CONJUGATE GRADIENT & NESTED SAMPLING

John Skilling Cambridge, Sept 2008.

Nested sampling proceeds by compressing enclosed volume X.

$n = 3$ samples.

New L^*, compression $\Delta \log X = -\frac{1}{n}$.

It plots $L(X)$, giving evidence $Z = \int L\,dX$ and posterior, weight$(\underset{\sim}{x}_i) = L_i \Delta X_i / Z$.

It needs uniform samples within $L(\underset{\sim}{x}) > L^*$.

We have relied on traditional exploration (Gibbs etc.). This ignores gradients.

Computation of $L(\underset{\sim}{x})$ can easily get ∇L also: a million times more info. for twice the cost.

But we seem to lose detailed balance.

Actually, we CAN do it!

The archetype problem is $\log L = -\frac{1}{2}\chi^2$.

$\chi^2 = |R\underset{\sim}{x} - \underset{\sim}{D}|^2$, D = data, R = "Opus".

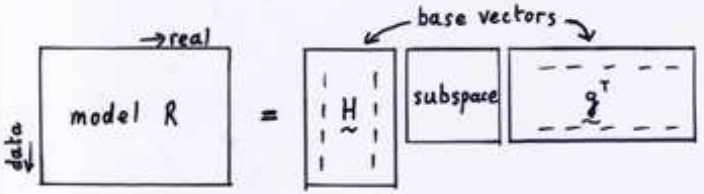
$\nabla L = -R^T(R\underset{\sim}{x} - \underset{\sim}{D})$, $\nabla\nabla L = -R^T R$.

MATRIX MODELLING (secrets from 20 yrs ago)

Seek to simulate matrix operations (inversion ...) on

$A = R^T R$

Feedback; "Opus" real-to-data; "Tropus" data-to-real



Wider subspace ⟶ closer modelling

Generate vectors from a seed:

$\underset{\sim}{q}_1$ = seed

$\underset{\sim}{q}_2 \sim \{\underset{\sim}{q}_1, R^T\underset{\sim}{H}_1\}$ ⟷ $\underset{\sim}{H}_1 \sim R\underset{\sim}{q}_1$

$\underset{\sim}{q}_3 \sim \{\underset{\sim}{q}_1, \underset{\sim}{q}_2, R^T\underset{\sim}{H}_2\}$ ⟷ $\underset{\sim}{H}_2 \sim \{\underset{\sim}{H}_1, R\underset{\sim}{q}_2\}$

⟷ $\underset{\sim}{H}_3 \sim \{\underset{\sim}{H}_1, \underset{\sim}{H}_2, R\underset{\sim}{q}_3\}$

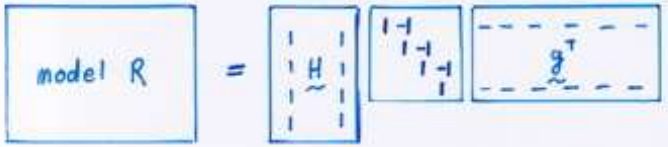
Orthogonal sets of vectors preferred.

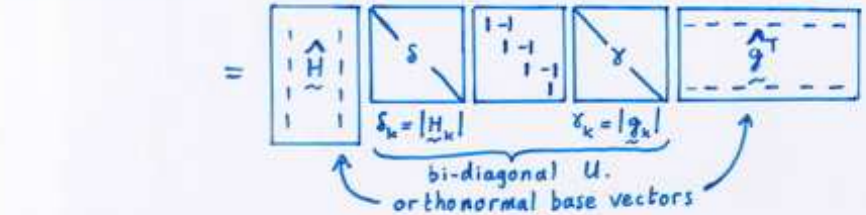
Simplify the code to **kernel**

$\underset{\sim}{H}_0 = \underset{\sim}{0}$	, $\underset{\sim}{q}_1$ = seed
$\forall k$: $\underset{\sim}{H}_k = \underset{\sim}{H}_{k-1} + \frac{R\underset{\sim}{q}_k}{\underset{\sim}{q}_k^T \underset{\sim}{q}_k}$	, $\underset{\sim}{q}_{k+1} = \underset{\sim}{q}_k - \frac{R^T\underset{\sim}{H}_k}{\underset{\sim}{H}_k^T \underset{\sim}{H}_k}$

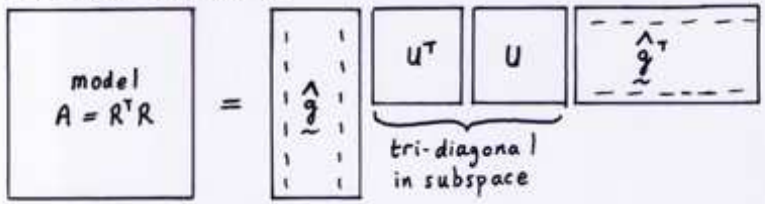
Check: $\underset{\sim}{q}_i \perp \underset{\sim}{q}_j$, $\underset{\sim}{H}_i \perp \underset{\sim}{H}_j$

These "nicest" vectors have

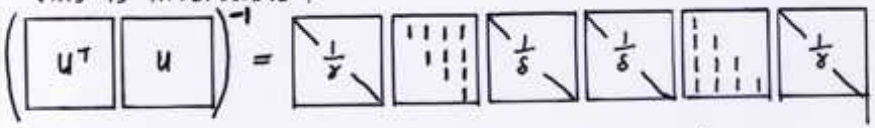




Feedback matrix is



This is invertable :



Check: With seed $\underset{\sim}{b}$, this models $\underset{\sim}{x} = A^{-1}\underset{\sim}{b}$

by maximising $q = 2\underset{\sim}{x}^T\underset{\sim}{b} - \underset{\sim}{x}^T A\underset{\sim}{x}$ in subspace.

Merge accumulation of $\underset{\sim}{x}$ and q into kernel.
Get standard conjugate gradient for $A^{-1}\underset{\sim}{b}$.

	$\underset{\sim}{g}_1 = \underset{\sim}{b}$, $\underset{\sim}{h}_0 = \underset{\sim}{0}$.	$\underset{\sim}{x}_0 = \underset{\sim}{0}$.	$q_0 = 0$
$\forall k$:	$\gamma_k^2 = \underset{\sim}{g}_k^T \underset{\sim}{g}_k$,	$\underset{\sim}{h}_k = \underset{\sim}{h}_{k-1} + \underset{\sim}{g}_k/\gamma_k^2$	$\underset{\sim}{H} = R\underset{\sim}{h}$
	$\delta_k^2 = \underset{\sim}{h}_k^T A \underset{\sim}{h}_k$,	$\underset{\sim}{g}_{k+1} = \underset{\sim}{g}_k - A\underset{\sim}{h}_k/\delta_k^2$	
	$q_k = q_{k-1} + 1/\delta_k^2$.	$\underset{\sim}{x}_k = \underset{\sim}{x}_{k-1} + \underset{\sim}{h}_k/\delta_k^2$	

1. As subspace widens,
$0 = q_0 < q_1 < q_2 < \cdots \leq Q = \underset{\sim}{b}^T A^{-1}\underset{\sim}{b}$

2. $\underset{\sim}{g}$ are gradients ∇Q at successive maxima $\underset{\sim}{x}$.

3. Precursors $\underset{\sim}{h}$ of $\underset{\sim}{H} = R\underset{\sim}{h}$ are conjugate
($\underset{\sim}{h}_i^T A \underset{\sim}{h}_j = 0$) —— hence name.

Modelling (J.S.) $\Longleftrightarrow$ Maximisation ("Establishment")

Regularisation (to stabilise small eigenvalues of A)

Instead of $\max\ q = 2\underset{\sim}{x}^T\underset{\sim}{b} - \underset{\sim}{x}^T A\underset{\sim}{x} \le Q = \underset{\sim}{b}^T A^{-1}\underset{\sim}{b}$

seek $\max\ q' = 2\underset{\sim}{x}^T\underset{\sim}{b} - \underset{\sim}{x}^T B\underset{\sim}{x} \le Q' = \underset{\sim}{b}^T B^{-1}\underset{\sim}{b}$, $B = \alpha I + A$

Base vectors $\hat{\underset{\sim}{q}}$ don't change, because subspace is same.

Scalars γ and δ do change, but can be simulated "free".

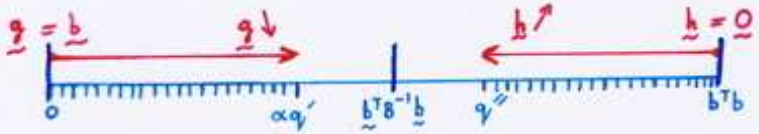
Get γ' and δ', hence $\underset{\sim}{x}'$ and q', for free.

Can also get for free

$\max\ q'' = 2\underset{\sim}{x}^T A\underset{\sim}{b} - \underset{\sim}{x}^T AB\underset{\sim}{x} \le Q'' = \underset{\sim}{b}^T AB^{-1}\underset{\sim}{b}$

But $\alpha Q' + Q'' = \underset{\sim}{b}^T\underset{\sim}{b}$ = known, so

$\alpha q''$ gives upper limit on $\underset{\sim}{b}^T B^{-1}\underset{\sim}{b}$.



Termination : Stop when 99 % of $\underset{\sim}{b}^T\underset{\sim}{b}$ accounted.

Extensions

Get controlled evaluation of $(\alpha I + A)^{-1}$ on any $\underset{\sim}{b}$

Also get $(\beta I + A)^{-1}$, even better controlled if $\beta > \alpha$.

Superpose : get $(\beta_1 I + A)^{-1} + (\beta_2 I + A)^{-1}$

Generalise : get $f(A) = \int_\alpha^\infty d\beta\, W(\beta)\,(\beta I + A)^{-1}$, $W \ge 0$

for "simulatable" function $\boxed{f(t) = \int_\alpha^\infty \frac{W(\beta)\,d\beta}{\beta + t}}$

Example : $W(\beta) = (\beta - \alpha)^{-k}$, $0 < k < 1$, gives $f(t) = (\alpha + t)^{-k}$

In particular, $k = \frac{1}{2}$ gives $\boxed{\underset{\sim}{s} = (\alpha I + A)^{-\frac{1}{2}}\underset{\sim}{r}}$

With r_i unit normal, covariance $\langle \underset{\sim}{s}\,\underset{\sim}{s}^T \rangle = (\alpha I + A)^{-1}$.

Example : $k \to 0$ (with offset) gives $f(t) = -\log(\alpha + t)$

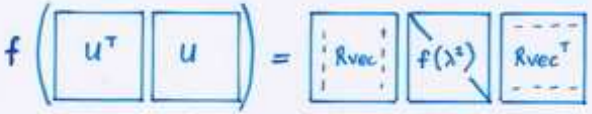
In particular, $\underset{\sim}{r}^T \log(kI + A)\,\underset{\sim}{r}$ with r_i unit normal

$= \sum r_i^2 \log(\alpha + \lambda_i)$ in eigencoords.

with expectation $\sum \log(\alpha + \lambda_i) = \boxed{\log\det(\alpha I + A)}$

To evaluate, diagonalise U to

so that



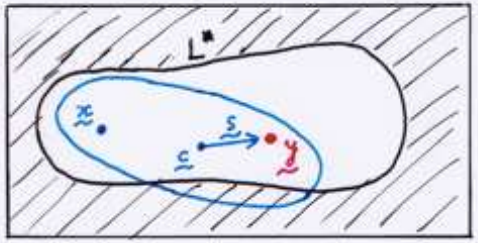
Upper and lower limits remain available.

NESTED SAMPLING

Space of $\underset{\sim}{x}$ has flat prior and convex

$$L(\underset{\sim}{x}) = \log \Pr(\text{data} \mid \underset{\sim}{x})$$

Current position is $\underset{\sim}{x}$ with $L(\underset{\sim}{x}) = L_0 > L^*$.



Seek new random point $\underset{\sim}{y}$ within $L(\underset{\sim}{y}) > L^*$.

Proposal : Build local quadratic model, regularised.

$$\tilde{L}(\underset{\sim}{y}) = L_0 + (\underset{\sim}{y}-\underset{\sim}{x})^T \underset{\sim}{g} - \tfrac{1}{2}(\underset{\sim}{y}-\underset{\sim}{x})^T B (\underset{\sim}{y}-\underset{\sim}{x})$$

where $\underset{\sim}{g} = \nabla L$, $B = \alpha I + \nabla\nabla L$ at $\underset{\sim}{x}$.

Centre of model is

$$\underset{\sim}{c} = \underset{\sim}{x} + B^{-1}\underset{\sim}{g} \quad \text{with} \quad L_c = L_0 + \tfrac{1}{2}\underset{\sim}{g}^T B^{-1}\underset{\sim}{g}$$

Propose $\underset{\sim}{y} = \underset{\sim}{c} + \underset{\sim}{s}$

within ellipsoid $L(\underset{\sim}{y}) = L_c + \tfrac{1}{2}\underset{\sim}{s}^T B \underset{\sim}{s} > L^*$.

Offset $\underset{\sim}{s}$ constrained within ellipsoid

$$\underset{\sim}{s}^T B \underset{\sim}{s} < 2(L_c - L^*) \equiv a^2$$

Random point on surface is $a B^{-\frac{1}{2}} \underset{\sim}{r}$

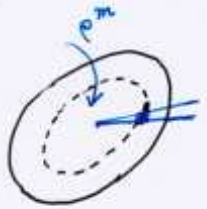
where $\underset{\sim}{r}$ is random on sphere $\underset{\sim}{r}^T\underset{\sim}{r} = 1$.

Random point within volume is

$$\underset{\sim}{s} = \rho \cdot a B^{-\frac{1}{2}} \underset{\sim}{r}$$

where $\rho^m = \text{Uniform}(0,1)$

(m = dimension).



Propose $\boxed{\underset{\sim}{y} = \underset{\sim}{x} + B^{-1}\underset{\sim}{g} + \rho a B^{-\frac{1}{2}}\underset{\sim}{r}}$

$$B = \alpha I + A$$

$$a = \sqrt{2(L_0 + \tfrac{1}{2}\underset{\sim}{g}^T B^{-1}\underset{\sim}{g} - L^*)}$$

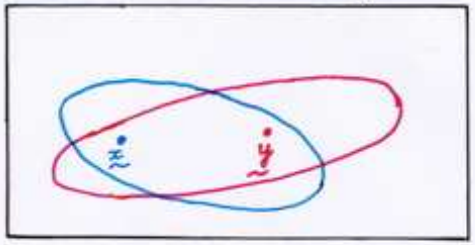
Volume of destination ellipsoid

$$V = \frac{\pi^{m/2}}{(\frac{m}{2})!}\frac{a^m}{\sqrt{\det B}}$$

Acceptance ?

If $L(\underset{\sim}{y})$ outside L^*, reject.

If OK, build new model around $\underset{\sim}{y}$.



If original $\underset{\sim}{x}$ outside new ellipsoid, reject.

If OK, accept in ratio of volumes (detailed balance).

But $V = \exp(m \times \ldots)$, so volume ratio not O(1).

Fixup: Adjust α to maintain

$$\frac{\log V}{m} = \log a - \frac{1}{2m} \log\det B = \log a - \tfrac{1}{2}\underset{\sim}{r}^T \log B\, \underset{\sim}{r}$$

= typical length in model.

Exploration loop:

Propose $\underset{\sim}{y} = \underset{\sim}{x} + B^{-1}\underset{\sim}{g} + \rho\alpha B^{-\frac{1}{2}}\underset{\sim}{r}$ at current α.

If $L(\underset{\sim}{y})$ outside L^*, reject.

If OK, build new ellipsoid of same V (adjust α).

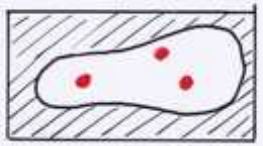
If new ellipsoid does not contain $\underset{\sim}{x}$, reject.

If OK, accept $\underset{\sim}{y}$.

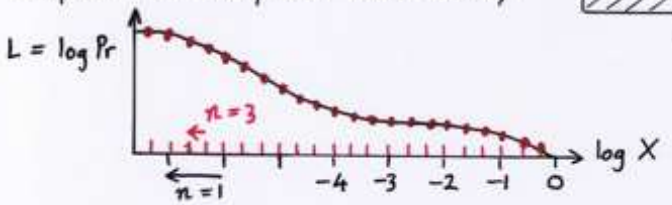
Oddity: Need to impose dimension m in $\rho = \text{Uniform}^{1/m}$.

Multiple samples

Nested sampling can use several samples to compress more slowly



e.g. $n = 3$.



Samples can evolve separately — or together in an ensemble with

$$\nabla L = \begin{pmatrix} \underset{\sim}{g} \\ \underset{\sim}{g} \\ \underset{\sim}{g} \end{pmatrix} \qquad \nabla\nabla L = \begin{pmatrix} A & & \circ \\ & A & \\ \circ & & A \end{pmatrix}$$

Random detail apart, the α's stay as before, and algorithm proceeds as before, in bigger space.

Each individual sample evolves as before, except

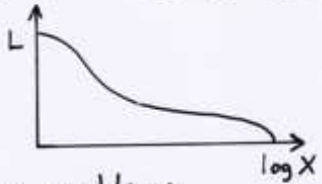
$\rho = \text{Uniform}^{1/3m}$

Sampling is biassed outward, giving slower compression.

SUMMARY

Nested sampling was designed for evidence value.

Actually, it plots $L(X)$.



Evidence and posterior are by-products.

Can cope with multi-phase problems.

Exploration need not rely on guessed directions, mostly orthogonal to useful directions.

Can use gradients and curvatures.

Control by volume, not point-wise values.

Compression can be arbitrarily slow (accuracy) or arbitrarily fast (speed).

High-dimensional problems need not be slow.

CPU $\propto$ size

J.S. Sept 2008