x)$ independent of τ.

$$N(x \to y) \simeq T\pi(x)\Omega(x \to y) + \cdots$$

$$P(x_0, \cdots, x_T) \simeq \exp\left\{ -T \boxed{ \sum_x \pi(x) \left[-\sum_y \Omega(x \to y) \log \Omega(x \to y) \right] } \right\}$$

where T is time duration. The part in the box is called entropy rate.

Hidden Markov Model (HMM)

Consider the case below: we can not see X but Y, with $q(y|x)$ = emission probability, from X_t to Y_t:

$$\begin{tikzcd}
{} \arrow[r] & \mathrm{X}_{t-1} \arrow[r] \arrow[d] & \mathrm{X}_t \arrow[r] \arrow[d] & \mathrm{X}_{t+1} \arrow[r] \arrow[d] & {} \\
{} \arrow[r] & \mathrm{Y}_{t-1} \arrow[r] & \mathrm{Y}_t \arrow[r] & \mathrm{Y}_{t+1} \arrow[r] & {}
\end{tikzcd}$$

$$\vec{X}_{t+1} \in \mathbb{R}^D = A\vec{X}_t + \vec{W}_t \qquad\qquad \vec{y}_t \in \mathbb{R}^{D'} = B\vec{X}_t + \vec{\varepsilon}_t$$

See: Kalmer filter article on the website. (Useless link... ...)

HMM N states $x^{(i)}, i = 1, \cdots, N$; M symbols $y^{(j)}, j = 1, \cdots, M$; $\Omega_{N\times N}$ transition matrix; $q_{M\times N}$ emission matrix. Initial state x_0.

Question-1: Given a set of observations $Y = (y_1, \cdots, y_T)$ and knowledge of Ω, q, x_0. What is the probability of Y, ie. $P(Y)$?

Question-2: Given Y and Ω, q, x_0. What are the most likely states $X = (x_1, \cdots, x_N)$.

Question-3: Given Y, what are Ω, q?

Question-1

$$P(Y,X) = \underbrace{\prod_{t=0}^{T-1} \Omega(x_t \to x_{t+1})}_{\mathrm{P(X)}} \prod_{t=1}^{T} q(y_t|x_t) \qquad M_{t+1}(x_{t+1}, x_t) = \Omega(x_t \to x_{t+1}) \cdot q(y_{t+1}|x_{t+1})$$

$$\begin{aligned}
P(Y) = \sum_X P(Y,X) &= \sum_{X_1,\cdots,X_T} M_1(x_1,x_0)M_2(x_2,x_1)\cdots M_T(x_T,x_{T-1}) \\
&= \sum_{X_1,\cdots,X_T} M_T(x_T,x_{T-1})\cdots M_2(x_2,x_1)M_1(x_1,x_0) = \sum_{x_T} \left(M_T \times \cdots \times M_2 \times M_1\right)(x_T, x_0)
\end{aligned}$$

Here we exploit matrix multiplication; and the property $M_2(x_2,x_1) \times M_1(x_1,x_0) = (M_2 \times M_1)(x_2,x_0)$, where 1st x_1 is column in M_2 and 2nd x_1 is rank in M_1.

Question-2

$$P(X|Y) = \frac{\prod_{t=0}^{T-1} \Omega(x_t \to x_{t+1}) \prod_{t=1}^{T} q(t_t|x_t)}{P(Y)} = \frac{M_T(x_T,x_{T-1})\cdots M_2(x_2,x_1)M_1(x_1,x_0)}{P(Y)}$$

How do we maximize over X in $P(X|Y)$? Two pass algorithm $X_0 \leftrightarrow X_T$.

Optimize over x_1: (N possible values of x_2)

$$x_1^*(x_2) = \text{argmax}\left[\underbrace{M_2(x_2,x_1)\cdot M_1(x_1,x_0)}_{\mathrm{P}_2^*(\mathrm{x}_2)}\right]$$

Optimize over x_2:

$$x_2^*(x_3) = \text{argmax}\left[\underbrace{M_3(x_3, x_2) \cdot M_2(x_2, x_1^*(x_2)) \cdot M_1(x_1^*(x_2), x_0)}_{\mathrm{P_3^*(x_3)}}\right]$$

We repeat the process, getting $p_T^*(x_T)$. Maximize it as x_T^* and we return to x_{T-1}^* by the relation we just found, until x_0, namely $X_0 \leftrightarrow X_T$.

Exercise: $P(x_{t=18}) = ?$, $P(x_{t=18}, x_{t=25}) = ?$

Question-3

We can compute "efficiently" $P_\theta(Y)$.

Procedure Expectation Maximization (EM)

Idea suppose we have estimate θ_e, then we find θ_b s.t $P_{\theta_b}(Y) \geq P_{\theta_e}(Y)$. I know how to sample $P_{\theta_e}(X|Y)$. I define

$$G(\theta) = \sum_X P_{\theta_e}(X|Y) \log \underbrace{P_\theta(X, Y)}_{\text{joint}}$$

Claim $\theta_b = \text{argmax} G(\theta)$ is a better estimate of θ than θ_e.

$$P_\theta(X, Y) = \prod_t \Omega(x_t \to x_{t+1}) \prod_t q(y_t|x_t)$$

Proof We know $P_\theta(Y) = P_\theta(X, Y)/P_\theta(X|Y)$, and then

$$\log P_\theta(Y) = \log P_\theta(X, Y) - \log P_\theta(X|Y)$$

$$\sum_X P_{\theta_e}(X|Y) \quad \to \quad \underbrace{\cancel{\sum_X P_{\theta_e}(X|Y)}}_{=1} \log P_\theta(Y) = \underbrace{\sum_X P_{\theta_e}(X|Y) \log P_\theta(X, Y)}_{\mathrm{G}(\theta)} - \sum_X P_{\theta_e}(X|Y) \log P_\theta(X|Y)$$

$$\forall \theta, \qquad \log P_\theta(Y) = G(\theta) - \sum_X P_{\theta_e}(X|Y) \log P_\theta(X|Y)$$

$$\theta_e : \qquad \log P_{\theta_e}(Y) = G(\theta_e) - \sum_X P_{\theta_e}(X|Y) \log P_{\theta_e}(X|Y)$$

$$\log P_\theta(Y) - \log P_{\theta_e}(Y) = G(\theta) - G(\theta_e) + \underbrace{\sum_X P_{\theta_e}(X|Y) \log \left[\frac{P_{\theta_e}(X|Y)}{P_\theta(X|Y)}\right]}_{\mathrm{D_{KL}} \geq 0}$$

$$\log P_\theta(Y) - \log P_{\theta_e}(Y) \geq G(\theta) - G(\theta_e)$$

How to maimize $G(\theta)$ over $\theta = (\Omega, q)$?

$$P_{\theta_e}(X|Y) = \frac{\boxed{\prod_t \Omega_e(x_t \to x_{t+1}) \prod_t q_e(y_t|x_t)}}{\sum_{X'} \prod_t \Omega_e(X'_t \to X'_{t+1}) \prod_t q_e(y_t|X'_t)}$$

$$G(\theta) = \sum_X P_{\theta_e}(X|Y) \log P_\theta(X,Y) = \sum_X P_{\theta_e}(X|Y) \left\{ \sum_t \log \Omega(X_t \to X_{t+1}) + \sum_t \log q(y_t|X_t) \right\}$$

max under constraints (i) $\sum_y \Omega(x \to y) = 1, \forall x$; (ii) $\sum_y q(y|x) = 1, \forall x$. We need to max:

$$\max \left\{ G(\Omega, q) - \sum_X \mu_X \times \boxed{(i)} - \sum_X \eta_X \boxed{(ii)} \right\}$$

$$\frac{\partial G}{\partial \Omega(x \to x')} = 0 = \sum_X P_{\theta_e}(X|Y) \left\{ \sum_t \frac{\delta_{x,x_t} \delta_{x',x_{t+1}}}{\Omega(x \to x')} \right\} - \mu_X = \frac{\langle N(x_t = x \to x_{t+1} = x') \rangle_{\theta_e}}{\Omega(x \to x')} - \mu_X$$

$$\Omega(x \to x') = \frac{\langle N(x_t = x \to x_{t+1} = x') \rangle_{\theta_e}}{\langle N(x_t = x) \rangle_{\theta_e}}$$

Similarly, for transition matrix, we change variables right of →:

$$q(y|x) = \frac{\langle N(x_t = x \to y_t = y) \rangle_{\theta_e}}{\langle N(x_t = x) \rangle_{\theta_e}}$$

9 09/12/22: Probabilistic graphical models & Boltzmann machine

Understanding how the many elementary components of a system, be they neurons, genes, species, etc., interact to produce a global behavior is the central scope of data-driven modeling. A simple way to characterize these interactions is to look at the correlations between pairs of components. Yet, pairwise correlations are potentially misleading, as they can reflect indirect effects mediated via third-body components, rather than direct interactions. A more sensible estimate of interactions is provided by the graph of dependencies we have introduced in the context of multivariate Gaussian distribution. Briefly speaking, the idea is to look at the conditional probability of a component, or variable y_i, given the other y_j. The question we will ask below is: how can this graph, or network, be reconstructed from a set of observations of the variables?

(Two lectures today... BON COURAGE!!!)

From Lecture 3, we see PCA. Consider interaction graph inference, position at i reads x_i, interaction shown as $L \times L$ matrix T_{ij} between sites i and j. Probability:

$$P(x) \propto \exp\left(-\frac{1}{2} \sum_{ij} x_i T_{ij} x_j \right)$$

Categorical variables → binary variables

Independent-site model Configuration of data reads $\sigma = (\sigma_1, \cdots, \sigma_N)$, and each σ_i has q values. Probability: $P(\sigma) = \prod_{i=1}^{N} p_i(\sigma_i)$. For each site i, its probability reads:

$$p_i(\sigma_i) = \frac{\exp\left[h_i(\sigma_i) + X\right]}{\sum_i \exp\left[h_i(\sigma_i) + X\right]}$$

with h_i the position weight matrix in transformation, and X can be any constant not influencing probability. Thereafter $q = 2$, we consider the following transformation:

$$\sigma_i = 1 \quad \text{or} \quad 2 \qquad \to \qquad \sigma_i = 0, 1$$

$$h_i(1) \quad \text{and} \quad h_i(2) \qquad \to \qquad h_i = h_i(2) - h_i(1)$$

$$p_i(\sigma_i) = \frac{e^{h_i \sigma_i}}{1 + e^{h_i}}$$

<u>Inference of h_i through MLE</u>

For data $\left\{\sigma_i^{a=1,\cdots,M}\right\}$, its log-likelihood reads

$$\sum_a \log p_i(\sigma_i^a) = h_i \sum_a \sigma_i^a - M \log\left(1 + e^{h_i}\right) = M\left\{h_i \langle\sigma_i\rangle - \log\left(1 + e^{h_i}\right)\right\}$$

$$\frac{\partial}{\partial h_i}(\bullet) = 0 \qquad \to \qquad \langle\sigma_i\rangle = \frac{e^{h_i \sigma_i}}{1 + e^{h_i}} = \sum_{\sigma_i = 0,1} \sigma_i p_i(\sigma_i)$$

called **Moment-matching condition**.

$$-LL = -\sum_a \log p_i^{\text{Model}}(\sigma_i^a) \qquad p_i^{\text{Data}}(\sigma_i) = \frac{1}{M}\sum_a \delta_{\sigma_i, \sigma_i^a} \qquad LL = -\sum_{\sigma_i = 0,1} p_i^{\text{Data}}(\sigma_i) \log p_i^{\text{Model}}(\sigma_i)$$

Here, LL = cross-entropy between $p^{\text{Data}}, p^{\text{Model}}$.

Connection with MaxEnt We maintain first moment conserved

$$\max_{\{p(\sigma)\}} \left\{-\sum_\sigma p(\sigma)\log p(\sigma) + \lambda\left[\sum_\sigma p(\sigma) - 1\right] + \mu\left[\sum_\sigma p(\sigma)\sigma - \langle\sigma\rangle^{\text{data}}\right]\right\}$$

$$\frac{\partial}{\partial p(\sigma)}(\bullet) = 0 = -\log p(\sigma) - 1 + \lambda + \mu\sigma \qquad \to \qquad p(\sigma) = e^{\lambda - 1} e^{\mu\sigma} \qquad \frac{\partial}{\partial p(\sigma)}\frac{\partial}{\partial p(\tau)}(\bullet) = -\frac{\delta_{\sigma,\tau}}{p(\sigma)}$$

Therefore, we maximize Shanon entropy (≤ 0) = we minimize cross entropy with $p^{\text{data}} \geq 0$. Both give the same result in model space, where model fitting data ($\langle\sigma\rangle^{\text{Data}}$).

Coupled-site model Only consider 2-body interactions

$$p(\sigma) = \prod_i p(\sigma_i) \prod_{i<j} p(\sigma_i, \sigma_j) = \frac{1}{Z} \exp\left(\sum_i h_i \sigma_i + \sum_{i<j} J_{ij} \sigma_i \sigma_j \right)$$

For instance, there would be 20 × 20 matrix for amino acides in exponential. But here we only consider binary parameters, hence we write $J_{ij}\sigma_i\sigma_j$. Cross-entropy reads:

$$\mathrm{Cross-entropy} = -M \left\{ \sum_i h_i \left\langle \sigma_i \right\rangle^{\mathrm{Data}} + \sum_{i<j} J_{ij} \left\langle \sigma_i \sigma_j \right\rangle^{\mathrm{Data}} - \log Z \left[\left\{ h_i, J_{ij} \right\} \right] \right\}$$

Minimize over $\left\{h_i, J_{ij}\right\}$

$$\frac{\partial \mathrm{Cross-entropy}}{\partial h_i} = 0 = \left\langle \sigma_i \right\rangle^{\mathrm{Data}} - \underbrace{\frac{1}{Z}\frac{\partial Z}{\partial h_i}}_{\langle\sigma_{\mathrm{i}}\rangle^{\mathrm{Model}}, \forall \mathrm{i}}$$

$$\frac{\partial \mathrm{Cross-entropy}}{\partial J_{ij}} = 0 = \left\langle \sigma_i \sigma_j \right\rangle^{\mathrm{Data}} - \underbrace{\frac{1}{Z}\frac{\partial Z}{\partial J_{ij}}}_{\langle\sigma_{\mathrm{i}}\sigma_{\mathrm{j}}\rangle^{\mathrm{Model}}, \forall \mathrm{i<j}}$$

Consider X in $N(N+1)/2$ dimension, N for σ_i, $N(N-1)/2$ for $\sigma_i\sigma_j$; similarly for $\theta = (h_i, J_{ij})^T$

$$p^{\mathrm{Model}}(X) = \frac{1}{Z(\theta)} \exp \sum_\alpha \theta_\alpha X_\alpha$$

$$\frac{\partial}{\partial \theta_\alpha} \frac{\partial}{\partial \theta_\beta} \mathrm{Cross-entropy} = \left\langle X_\alpha X_\beta \right\rangle^{\mathrm{Model}} - \left\langle X_\alpha \right\rangle^{\mathrm{Model}} \left\langle X_\beta \right\rangle^{\mathrm{Model}} \geq 0$$

Consider L_2 prior: $\exp\left(-\frac{\gamma}{2}\sum_\alpha \theta_\alpha^2\right)$. Additive contribution to cross-Ent = $+\frac{\gamma}{2}\sum_\alpha \theta_\alpha^2$. Additive contribution to Hessian matrix = $+\gamma\delta_{\alpha\beta}$, see second derivative.

Boltzmann Machine Learning (1986)

Idea:

- Start from guesses for h_i^0, J_{ij}^0.
- Compute $\langle\sigma_i\rangle^{\mathrm{Model}}$, $\left\langle\sigma_i\sigma_j\right\rangle^{\mathrm{Model}}$, with Monte Carlo.
- Update parameters with small parameter η called **Learning rate**:

$$h_i^{t+1} = h_i^t - \eta\left(\langle\sigma_i\rangle^{\mathrm{Model}} - \langle\sigma_i\rangle^{\mathrm{Data}}\right) \qquad J_i^{t+1} = J_i^t - \eta\left(\left\langle\sigma_i\sigma_j\right\rangle^{\mathrm{Model}} - \left\langle\sigma_i\sigma_j\right\rangle^{\mathrm{Data}}\right)$$

$$\theta^{t+1} = \theta^t - \eta\frac{\partial \mathrm{Cross-Ent}}{\partial\theta}$$

Perhaps we jump far away from the local minimum we want due to large first-order derivative, thus we can consider second derivative and decrease θ while differential.

Approximate inference: Mean Field Callen identities: with $H_i = \sum_{j\neq i} J_{ij}\sigma_j + h_i$

$$\langle\sigma_i\rangle = \left\langle \frac{\exp H_i(\{\sigma_j\})}{1+\exp H_i(\{\sigma_j\})}\right\rangle^{\text{Model}}_{\{\sigma_{j\neq i}\}}$$

Mean Field:

$$\langle\sigma_i\rangle^{\text{MF}} \simeq \frac{e^{\langle H_i\rangle}}{1+e^{\langle H_i\rangle}} = \frac{e^{\sum_j J_{ij}\langle\sigma_j\rangle^{\text{MF}}+h_i}}{1+e^{\sum_j J_{ij}\langle\sigma_j\rangle^{\text{MF}}+h_i}}$$

there are N unknown equations, for fixed J_{ij}, h_i.

Consider response matrix $R_{ij} = \frac{\partial\langle\sigma_i\rangle^{\text{MF}}}{\partial h_j}$.

- Fluctuation-dissipation theorem:

$$\langle\sigma_i\rangle = \frac{\partial}{\partial h_i}\log Z = \frac{1}{Z}\frac{\partial Z}{\partial h_i}$$

$$R_{ij} = \frac{\partial\left\langle\sigma_i\right\rangle^{\text{MF}}}{\partial h_j} = \frac{\partial}{\partial h_j}\left(\frac{1}{Z}\frac{\partial Z}{\partial h_i}\right) = \frac{1}{Z}\frac{\partial^2 Z}{\partial h_i\partial h_j} - \frac{1}{Z^2}\frac{\partial Z}{\partial h_i}\frac{\partial Z}{\partial h_j} = \left\langle\sigma_i\sigma_j\right\rangle - \left\langle\sigma_i\right\rangle\left\langle\sigma_j\right\rangle = C_{ij}$$

- Using MF

$$R_{ij} = \frac{\partial\left\langle\sigma_i\right\rangle^{\text{MF}}}{\partial h_j} = \frac{\partial}{\partial h_j}F\left(\sum_k J_{ik}\left\langle\sigma_k\right\rangle^{\text{MF}} + h_i\right) \qquad F(u) = \frac{e^u}{1+e^u}$$

$$R_{ij} = F'\left(\left\langle H_i\right\rangle^{\text{MF}}\right)\cdot\left(\sum_k J_{ik}\times R_{kj} + \delta_{ij}\right) \qquad R = D\left(JR+\mathbb{I}\right) = C \qquad D^{-1} = J + R^{-1}$$

Here D is a diagonal matrix, expressed by $F'\left(\left\langle H_i\right\rangle^{\text{MF}}\right)$. If $i\neq j$, $D_{ij}=0$ leads to

$$J_{ij} = -(C^{-1})_{ij}$$

Pseudo-likelihood approximation Data $\sigma^{a=1,\cdots,M}_{i=1,\cdots,N} \to \{h_i, J_{ij}\}$. $P(\sigma_i|\{\sigma_{j\neq i}\})$ depends on all J_{ij}. MLE on one row of J_{ij}:

$$LL = \sum_{a=1}^{M}\log P(\sigma_i^a|\{\sigma_{j\neq i}^a\};h_i;\{J_{ij}\}) \qquad P(\sigma_i^a|\{\sigma_{j\neq i}^a\};h_i;\{J_{ij}\}) \propto \frac{e^{\sigma_i^a\left(\sum_j J_{ij}\sigma_j^a+h_i\right)}}{1+e^{\sum_j J_{ij}\sigma_j^a+h_i}}$$

max over H_i, over J_{ij} for $j\neq i$.

$$LL = M\left\{h_i\left\langle\sigma_i\right\rangle^{\text{Data}} + \sum_j J_{ij}\left\langle\sigma_i\sigma_j\right\rangle^{\text{Data}} - \left\langle\log\left(1+e^{\sum_j J_{ij}\sigma_j^a+h_i}\right)\right\rangle^{\text{Data}}\right\}$$

$$\frac{\partial LL}{\partial h_i} = 0 = \left\langle\sigma_i\right\rangle^{\text{Data}} - \left\langle\frac{e^{\sum_k J_{ik}\sigma_k^a+h_i}}{1+e^{\sum_k J_{ik}\sigma_k^a+h_i}}\right\rangle^{\text{Data}} = \left\langle\sigma_i\sigma_j\right\rangle^{\text{Data}} - \left\langle\sigma_j\frac{e^{\sum_k J_{ik}\sigma_k^a+h_i}}{1+e^{\sum_k J_{ik}\sigma_k^a+h_i}}\right\rangle^{\text{Data}}$$

To prove that only few J_{ij} are non-zero, we use L_1 norm and substrate $\gamma\sum_j\left|J_{ij}\right|$ in LL.

Second Lecture

Exam on Friday, Jan 6th. **NOT** at SU, bad network; but ENS 3rd floor (L357 or L367), from 2 to 5pm. There would be one theoretical question like those from lectures, and one practical question like TD. Note, lecture notes & TDs are allowed, as well as personal laptop. :-)

...

Consider the following model with $\sigma_i, i = 1, \cdots, N$, and $h_\mu, \mu = 1, \cdots, M \leq N$.

$$P(\sigma) = \frac{1}{2} \exp\left(\sum_i g_i \sigma_i + \frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j \right)$$

$$J_{ij} = \sum_{\mu=1}^{N} \lambda_\mu v_{i,\mu} v_{j,\mu}$$

v_μ is normalized eigenvector associated to λ_μ. And h_μ is hidden variables.

$$P(\sigma) = \frac{1}{2} e^{\sum_i g_i \sigma_i} \prod_\mu \exp\left[\frac{1}{2} \lambda_\mu \left(\sum_i v_{i\mu} \sigma_i \right)^2 \right] \qquad \rightarrow \qquad \int_{-\infty}^{+\infty} \frac{dh_\mu}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} h_\mu^2 + h_\mu \sqrt{\lambda_\mu} \sum_i v_{i\mu} \sigma_i \right)$$

This is called **Restricted Boltzmann machine**.

Define joint probability: $P(\sigma, h)$, and $P(\sigma) = \int dh p(\sigma, h)$ marginal for cond. probability.

$$P(h|\sigma) \propto \prod_\mu \exp\left[-\frac{1}{2} h_\mu^2 + h_\mu \underbrace{\left(\sqrt{\lambda_\mu} \sum_i v_{i\mu} \sigma_i \right)}_{\mathbb{I}_\mu(\sigma)} \right] \qquad P(\sigma|h) \propto \prod_i \exp\left[\sigma_i \left(g_i + \underbrace{\sum_\mu \sqrt{\lambda_\mu} v_{i\mu} h_\mu}_{\mathbb{I}_\mathrm{i}(\mathrm{h})} \right) \right]$$

This is much stronger than PCA, since we can change h_μ. We do projection of data towards $\mathbb{I}$.

Consider a machine with input $x \in \mathbb{R}^N$ and output $x \in \mathbb{R}^N$, we insert one layer $y \in \mathbb{R}^K$ as $K \ll N$. We write $y_\ell = f(\sum_i W_{\ell i} x_i)$ as *Encoder*, $\theta_j = g(\sum_\ell v_{j\ell} y_\ell) \simeq x_j$ as *Decoder*. **Problem**: find weights W, v such that on a data distribution:

$$\frac{1}{M} \sum_{a=1}^{M} \left\{ \sum_j \left[x_j^a - g\left(\sum_\ell v_{j\ell} f(\sum_i v_{\ell i} x_i^a) \right) \right]^2 \right\}$$

is minimum. Namely we have to find several planes in $\mathbb{R}^K$ with largest variances of projected data from $\mathbb{R}^N$.

If $f(u) = u, g(u) = u$, so $X \simeq V_{N\times K} \cdot W_{K\times N} \cdots X$. Consider the data with noise, we could only keep eigenvalues larger than noise amplitude. $\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \rightarrow$ eigenvalues :$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$.

Consider one machine with hidden units in dim $K = N$, **but** many $y_\ell = 0$.

......

Below are course based on Slides

Images:

12000 Images = 6000 natural + 6000 man-made, 256*256 pixels.

Principal component analysis:

compute correlation matrix C(r,r'), where r is 2D vector

Diagonalize and find top components: 8 patterns thought

Similar to 2D Fourier modes (plane waves)

Comes from statistical properties of images. Due to statistical properties of images: translation invariance approximate rotation invariance.

Sparse auto-encoder:

See Ng, 2001. 100(+1) input units, 100 output units, Images=10*10 pixels. 100 hidden units

$$\sigma_\mu = \frac{1}{1 + \exp\left(-\sum_i W_{\mu i} x_i\right)}$$

Sparse auto-encoder:

trained on 100 natural images, we get 10*10 patterns with bright-dark patterns by $\frac{w_{\mu i}}{\sum_i w_{\mu i}^2}$

called **Sparse-dictionary learning**

Sparse-dictionary learning:

180 basis functions, 12*12 pixel images, 10000 natural images $S(x) = \log\left(1 + x^2\right)$

Some images from Olshausen , Field 1996.

Receptive Fields in Macaque V1:

Ringach, 2002. Zylberberg, Murphy, DeWeese 2011

Restricted Boltzmann Machines:

Graphical model constituted by two sets of random variables that are coupled together:

$$P(v,h) = \frac{1}{Z}\exp[-E(v,h)] \qquad E(v,h) = -\sum_i g_i v_i + \sum_\mu U_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu$$

... Smolensky 1986

Not necessarily we put $U(h) \sim h^2$, but we can add $\gamma\,|h|$ for instance. Thus we modifchange behavior near the origin.

MNIST: Unsupervised learning of synthetic digits:

60000 images of digits with 28*28 pixels.

Easy for $\langle \sigma_i \rangle$, but no correlation? Hard, we consider $\left\langle \sigma_i^{\text{Data}} h_\mu \right\rangle$.

Weights:

Bernoulli RBMs, Trained on MNIST

Fischer & Igel. Training Restricted Boltzmann Machine: An Introduction, 2014.

Learning continuous invariances:

- Features reflect the data distribution in a non-trivial way
- What happens for data distribution with continuous invariances?

- Do not want to hardwire the symmetry eg convolutional architecture
- Show very simple example
- Problem solved by the brain ...

Learning with RBM: dynamical symmetry restoration:

Ising 1D ring, many config.s toward 1 hidden units (most stupid machine). It could reflects correlation length! W looks like a peak, but the peak position would move as learning time increases.

More hidden layers for the same system? They would move together, with different peak position but same frequency

Representation of space in the brain:

The Nobel Prize in Physiology or Medicine 2014

John O'Keefe & "place cells" in Hippocampus