8.11.03 Nips | See HMC folder for earlier work on HMC

HMC

(293–297, 303, 304)

0: 303.40 Then t. "pc per datum" is $\propto \frac{\Delta}{\sigma}$, : $\frac{\sigma}{\Delta}$ is t. no. of poss. values — it is t. "no. of choices" per prodn $\equiv$ rcost.

One way: The "constant bias" "coif" is common to all models, so we put it on "free" (except for precision part — which is minimized alay on other coifs)

T. "next" coeff is given an a priori that is given by a mean of ϕ (?) σ^2 that is t. σ^2 of t. other coifs. — possible values are more likely for first coif.

So: what is t "σ^2" for t first (non-bias) coif? Make it 1.

---

08 My present impression: 33.27 is wrong: previous is what is right. → .32

Consider $2^{-\text{kause}(x)}$ — x is an integer. This is t. pc of t. integers.

$2^{-l(x)^*}$ trys to approximate it.

.10 converges to zero for $x \to \infty$, perhaps more slowly than any recursive funct., yet $\sum_0^\infty .10 < \infty$.

We are interested in 3 aspects of .10 at least

1) Value at x=1 (small x behavior)
2) behavior at x=∞ (large x behavior)
3) pt. at which transition from 1) to 2) occurs.

---

20 By going beyond prim. rec. functs, we get closer to t. behavior at ∞ ... but we almost never have x large enough for very Ackerman or Ak^{-1} or $(Ak)^{-1}$.

1) $\frac{a}{x}$ may be fine for small x, but value of $\underline{\underline{a}}$ is machine dependent (i.e. depends on experience)

3) Transition pt. is prob. machine dependent: depends much on "experience".

Hm. is .10 really relevant to HMC? My earlier impression was that t. problem was comparing t. volume of a k dim. to a k+1 dim space — that unless t. bounds were given, this comparison couldn't be made.

30 But consider t. finite resoln (Gaussian) spaces for data. This makes it an apparently well defined (tho unsolvable!) problem — getting short codes for a finite digital sequence.

---

32 (08) The idea is a GRU soln. Find a situations in past, w. decisions about "HMC".

Devise ways to categorize HMC problems to make a give common classes w. common "devices", so we can "pool" their HMC info. It's important to write this up for GENERAL cases; It's a very impt. problem!

$e^5 = 148.413$

We have cost of defining t. classes, and t. resultant savings by defining t. classes.

In case of HMC, t. amt. "saved" is unclear because we don't know what t. default code is!

$\cos^{-1} x = \sin^{-1} x$

$\sin^{-1}\cos^{-1} x = x$

$\sin^{-1}c^* = \frac{\pi}{2} - x = r = r^{-1}$

$s^{-3}r\, c^3 = x$

$r\, c^3 x = s^3 x$

$x = sx$

$cx = \sqrt{1-x^2}$

$r\, c^2 x^{1/2} = s^2\sqrt{1-x^2}$

$\cos 2 \quad \sin .7$

.7487

.9686

$\cos^{-1}\sin x = \frac{\pi}{4} - x = r$

$\sin^{-1}\cos x = \frac{\pi}{4} - x = r$

$r\,\sin^3 x = \cos^3 x$

$r\,\sin^2 x = \cos^2 \; r x$

$\cos \cos r x$

$= \cos \sin x$

$\cos^4 x = \sin^4 x$

$\cos^4\cos^{-1} x = \sin^3 \sin x \cos^{-1} x$

$\cos^3 x = \sin^3 r = \sin^2\cos x$

$\sin^2 x = \cos^2 x$

$\sin x \cdot \sin x \cdot \sin^{-1} x =$

$\cos \cdot \cos \cdot \sin^{-1}$

$\cos x = \sin x \quad x = \frac{\pi}{4}$

HMC HMC? 10 oclock Sept 6-7 Oct 8

10: (Spec 297.40): If the ratio of rootspace to coeff space densities is constant (with change of no. of coeffs) then if rootspace ratio is $\frac{\pi}{2}+2$ ($\approx$ 3.57), then the ratio has to be/constant in same coeff space as well — which seems unlikely, but may be true.

I think I ran into folk difficulty trying to compute hvol. of rootspace: I took random pts in root space & took mean of the Jacobian. Unfortly, while this would eventually converge, it converges very slowly. This is because there are several (I don't remember how many) very small regions of very hy Jacobian — the empirical result was that the variance of "estimate up to now" seemed to ↑ as the n ↑. — So I'd never (or for a long time) ever find true mean w/ any precision.

$\pi \div 2 =$

~~3.14885~~

3.57079

$\sqrt{\ } = 1.8896 \approx 2$

$\frac{\pi}{2} = 1.5708$

0 Anyway, it may well be that for large n (no. coeffs) the results for coeffs space are quite diffrnt from results for root space.

Also, it's not clear/as to why Jacobians should get particularly large in any regions — even on edges of root space!

15 Anyway, this "Grain prob" problem is interesting & important! I should study it & w problems (Grue; STEIN Betting averages (297.38), etc).

I think the whole pitch of Non-parametric statistics was to take params on variables in which people agreed on apriori. I think there may be many params ~~that~~ for which people agree on Apriori. These may be "builtin" to humans or into ~~our~~ "the physical laws" of our universe

→ { I think that people normally are unaware of/the arbitraryness of the apriods they assume.

→ I am pretty much more able to work on these (.15) problems than ever before.

26 The idea of the Scientist continually updating his ~~~~ apriod is important.

27 **ON THE OTHERHAND!** : [See 304.03] In [redacted] SolLM78 I had this [redacted]

[redacted] b cost astron equation of $\boxed{2+Ab \log n}$ b = no. of coeffs, n being no. of data pts

30 "A" was a no. characteristic of the model. It may be that $A = \frac{1}{2}$ for most linear & non-linear cases w. large n. so $\text{cost} \propto n^{Ab}$. ~~The~~ An interesting thing was that the mean bcost per new coeff was just $A \log n$ — same for every coeff!

For me thing, coeff space is equivt. in all dimension.

It would seem that the st. ~~~~ is crazy! That the only thing one needs to describe the models is the set of coeffs (w. prob dist) w. bcosts that depend on precision need — .. dependance on "n" in a simple ~~~~ way!

HMM! The Apriod of σ^2/ d.f. need not be simple. — it is a f'n of no. of coeffs, so ~~we compare it~~. So we compare σ^2's ~~for~~ diffrnt no. of coeffs. The smaller σ^2 is, the less info we need in each predn. Say we have a "Grain size" of Δ, Spec. 304.00

Nips HMC

"Because" the constraints on the roots, I'd guess at the d.f. being uniform in Rootspace —

→ But the Real Q (for correct answer) is How to use "previous Experience"?

However, because the real — complex roots being treated differently (296.20–28), it's not entirely clear as to what a "uniform d.f. in root space" means.!

Anyway, the solution to the Geom proby problem is the "same" as Soln to "Grue":

In the Case of Grue, we have data of past known universe, sometimes using the "Grue" (data) Concept, sometimes not. This gives us a pc of the symbol "Grue" as used in coding/predn. This pc extrapolates into future & is updated each time we use it for predn.

In a similar way, we can "parametrize" the selection of "kind of coeff".

For small sz, use only 1 param. — say pc is same for "1 more param" indep of n. I guess there are many other parametrizations possl. — uniformly in root or coeff space are two others: Also distance of set of coeffs from "stability" manifold boundary might be of interest.

Also, to add a new coeff: look at d.f. of previous coeffs, this would give a reasonable d.f. for the new one! Say, use Gaussian d.f. so μ, σ would characterize it. If coeffs tend to have other d.f. (not Gaussian) use Real d.f. w. ~~2~~ 2 params — maybe more! (instead of e^{-x^2} use $e^{-|x|^k}$ to get tails of varying "fatness")

What Corpus should be used to parametrize our "Geom proby"? In the case of SM, use any securitys that seem "near" to those being predicted.

A not bad corpus would ~~seem~~ be simply the more distant past of the present stock.

— Too far back to use the same coeffs over entire corpus, but o.k. to use to parametrize choice of next coeff: (Say we use x window-out. ~~at~~ any other window of essentially finite width). — (For a given window width, k, we could only have a certain maximum no. of coeffs!) WOOPS!

Actually there is interaction betw. W (window width) and no. of usable coeffs, so ~~we would~~ the way d.f. of no of coeffs is a way to tell what W is (to some extent but probly not exactly). (= ssz) .so — not at all clear! This might be possl. — for large W, only 3 coeffs are possl. & predns not so good for such a large corpus. For smaller W, we can use more coeffs, that ~~may~~ reflect peculiarities of recent data. not obviously correct!

It might be instructive to calculate densities for a few cases & contrast root space w. coeff space results.

38 (~SN) In that STEIN, Baseball batting averages, case, I had results critically dependent on behavior of d.f. for large ssz ($=\infty$) or something. "behavior at ∞" characteristic. How should this be treated?

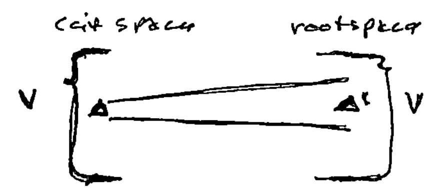
Spec 303.00

8.7.03
Nips

$\boxed{\text{HMC}}$ Import New Tack !

00 : I know the H-volume in Root space of non-divergent solns) (sets of coeffs)

For a particular set of coeffs I know the h-vol in coeff space of the "fitting" sets of coeffs (this depends on ssz as well as nature of data itself). I also know the Jacobian at that pt. Is this enough info? : How does it compare w.r.t. "other soln"?



Δ is size of acceptable region in coeff space; Δ is known.

Δ' is size of acceptable region in Root space: since the Jacobian known, Δ' is known.

V' is H vol of convergent region of Root space: known.

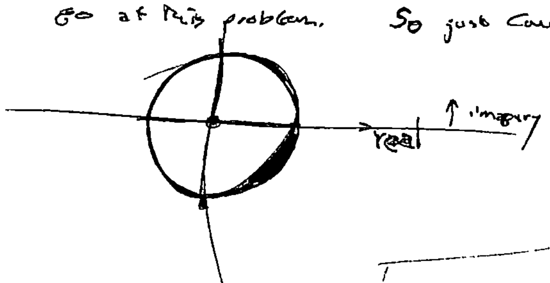
10 V " " " " " coeff space: knowable but very diff to calculate.

The "info" from V, Δ perspective is $\ln \frac{V}{\Delta}$
" " " V', Δ' " " $\ln \frac{V'}{\Delta'}$
} No reason to believe they are the same.

Its pretty much the problem of $\boxed{\text{Geom probability}}$. e.g. does specific Gravity or Specific Volume have "uniform apriori"?

.17 There seems to be some diffy in computing H-vol in rootspace because the complex roots are always paired... I think I may have figured out how to deal w. this at my last go at this problem. So just consider Quadratic Equ:

20



So possibilities are ① 1 pt. in upper half ⊙, ② 2 pts on real segment [−1, ..+1]

So maybe sum of areas? $\frac{1}{2}\pi + 4$?

only $\boxed{\frac{1}{2}\pi + 2}$ because the order of the 2 real roots is irrelevant.

So total H volume : $\boxed{\left(\frac{1}{2}\pi+2\right)^k \cdot 2^\alpha}$ k = no. of root pairs: α = 0 if no. of roots is even, α = 1 " " " " odd.

28 Same Q about redundancy in ordering of Quad roots.

I have no idea as to order of magnitude of Jacobian. I think that at certain finite no. of pts, the values were very hi or very low, but did not ≡ ∞ or 0. 30

But re: the Geom proby problem: My impression was that it was normally solved by using "previous experience". One form of such "previous experience" would be the probability distribution over "no. of coeffs", for probs in the "Macro Corpus" of interest.

Actually, the very General "Geom proby" problem is of much (interest impt) — Just how does

38 → one use past experience to get the needed Apriori?

.38 is the more General problem, but "Geom proby" is a very common (diff) manifestation.

8.6.03 ~~Wavering sign?~~ 295

N.P.B. HMC (How Many Coeffs)

20 : One could express each set of k successive z's as a fixed matrix times the previous set of k successive z's. So we have to do powers of the matrix — which converges if all of its |eiv's| are < 1.

$$\begin{bmatrix} a_1 + \lambda & a_2 & \cdots & a_k \\ & +\lambda & & \\ & & +\lambda & \\ & & & \ddots \\ & & & & +\lambda \end{bmatrix}$$

what's Det of this Matrix? looks like $(a_1+\lambda)\lambda^{k-1}$.

Getting the $k \times k$ matrix might be a fast way to decide convergence because one could successively square the matrix! This amounts to doubling the jump each time — even tho matrix mult takes k^3 steps, the squaring effect will always win! So, if any matrix element got > 1000, say, we assume divergence.

10 Another trick I remember: for a $k \times k$ matrix, the first k powers have all info one needs for higher powers (— Is this true?)

To square a $k \times k$ matrix k^3 mults or k^2 dot products, so doing it n times is nk^2 dot products, equiv to $k2^n$ jumps, which would normally take $k2^n$ dot products, so the factor of $\frac{k2^n}{nk^2} = \frac{2^n}{kn}$

so for large k, it's not immediately efficient say $k=10$, $n=10$ $\frac{2^{10}}{10 \cdot 10} = 10$. for $k=5$ $n=10$ it's 20.

The correct matrix for $\vec{z}(t) \to \vec{z}(t+1)$ is

$$\begin{bmatrix} 0 & 1 & \cdots & & \\ 0 & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ v_1 & v_2 & v_3 & v_4 \cdots & v_k \end{bmatrix} \qquad \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & 1 & \\ & & & \ddots & \ddots \\ & & & & \lambda & 1 \\ v_1 & v_2 & \cdots & & & v_k + \lambda \end{bmatrix}$$

← This would be the det matrix. anyway: removing the λ's to get the matrix one can be repeatedly squared.

The multiple squarings the first few times has lots of 0's which may waste time.) Put separate cond. to see if either of factors is ϕ.

20 To write this prog, I have to know max/min of each coeff. Ex: $\pm$ Σ $v_n = 1$.

$n=1$ $(x-\lambda_1)$ [redacted] $V_1 = 1$ $V_0 = \pm 1$

25 $$(x-\lambda_1)(x-\lambda_2) = \underset{V_2}{x^2} - \underset{V_1}{(\lambda_1+\lambda_2)x} + \underset{V_0}{\lambda_1\lambda_2}.$$

$V_0 = \pm 1$ $V_1 = \pm 2$ $V_2 = 1$

but these at least are correlated.

$V_0 = +1$ $V_1 = +2$
$V_0 = +1$ $V_1 = -2$ } are both possl.

$V_0 = -1$ $V_1 = 0$

I worked all of this out in previous work on HMC: perhaps try to find it.

For $k = 2, 3$ maybe 4 I can "do it by hand".

30 For about ½ of the coeffs, the full $\pm$ range will often be possl. For ½ of the others, maybe only ½ is avail. be so the ~~true~~ volume of stable pts will be $\sim 2^{-\frac{k}{2}}$

for $k=10$ this is $2^{-5} = \frac{1}{32}$. — wdely

Also note that many of the divergent pts will diverge rapidly, so little time spent on them.

BUG?? ! true(25R) If I use $k=3$, coeffs $= 1, 2, 1$ The roots are $-1, -1$ and the seq. should not diverge! But since the coeffs are all ≥ 1, it would seem that iteration would always give values that would diverge!

Start w. 1 → 1, 2, 3, 6, 12,

$1 + 2A^{\alpha} + A^2 =$

$$\left\{ \begin{aligned} &A(n+1) = A(n) + 2A(n-1) + A(n-2) \\ &A^{n+1} = A^n + 2A^{n-1} + A^{n-2} \\ &A^3 = A^2 + 2A + 1 \end{aligned} \right.$$

⟹ I may have bad program on char. funcs wrong! check this out →

Nies

HMC

00 : On my work on the How many Coeffs problem: I assumed that the coeffs were real, the roots had to be in conj. pairs / Is this true?

~~[illegible]~~ $x^3+1=0$ has 3 roots: 2 conj pairs, 1 real.

$(x-i)(x-i) = x^2-2ix-1$

A possl. way to get density of ~~soluti~~ ~~[illegible]~~ "stable" sorns in coeff space:

pick random pts in the finite space in which each coeff is uniform betw its limits.

For each pick, we can either solve the polynomial or do an iteration of the coeffs (zero noise) for a long distance to see if it blows up (like Mandelbrot sets). The edges may be fractal, but the ~~[illegible]~~ volume of the legal region may not be hard to estimate.

10 For k coeffs, do we start w. a random k-vector?

Obtain an analytic expression for the r^{th} iteration: perhaps there is a fast way to compute it.

Pentium MMX ~~[illegible]~~ does vector dot products

13 $a_1 \ldots a_k$: $a_{k+1} = \sum' a_1 v_1 + a_2 v_2 \ldots a_k v_k$

$a_{k+2} = \quad a_2 v_1 + a_3 v_2 \ldots a_{k+1} v_k \quad = \quad a_2 v$

15 Its a circular pgm: we only keep the ~~[illegible]~~ k values ~~of~~ a_i, and overwrite one value at each new dot product.

There may be a fractal effect wrt. the starting vector $a_1 \ldots a_k$. I can try this empirically;

Say I have a fixd $\vec{V}$: I try various initial $\vec{a}$ values & see if certain ones converge, others diverge.

I think its possl. to figure out what will happen, from simple analysis — i.e. Why the roots outside unit ⊙ cause divergence.

22\. From the analysis of .13 – .15 I think the value at any pt. is a linear function of the values of $a_1 \cdots a_k$, so we always have k coeffs & correspondingly, k roots. So maybe the roots ~~of~~ of each successive iteration are a simple function of the previous set of roots!

If .22 is true, I can do an analysis of $\vec{a} = 1,0,0,0 \ldots$, $\vec{a} = 0,1,0,0 \ldots$, $\vec{a} = 0,0,1,0,0 \ldots$

I do ~~k~~ runs. If any of them diverge, then I assume divergence.

try $\vec{a} = 1,0,0 \ldots$; $a_{k+1} = v_1$; $a_{k+2} = v_1 \cdot v_k$, $a_{k+3} = \underbrace{v_1 \cdot v_{k-1} + \cancel{v} \; v_1 \cdot v_k \cdot v_k}$ | $a_{k+4} = v_{k-2} \cdot v_1 + v_k \cdot v_{k-1} v_1 + v_1 v_{k-1} v_k + v_1 \cdot v_k^3$

v_1; 1; 1, 0; 1, 1, 0²

v_{k-1} v_k

v_{k-2} v_{k-1} v_k

1 ; v_k ; $v_{k-1} + v_k^2$; $v_{k-2} + \underbrace{v_k v_{k-1} + v_{k-1} v_k}_{2} + v_k^3$ — looks like Pascal △ 1, 11, 121, etc. ← Maybe not!

1; 1 1; 1 2 1

v_{k-2} v_{k-1} v_k

$v_k v_{k-2} + \underbrace{v_{k-1}^2 + v_{k-1} v_k^2 + v_{k-2} v_k^2 + v_k^2 v_{k-1}}_{} + v_k^4$

$v_{k-1}^2 \qquad 4 v_{k-1} v_k$

.40 $0 \cdot 2 + 1^2 + 4 \cdot 1 \cdot 0^2 + 0^4$

So omit v's & instead of $k-1$ write "1". Mo ~~[illegible]~~ & on line .40;

With a suitable compact notation, I may be able to write a lot of terms & their general form of term.

8.5.03

NIPS

ALP General. "IF & ONLY IF" for ALP & Conv. Prm.

50 : Ont. "Inverse Conv. Prm." (1) IS ALP t. any soln? (2) If μ, t. seq. generator is known w. $P<=P(\mu)$; is KL distance of soln. w. μ $-\ln P(\mu)$ t. BEST that can be done?

(3) If we allow larger errors, Bayes error is $\leq$ greater than $-\ln P(\mu)$ by factor, k or by additive amts α.

(4) I'm not talking about "assymtotic" results for infinite seq length, but short finite Seqs.

(5) How large is t. dom. space of pgms that give seqs a distance d, from t. pgm generated by μ? This could be very useful!

(6) Perhaps Get Marcus to Work on it!

(7) A Guess; That for large SSZ, (i.e. big corpus)

10

Some Quick notes on Univl dist. on $\boxed{\text{Continuous functions}}$ of a $\boxed{\text{real (or complex)}}$ variable:

We can descr. Continuous functions w. a finite no. of bits.

One big uncertainty is in t. a prior of continuous params.

Alex Suggested Using floating pt. notation! This does give a finite sized space (on high & low ends).

Perhaps use Very large word. Since one integrates over all derns of functions that are consist w. t. data, ~~Using~~ t. size of word may become irrelevant for word sizes > a certain amt.

Hvr, my Guess, is that If t. space of interest Gets Very large we have to pay for it — by t. apr. p of smaller values decreasing.

Another tack: Somehow by t. Context of t. problem, one has some idea as to t. Size of t. space.

→ Re floating pt. notation: My intuitive idea is that One should not use ordinary floating pt. That one initially tells what t. upper bound of t. space is, and then if one wants precision, you get it by integration over many possl. $\underline{\underline{U}}$ values.

Better: $\pm$(upper bnd) one has a priori Uniform dist.

26 One apparent Serious diffy: Linear regression: when one introduces 1 more coef, the cost depends much on one's upper bnd for t. coef.

A possl "Soln" to 26! t cost of one more coef is always "α". α is found by trying various values on one's corpus & selecting a value that "works best".

30 Hvr: T. value of α may depend critically on t. corpus & on how many coeffs are needed to predict it!

For ~~[illegible]~~ binary variables, small values are always of hyper apr.p.

For linear Regression, one can assume "stability" so roots of char. eq. are within (unit + ε) circle ("+ε" to allow for some "instability"). For each coeff, one can get max and min values. A rough guide would be Uniform apr p.d. betw. Min & Max for all coeffs.