26 Lecture 26, Dec 2

Announcements

- HW8 (ranking MLB teams) returned. Feedback:
  - Bradley-Terry model. KL 12.6 or David Hunter's paper.
  - Non-concavity of the log-likelihood in $\boldsymbol{\lambda}$ parameterization. Enough to find a simple counter-example. However the negative log-likelihood is an example of geometric program, a branch of convex programming.
  - Vectorization of MM update: `outer` function.
  - Concavity of the log-likelihood function in $\boldsymbol{\gamma}$ parameterization. "log-sum-exp" terms are convex.
  - Implementation of Newton's method. Hessian is singular due to identifiability. Setting γ_1.
  - Problem structure: sparsity in large league.
  - Check David Hunter's code for taking advantage of sparsity.

  Sketch of solution: `http://hua-zhou.github.io/teaching/st758-2014fall/hw07sol.html`
- HW9 (simulation project) due Dec 9 @ 11A.
- FAQs on HW9 (simulation project)
- Regular office hours this week: Tue (Hua), Thu (Hua) and Fri (William). No office hours next week.
- **Course evaluation**!: `https://classeval.ncsu.edu/`

Last time

- Pre-conditioning for conjugate gradient (PCG) method.
- Nonlinear conjugate gradient for optimization.
- Concluding remarks on optimization.

Today

- Introduction to Markov chains and MCMC.
- Fast algorithms: sorting, FFT.

Introduction to MCMC

Some topics I'll briefly talk about.

- History of Markov chain
- History of Monte Carlo and birth of MCMC
- Convergence rate of Markov chain

Markov chains

- Markov chain is a stochastic process $X_0, X_1, X_2, \ldots$ with the *Markov property*

$$\mathbf{P}(X_{t+1}|X_t, X_{t-1}, \ldots, X_0) = \mathbf{P}(X_{t+1}|X_t).$$

  Given current state, the future is independent of the past.

- Stochastic analog of ordinary differential equations.

$$\frac{dx(t)}{dt} = F(x(t))$$
$$\mathcal{L}(X_{t+1}|X_t = x) = K(x, \cdot)$$

- Notations (for discrete time, finite Markov chains)
  - a finite state space $\mathcal{X}$
  - transition matrix $K(x, y)$, $x, y \in \mathcal{X}$
  - stationary distribution π on $\mathcal{X}$, defined as a probability vector that satisfies

$$\pi^T K = \pi^T.$$

    Existence of such π is guaranteed by the Perron-Frobenius theorem.
  - $K^l(x, y)$ denotes the l-step transition probabilities

- Markov chains in early years:
  - Fermat and Pascal (circa 1654): Gambler's ruin.
  - Bernoulli (1769) and Laplace (1812): Urn model.
  - I. J. Bienaymé (1845), and later Sir Francis Galton and Watson (*Educational Times*, 1873): Branching process.
  - Paul and Tatiana Ehrenfest (1906): Statistical physics.
  - Poincaré (1912): Card shuffling. *Calcul des Probabilités*

  - Markov's contribution
    - Markov, A. A. (1906)
      Extension of the law of large numbers to dependent quantities [in Russian], *Izv. Fiz.-Matem. Obsch. Kazan Univ. (2nd Ser.)*
    - St. Petersburg School vs Moscow School.
    - Example: 20,000 letters in Pushkin's *Eugene Onegin*.

$$\begin{array}{c} \\ \text{vowel} \\ \text{consonant} \end{array} \begin{array}{c} \begin{array}{cc} \text{vowel} & \text{consonant} \end{array} \\ \begin{pmatrix} 0.128 & 0.872 \\ 0.663 & 0.337 \end{pmatrix} \end{array} \quad \begin{array}{c} \pi \\ \begin{pmatrix} 0.432 \\ 0.568 \end{pmatrix} \end{array}$$

    - See Seneta's interesting account on the history (Seneta, 1996)
  - Some other leading pioneers: Kolmogorov, Fréchet, and Doeblin.
- Why are Markov chains important in statistics?

  - Modeling tool.
    E.g., Marc Coram's "jail message" example: Markov model for letter sequence; PageRank's (imaginary) random web surfer; ...
  - A methodology that revolutionized statistics: Markov Chain Monte Carlo (MCMC).
    * Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953)
      Equation of state calculation by fast computing machines, *The Journal of Chemical Physics*, 21, 1087–1092.
    * Want to sample from a distribution $\pi(x) \propto f(x)$.
      Metropolis algorithm constructs a Markov chain that converges to π.
    * Marc Coram's "jail message" example.

Markov chain Monte Carlo (MCMC)

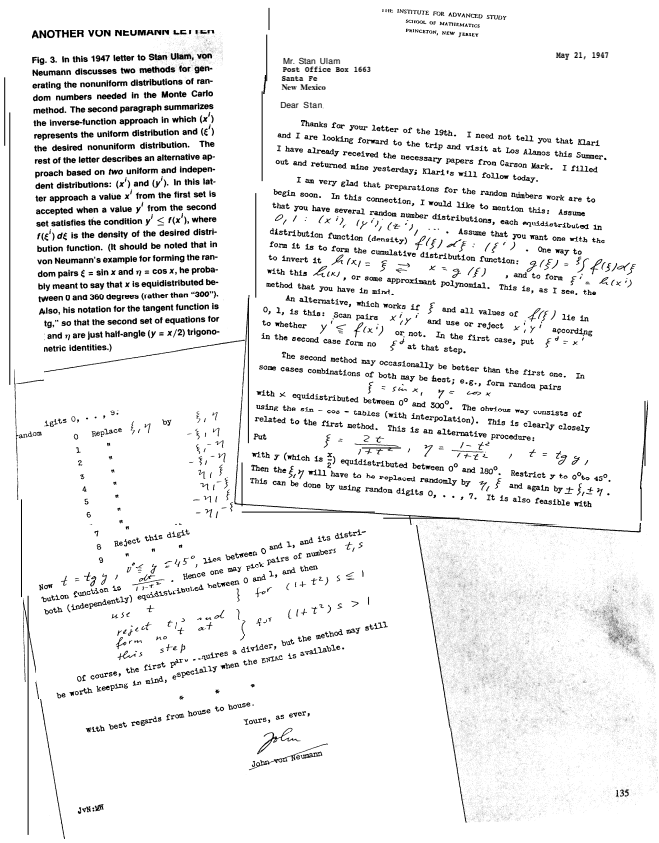
- *Monte Carlo method* is a generic name for "computational algorithms that rely on repeated random sampling to obtain numerical results". They are in contrast to the deterministic algorithms. They have wide applications in
  - integration
  - drawing sample from a distribution. E.g., HW9 (simulation study).
  - optimization (simulated annealing). E.g., Marc Coram's "jail message" example, traveling salesman, Soduku, ...
- Monte Carlo in early years

Von Neumann in the 1940s

Stanislaw Ulam

– Stanislaw Ulam conceived it in 1946 while playing solitaire in hospital bed. He wanted to know the probability of getting a perfect solitaire hand, and wondered whether computers can help answer this.

– John von Neumann was intrigued by the idea and developed a way to generate pseudorandom numbers (inversion, importance sampling, acceptance-rejection sampling) on electronic digital computers (ENIAC) to realize Ulam's idea, for the neutron fission and diffusion problem.



ANOTHER VON NEUMANN LETTER

Fig. 3. In this 1947 letter to Stan Ulam, von Neumann discusses two methods for generating the nonuniform distributions of random numbers needed in the Monte Carlo method. The second paragraph summarizes the inverse-function approach in which (x') represents the uniform distribution and (ξ') the desired nonuniform distribution. The rest of the letter describes an alternative approach based on *two* uniform and independent distributions: (x') and (y'). In this latter approach a value x' from the first set is accepted when a value y' from the second set satisfies the condition $y' \leq f(x')$, where $f(\xi')\,d\xi$ is the density of the desired distribution function. (It should be noted that in von Neumann's example for forming the random pairs $\xi = \sin x$ and $\eta = \cos x$, he probably meant to say that x is equidistributed between 0 and 360 degrees (rather than "300"). Also, his notation for the tangent function is tg," so that the second set of equations for ξ and η are just half-angle ($y = x/2$) trigonometric identities.)

THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: (x'), (y'), (z'), ... Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: (ξ'). One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(s)\,ds$ to invert it $h(x) = \xi \Rightarrow x = g(\xi)$, and to form $\xi' = h(x')$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in 0, 1, is this: Scan pairs x', y' and use or reject x', y' according to whether $y' \leq f(x')$ or not. In the first case, put $\xi' = x'$ in the second case form no ξ' at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs

$$\xi = \sin x, \quad \eta = \cos x$$

with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} \frac{y}{2}$,

with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0°$\leq$ 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits 0, . . , 7. It is also feasible with

digits 0, . . , 9.
random
0 Replace ξ, η by ξ, η
1 " $-\xi, \eta$
2 " $\xi, -\eta$
3 " $-\xi, -\eta$
4 " η, ξ
5 " $\eta, -\xi$
6 " $-\eta, \xi$
7 " $-\eta, -\xi$
8 Reject this digit
9 " " "

Now $t = \operatorname{tg} \frac{y}{2}$, $0° \leq y \leq 45°$, lies between 0 and 1, and its distribution function is $\frac{2t}{1+t^2}$. Hence one may pick pairs of numbers t, s both (independently) equidistributed between 0 and 1, and then

use t for $(1+t^2)\,s \leq 1$
reject t, s and form no ξ, η at this step for $(1+t^2)\,s > 1$

Of course, the first pair requires a divider, but the method may still be worth keeping in mind, especially when the ENIAC is available.

* * *

with best regards from house to house.

Yours, as ever,

John von Neumann

JvN:mW

135

– Ulam and von Neumann, working on Manhattan Project, used the code name *Monte Carlo*. It is a casino in Monaco where Ulam's uncle frequented.

– Nicholas Metropolis, fascinated by the Monte Carlo idea too, designed and built computing devices (MANIAC) to handle such calculations. His paper with Ulam in JASA (Metropolis and Ulam, 1949) formed the basis of modern sequential Monte Carlo methods.

– Birth of MCMC (Metropolis et al., 1953):

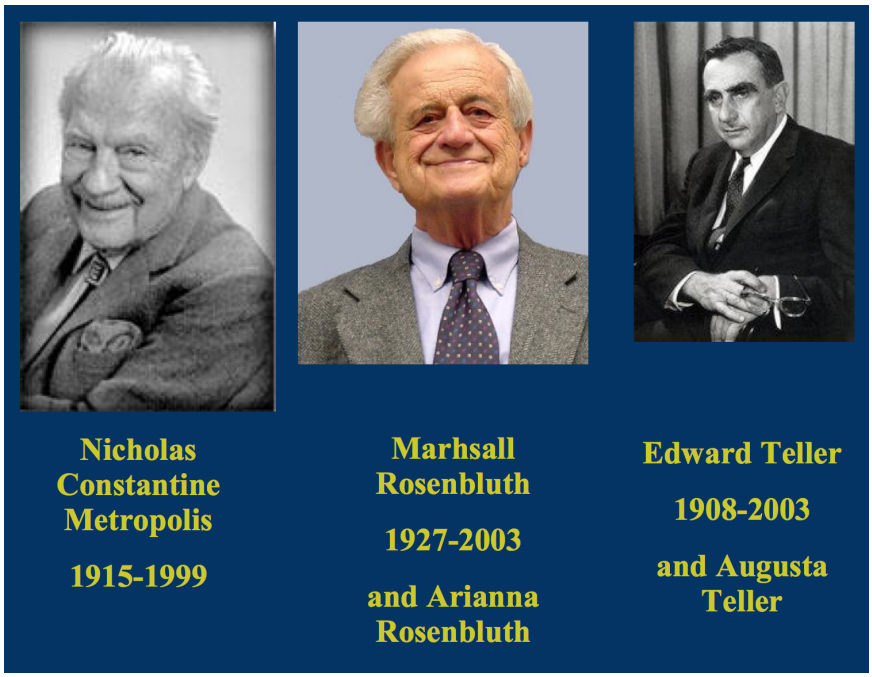
  ∗ Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) Equation of state calculation by fast computing machines, *The Journal of Chemical Physics*, 21, 1087–1092.

  ∗ Want to sample from a distribution $\pi(x) \propto f(x)$. Metropolis algorithm constructs a Markov chain that converges to π.

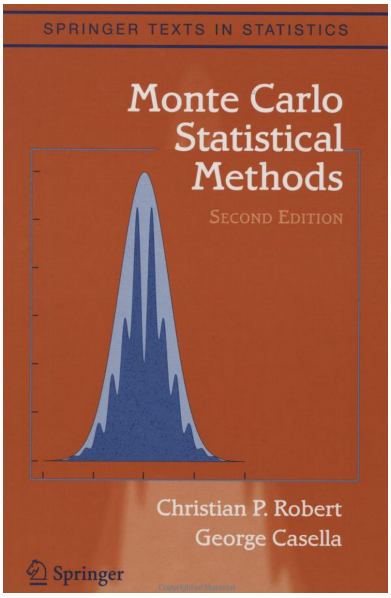
  ∗ *Metropolis chain*: From current state x, generate a new state x' (from a

proposal distribution $p(x, x')$ such that $p(x, x') = p(x', x)$) and accept x' with probability $\min\left\{\frac{f(x')}{f(x)}, 1\right\}$.
Fact: Metropolis chain has π as stationary distribution.

* Marc Coram's "jail message" example.



- Given π, generic ways to construct a Markov chain K that has π as stationary distribution:
  - Metropolis algorithm: (Metropolis et al., 1953)
  - Hastings algorithm: (Hastings, 1970)
  - Gibbs sampler: Glauber dynamics (1963), Tanner and Wong (1987), Gelfand and Smith (1990)
- See KL Ch25-27 and JM Ch13 for a general introduction, or take a Bayesian course. A comprehensive textbook is (Robert and Casella, 2004).



Convergence rate of Markov chains

- Classical result: For finite, irreducible, and aperiodic Markov chains

$$\lim_{l\to\infty} K^l(x,y) = \pi(y).$$

  In practice, we often want to know *how many* steps to make the difference between $K^l(x,\cdot)$ and π small?

- Example: "How many shuffles do I need to do to mix a deck of 52 cards?"

Consider *riffle shuffle*. GilbertShannonReeds model: bionomial(52, 0.5) cut + cards drop according to probability $L/(L+R)$, where L and R and the number of cards in the left and right hand respectively.

- Example: "How long do I need to run my Gibbs sampler?"

  Consider Beta-Binomial Gibbs sampler

  - Likelihood $f(x \mid p) \sim \text{Bin}(n, p)$ and prior $\pi(p) \sim \text{Beta}(\alpha, \beta)$
  - Want to sample from joint density $f(x, p) = f(x \mid p)\pi(p)$
  - *Gibbs sampler*: Repeat the following
    - Sample x from $\text{Bin}(n, p)$
    - Sample p from $\text{Beta}(x + \alpha, n - x + \beta)$
  - $(X_l, p_l)_{l \geq 1}$ form a Markov chain on $\{0, 1, \ldots, n\} \times [0, 1]$
  - Let $\tilde{K}(x, p; x', p')$ be the transition density
  - How many steps (obviously depending on n, α, β) does this Markov chain converge to the stationary distribution?

- In a typical Bayesian course, we learn many convergence diagnostics that are often heuristic. Is there anything rigorous we can say about convergence rate of Markov chains?

- Distances between distributions:

  - *Total variation distance*:

$$\begin{aligned} \|\mu - \pi\|_{\text{TV}} &= \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \pi(x)| \\ &= \max_{A \subset \mathcal{X}} |\mu(A) - \pi(A)| \\ &= \frac{1}{2} \sup_{\|f\|_\infty \leq 1} |\mu(f) - \pi(f)|. \end{aligned}$$

  - L^p *distance* wrt π:
    Let $f(x) = \frac{\mu(x)}{\pi(x)}$ and $g(x) = \frac{\nu(x)}{\pi(x)}$. For $1 \leq p < \infty$,

$$d_{\pi,p}(\mu, \nu) = \|f - g\|_{L^p(\pi)} = \left(\sum_{x \in \mathcal{X}} |f(x) - g(x)|^p \pi(x) \right)^{1/p}$$

- The usual limit theorems are useless in practice:

  "There exist constants $A, B > 0$, $\rho \in (0, 1)$ such that $\|K^l(x, \cdot) - \pi\|_{TV} \leq A\rho^{Bl}$."

A, B, and ρ are some mysterious constants.

Can we get some more quantitative, useful bounds?

- Cutoff phenomenon (Diaconis, 1996)

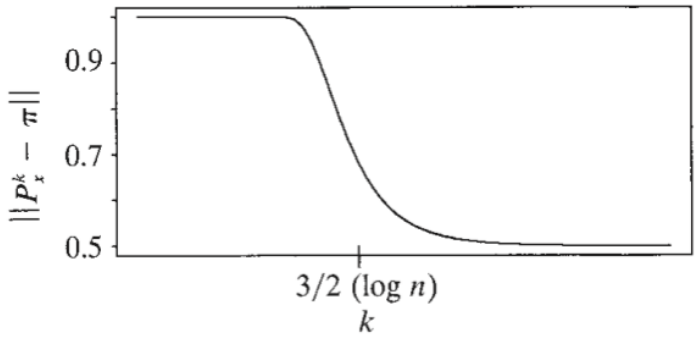


FIG. 1. **The cutoff phenomenon for repeated riffle shuffles of $n = 52$ cards.**

  - Riffle shuffle.

l	1	2	3	4	5	6	7	8	9	10
$\|K^l - \pi\|_{TV}$	1.000	1.000	1.000	1.000	0.924	0.624	0.312	0.161	0.083	0.041

  - Many more examples: Ehrenfest chain, random transposition, Gibbs sampler, ...

    "Cutoff phenomenon for XXX chain."

  - Not every Markov chain has a cutoff.
    A chain without cutoff: simple random walk on the integers mod n.

- Generic methods for studying convergence rates:

  - Algebraic methods (spectral analysis) - E.g, random walks on groups (shuffling cards), some Gibbs samplers, ...
  - Analytic methods - Geometric Inequalities.
  - Probabilistic methods - Coupling and strong stationary times.

- Algebraic method

– *Reversible* Markov chains.
If π is a probability distribution on $\mathcal{X}$ and

$$\pi(x)K(x,y) = \pi(y)K(y,x), \quad \text{for all } x, y \in \mathcal{X},$$

then π is the unique stationary distribution of K. E.g., Metropolis chain.

– K operates on $L^2(\pi) = \{f : \mathcal{X} \mapsto \mathbb{R}, \mathbf{E}_\pi[f^2] < \infty\}$ by

$$Kf(x) = \sum_{y \in \mathcal{X}} K(x,y) f(y).$$

– Reversibility of K is equivalent to *self-adjointness* of the operator K.

– By standard spectral theorem for self-adjoint operators, K has eigenvalues $1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{|\mathcal{X}|-1} \geq -1$ with (right) eigenfunctions $\{\phi_0, \ldots, \phi_{|\mathcal{X}|-1}\}$ that are orthonormal on $L^2(\pi)$

$$< \phi_i, \phi_j >_{L^2(\pi)} = \sum_{x \in \mathcal{X}} \phi_i(x)\phi_j(x)\pi(x) = 1_{\{i=j\}}.$$

– If we know all the spectral information (lucky!), then

$$d^2_{\pi,2}(K^l(x,\cdot),\pi) = \sum_{i=1}^{|\mathcal{X}|-1} \beta_i^{2l} \phi_i^2(x).$$

– Usually the upper bound is tight.

$$\|K^l(x,\cdot) - \pi\|_{TV} \leq \frac{1}{2} d_{\pi,2}(K^l(x,\cdot),\pi)$$

– When are we lucky? In presence of *symmetry*.
E.g., for random walks on groups, we only need eigenvalues, which can be derived from the irreducible representations of the group.
Definite reference for this topic is the book Diaconis (1988)

- Algebraic method for analyzing the baby Gibbs sampler

– The joint chain $\tilde{K}(x,p;x',p')$ is *irreversible*.

– The x-marginal chain $K(x,x')$ is *reversible*, with $m \sim$Beta-Bin(n,α,β) as stationary distribution. And

$$\|K^l_x - m\|_{TV} \leq \|\tilde{K}^l_{x,p} - f\|_{TV} \leq \|K^{l-1}_x - m\|_{TV}.$$

Thus sufficient to study convergence rate of the x-marginal chain.

- Some analysis (Diaconis et al., 2008) shows that K has
  - eigenvalues: $\beta_0 = 1$, $\beta_j = n_{[j]}/(n+\alpha+\beta)_{(j)}$, $j = 1, \ldots, n$.
  - eigen-functions: ϕ_j are the *Hahn polynomials*.
- Doing the summation gives

$$0.5\beta_1^l \leq \|\tilde{K}_{0,p}^l - f\|_{TV} \leq 3\beta_1^{l-1/2}.$$

- Cutoff phenomenon: $\frac{n+\alpha+\beta}{2(\alpha+\beta)}$ steps are necessary and sufficient for convergence.
- Similar analysis can be carried out for all following conjugate pairs (see KL 27.9)

TABLE 26.1. Conjugate Pairs

Likelihood	**Density**	**Prior**	**Density**
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	Beta	$\frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
Poisson	$\frac{\lambda^x}{x!} e^{-\lambda}$	Gamma	$\frac{\beta^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda}$
Geometric	$(1-p)^x p$	Beta	$\frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
Multinomial	$\binom{n}{x_1 \ldots x_k} \prod_{i=1}^k p_i^{x_i}$	Dirichlet	$\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}$
Normal	$\sqrt{\frac{\tau}{2\pi}} e^{-\tau(x-\mu)^2/2}$	Normal	$\sqrt{\frac{\omega}{2\pi}} e^{-\omega(\mu-\theta)^2/2}$
Normal	$\sqrt{\frac{\tau}{2\pi}} e^{-\tau(x-\mu)^2/2}$	Gamma	$\frac{\beta^\alpha \tau^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\tau}$
Exponential	$\lambda e^{-\lambda x}$	Gamma	$\frac{\beta^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda}$

- Analytic method
  - Upper bound through spectral gap

$$4\|K^l(x,\cdot) - \pi\|_{TV} \leq d_{\pi,2}^2(K^l(x,\cdot),\pi) \leq \frac{1}{\pi(x)} \beta_*^{2l},$$

    where $\beta_* = \max\{|\beta_1|, |\beta_{|\mathcal{X}|-1}|\}$.
  - Use some geometric inequalities to bound β_*.
  - Where to look up the material? Lecture notes by Saloff-Coste (1997).
  - (K, π) reversible on $\mathcal{X}$ (finite). Dirichlet form

$$\epsilon(f,g) = < (I-K)f, g > .$$

Fact

$$\epsilon(f,f) = \frac{1}{2}\sum_{x,y}[f(x)-f(y)]^2\pi(x)K(x,y).$$

- Lemma

$$1-\beta_1 = \min_{f\neq 1}\frac{\epsilon(f,f)}{\mathrm{var}(f)}$$
$$1-\beta_{|\mathcal{X}|-1} = \max_{f\neq 0}\frac{\epsilon(f,f)}{\mathrm{var}(f)}.$$

- Definition: Poincaré inequality

$$\mathrm{var}(f) \leq A\epsilon(f,f), \quad \text{for all } f \in L^2(\pi).$$

- We can bound β_1 by finding A

$$\beta_1 \leq 1 - \frac{1}{A}.$$

- Theorem (Poincaré Ineq for Markov Chains (Diaconis and Stroock, 1991)):

$$\beta_1 \leq 1-\frac{1}{A}, \quad \text{where } A = \max_e \frac{1}{Q(e)}\sum_{\gamma_{x,y}\ni e}|\gamma_{x,y}|\pi(x)\pi(y).$$

- Apply Poincaré inequality to the baby Gibbs sampler gives a horrible bound (something *exponential* ...)
- Other geometric inequalities: Cheeger, Nash, Log-Soblov inequalities (in a series of papers by Diaconis and Saloff-Coste).

- Probabilistic methods
  - By cleverness, we can get good bounds for convergence rate without using all those analytic methods.
  - Good starting point is the book (Diaconis, 1988) and the unpublished book by Aldous and Fill (available on Aldous' website).
  - *Coupling* - Wolfgang Doeblin.

Fig. 20: Cover of the DVD: "Wolfgang Doeblin. A mathematician rediscovered"

Fig. 21: The archive of the "Académie des Sciences

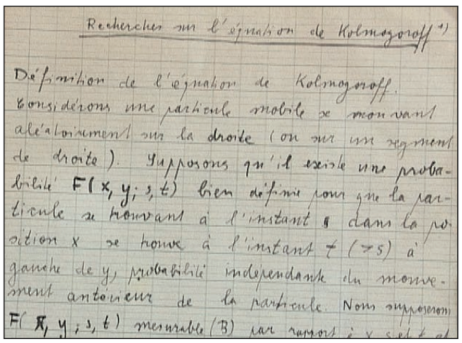


Fig.22: First page of the "pli cacheté no. 11668"

– For the baby Gibbs sampler, coupling gives an upper bound of order $n \ln n$ (off by $\ln n$).
– Yields useful bounds for hierarchical random effects model (Hobert and Geyer, 1998).
– Let's work on a simpler example: Borel's Shuffle (random to top, random to bottom, ...).
– *Coupling*: Two processes evolve until they are equal. *Coupling time* T. Coupling inequality

$$\|K_x^l - \pi\|_{TV} \leq \mathbf{P}(T > l).$$

– For Borel's shuffle, bound on coupon collector problem gives

$$\|K_x^l - \pi\|_{TV} \leq n\left(1 - \frac{1}{n}\right)^l.$$

Thus $l = n \ln n$ steps suffice.

- Summary
  - Keep Markov chains in your toolbox - useful for modeling, simulation, and combinatorial optimization.
  - Convergence rate of Markov chains is an interesting applied probability problem that often gives more insights into the chains.
  - A lot remains to be done for analyzing many MCMC algorithms being used.

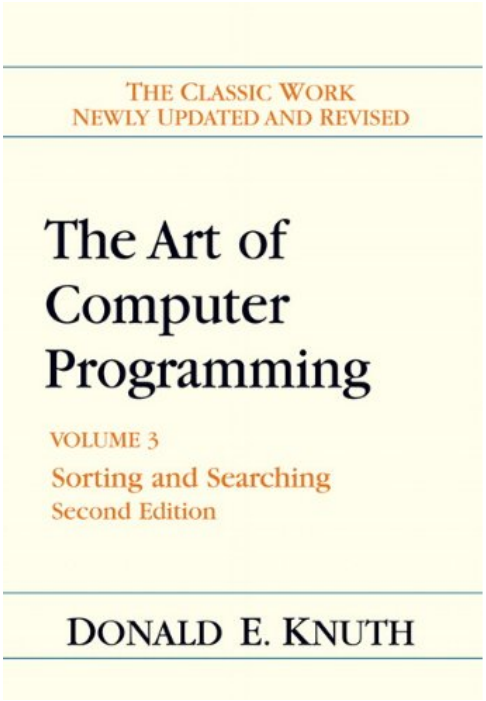
Sorting algorithms (JM 14.3, KL 1.10)

- Applications: order statistics (median, quantiles), QQ-plot, multiple testing (sorting p-values), Wilcoxon rank-sum test, ...
- *Bubble sort*: Locate maximum and put on top; find maximum in the $(n-1)$ list and put on the top second position; ...

  $O(n^2)$ average cost.
- Think about sorting *massive* data $n = 10^{12}$. On a teraflop computer. n^2 flops take 10^{12} seconds ≈ 31710 years, while $n \ln n$ flops take $10^{12} \log(10^{12})/10^{12} \approx 27$ seconds.
- Key idea: Divide and conquer.
- *Merge sort*: Recursively partition into two lists, sort them respectively, and then merge. $T(n) = 2T(n/2) + O(n)$. Solution is $T(n) = O(n \log_2 n)$.
- *Quick sort*: Randomly select a pivot element, split into 3 lists, and do some swaps so that the pivot is in the right position.

$$T(n) = \frac{1}{n} \sum_{j=1}^{n-1} [T(j-1) + T(n-j)] + n - 1 = \frac{2}{n} \sum_{j=1}^{n-1} T(j) + n - 1.$$

  Solution is $O(n \ln n)$.
- Sorting is a well-trodden area in computer science. Mature functions/libraries in standard softwares. The "bible" on this topic is (of course) Knuth (2005).



Fast Fourier transform (FFT) (KL Chapter 20, JM 14.5)

- History: Cooley and Tukey (1965)

  John Tukey: "bit", box-plot, "learning from the experience of the others", multiple comparison, FFT, ...

  Tukey conceived the FFT algorithm during meetings with President JFK's Science Advisory Committee. They need fast ways to analyze seismic waves to detect nuclear weapon tests in Soviet Union. Richard Garwin of IBM immediately realize the potential of this fast algorithm and referred Tukey to Cooley to implement it.

  People also believe Gauss essentially used the same strategy when solving his least squares problem!

- Applications in statistics: convolution, time series, branching process, ...

- Consider two independent random variables on $\{0, 1, \ldots, N-1\}$:

$$X \sim \{p_0, \ldots, p_{N-1}\}, \quad Y \sim \{q_0, \ldots, q_{N-1}\}.$$

  What's the distribution of the sum $Z = X + Y$?

- $z_k = \sum_{j=0}^{k} p_j q_{k-j}$, $k = 0, \ldots, 2N-2$. $O(N^2)$ computation.
- Do DFT of both sequences, multiply together, and inverse DFT. $O(N \ln N)$ computation!

- Discrete Fourier transform (DFT) of a vector $\boldsymbol{x} \in \mathbb{R}^N$.

$$a_k = \sum_{j=0}^{N-1} w^{jk} x_j, \quad k = 0, \ldots, N-1.$$

where $w = e^{-2\pi\sqrt{-1}/N}$. Note w is an N-th root of 1. DFT is essentially matrix-vector multiplication

$$\boldsymbol{a}^\top = \boldsymbol{x}^\top \boldsymbol{W}, \quad \boldsymbol{W} = (w^{jk}),$$

which usually costs $O(N^2)$ flops.

- Suppose $N = N_1 N_2$. Index rewriting:
  - $j \leftarrow j_1 N_2 + j_2$ (fill out N_1-by-N_2 matrix in row major),
  - $k \leftarrow k_2 N_1 + k_1$ (fill out N_1-by-N_2 matrix in column major),
  - $j_1, k_1 \in \{0, \ldots, N_1 - 1\}$, $j_2, k_2 \in \{0, \ldots, N_2 - 1\}$.

  Then

$$\begin{aligned}
a_k &= a_{k_2 N_1 + k_1} = \sum_{j=0}^{N-1} w^{jk} x_j \\
&= \sum_{j_1=0}^{N_1-1} \sum_{j_2=0}^{N_2-1} w^{(j_1 N_2 + j_2)(k_2 N_1 + k_1)} x_{j_1 N_2 + j_2} \\
&= \sum_{j_1=0}^{N_1-1} \sum_{j_2=0}^{N_2-1} w^{j_1 k_1 N_2 + j_2 (k_2 N_1 + k_1)} x_{j_1 N_2 + j_2} \\
&= \sum_{j_2=0}^{N_2-1} w^{j_2 (k_2 N_1 + k_1)} \sum_{j_1=0}^{N_1-1} (w^{N_2})^{j_1 k_1} x_{j_1 N_2 + j_2} \\
&= \sum_{j_2=0}^{N_2-1} (w^{N_1})^{j_2 k_2} w^{j_2 k_1} \sum_{j_1=0}^{N_1-1} (w^{N_2})^{j_1 k_1} x_{j_1 N_2 + j_2}.
\end{aligned}$$

- Essentially we need to do N_2 DFT of length-N_1 sequences and then do N_1 DFT of length-N_2 sequences. Total cost

$$T(N) = T(N_1 N_2) = N_2 T(N_1) + N_1 T(N_2).$$

Suppose N is a power of 2. Then $T(N) = (N/2)T(2)+2T(N/2)$ and the solution is $T(N) = O(N \ln N)$!.

- Inverse DFT. $\boldsymbol{W}^{-1}$ has entries w^{-jk}/N. Then

$$\boldsymbol{x}^\top = \boldsymbol{a}^\top \boldsymbol{W}^{-1}.$$

- Variants for prime N: still $O(N \ln N)$. But always a good idea to pad with zero to get N as a power of 2.

- Generalizations to 2D, 3D FFT available.

- Mature libraries/functions for both CPU and GPU.

- Galton-Watson process. Survival of families. Lotka data (Lotka, 1931a,b). Using 1920 census data, the *progeny generating function* for a white male

$$\begin{aligned} P(s) &= .4982 + .2103s + .1270s^2 + .0730s^3 + .0418s^4 + .0241s^5 \\ &\quad +.0132s^6 + .0069s^7 + 0.0035s^8 + .0015s^9 + .0005s^{10}. \end{aligned}$$

PGF for the first generation $P_1(s) = P(s)$. PGF for the second generation $P_2(s) = \sum_k p_k P(s)^k = P(P(s))$. In general, PGF for the i-th generation is $P_i(s) = P(\cdots P(s))$ (i recursions).

Extinction probability of a family: $\lim_{i\to\infty} P_i(0) = P(\cdots P(0)) = 0.88$, or solving for $P(s) = s$.

What if we want to know the distribution of the i-th generation? Extend the generating function P_i to unit circle $P_i(w^k) = \sum_j p_j w^{jk}$, $w = e^{-2\pi\sqrt{-1}/N}$, where $k = 0, \ldots, N-1$ for N large. So $P_i(w^k)$ is the DFT of distribution p_j of i-th generation. Then apply inverse DFT to retrieve p_j. $O(N \log N)$ cost!

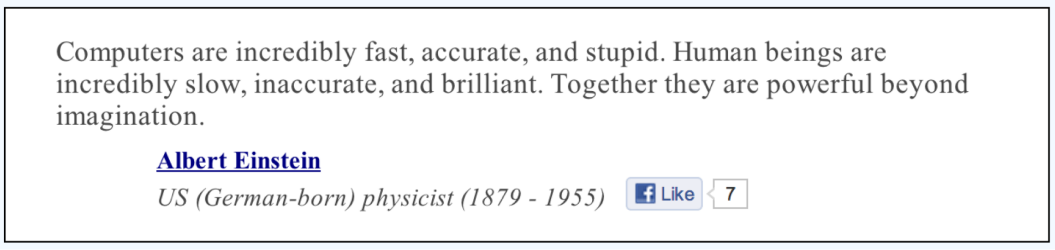
- Continuous-time branching process. Solve differential equation for $P_t(w^j)$ at any time t. Then apply inverse DFT.

- See JM 14.7 for more applications of FFT in statistics.

Take-home messages from this course

- Statistics, the science of *data analysis*, is the applied mathematics in the 21st century
  - Read the first few pages and the last few pages of Tukey (1962)'s *Future of data analysis* (`http://www.stat.ncsu.edu/people/zhou/courses/st810/notes/Tukey61FutureDataAnalysis.pdf`).
- *Big data* era: Challenges also mean opportunities for statisticians
  - methodology: big p
  - efficiency: big n and/or big p
  - memory: big n, distributed computing via MapReduce (Hadoop), online algorithms

- Being good at computing (*both* programming and algorithms) is a must for today's working statisticians.



- HPC (high performance computing) $\neq$ abusing computers.

  Always optimize your algorithms *as much as possible* before resorting to cluster computing resources.

- Coding

  - Prototyping: R, Matlab, Julia
  - A "real" programming language: C/C++, Fortran, Python
  - Scripting language: Python, Linux/Unix script, Perl, JavaScript

- Numerical linear algebra – building blocks of most computing we do. Use standard *libraries* (BLAS, LAPACK, ...)! Sparse linear algebra and iterative solvers such as conjugate gradient methods are critical for exploiting structure in big data.

- Optimization

  - *Convex programming* (LS, LP, QP, GP, SOCP, SDP). To do in ST790-003. Convex programming is becoming a *technology*, just like least squares (LS).
  - Specialized optimization algorithms for modern statistical learning problems. To do in ST790-003.
  - Generic nonlinear optimization tools: Newton, quasi-Newton, (nonlinear) conjugate gradient, ...
  - Specialized tools in statistics: EM/MM, Fisher scoring, Gauss-Newton, simulated annealing, ...

  - Combinatorial optimization techniques: divide-and-conquer, dynamic programming, greedy algorithm, ...

- MCMC: take a Bayesian course!