"Partitioning the likelihood" †

SOMA DHAVALA
"Dept. of Statistics"
Texas A & M University

to be presented Stat 613 class

† based on "partitioned algorithms for maximum likelihood other non-linear estimation" by Gordon Smyth, Statistics & Computing, 6, 201-216, 1996

Scope & Context of the talk:

$f_\theta(x)$ ← Model formulation (0/3)

present talk ← compute ML estimates

← Analyze bias / variance / robustness / consistency / Model selection (0/3)

Maximum likelihood estimation

$y_i \overset{iid}{\sim} p(\theta)$: θ: parameter vector of dimension "p"

y: observed data

p: a probability density function parametrized by θ

$$L(\theta) = \prod_{i=1}^{n} p_\theta(y_i) \quad \text{(likelihood)}$$

$$l(\theta) = \log L(\theta) \quad \text{log likelihood}$$

$$\boxed{\hat{\theta}_{mle} = \max_\theta l(\theta)}$$

So is MLE a turn-the-crank procedure?!

$$\hat{\theta}_{mle} = \max_{\theta} \ l(\theta)$$

is just the beginning of the harsh reality

Explore partitioning as a means to efficiently solve an otherwise difficult optimization problem

Partitioning space : data, parameter, objective function

ways : inferential
(eg : nuisance parameter)

: computational
(uncorrelated parameters / orthogonal parameters)

:

essentially an optimization problem

→ stability

→ speed of convergance

→ accuracy

Parameter space partion:

- reduced objective function
- nested iterations
- zig-zag iterations
- leap-frog iterations

Data-partion:

- Distributed ML estimation
  - EM
  - Sub-object gradient methods

- reduced objective function

$$r(\theta_2) = l\left(\hat{\theta}_1(\theta_2), \theta_2\right)$$

→ dimensionality reduction

→ possible when $\hat{\theta}_1(\theta_2)$ is in closed-form

→ converges faster than full objective function (curse-of-dimensionality)

* Profile likelihood when θ_1 is a nuisance parameter

$$\theta_1 = F_1(\theta_1, \theta_2)$$
$$\theta_2 = F_2(\theta_1, \theta_2) \rightarrow (1)$$

Nested algorithms:

$$\theta_2^{k+1} = F_2\left(\hat{\theta}_1(\theta_2^k), \theta_2^k\right)$$

F_2 does involve θ_1 and θ_2 unlike reduced objective function

Zig- Zag :

$$\theta_1^{k+1} = \hat{\theta}_1(\theta_2^{k})$$

$$\theta_2^{k+1} = \hat{\theta}_2(\theta_1^{k+1})$$

eg: $\mu_i = \beta_1 + \beta_2 x_i$ and $\sigma_i = \exp(\gamma_1 + z\gamma_2)$

both $\underline{\beta}$ and $\underline{\gamma}$ are estimated regressing given others.

Leap-frog :

$$\theta_1^{k+1} = F_1(\theta_1^k, \theta_2^k)$$

$$\theta_2^{k+1} = F_2(\theta_1^{k+1}, \theta_2^k)$$

F_1 & F_2 are updation equations

not estimates

Properties :

1) Reduced objective function: faster, stable

2) Nested: can not be worse than full-iterations

3) Zig-zag: stable but slow, faster if parameters are uncorrelated

4) leap-frog: less guaranteed, similar to zig-zag if parameters are orthogonal

Scoring Method :

$$\theta^{k+1} = \theta^{k} - \ddot{l}^{-1}(\theta^{k})\, \dot{l}(\theta^{k})$$

$\ddot{l}^{-1}(\theta^{k}) \downarrow$ observed information matrix

$\dot{l}(\theta^{k}) \downarrow$ score

$$\theta^{k+1} = \theta^{k} - I^{-1}(\theta^{k})\, \dot{l}(\theta^{k})$$

$I^{-1} \downarrow$ expected IM

Partitioned (θ_1, θ_2)

$$\theta_2^{k+1} = \theta_2^{k} - \ddot{l}_{2,1}^{-1}\, \dot{l}_2$$

Contd..

where $\ddot{l} = \begin{pmatrix} \ddot{l}_{11} & \ddot{l}_{12} \\ \ddot{l}_{21} & \ddot{l}_{22} \end{pmatrix}$ and

$$\ddot{l}_{2\cdot1} = \ddot{l}_{22} - \ddot{l}_{21}\,\ddot{l}_{11}^{-1}\,\ddot{l}_{12}$$

Newton Raphson of profile likelihood

Nested Fisher-scoring method

$$A^{-1} = \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{2\cdot1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix}$$

$$A_{2\cdot1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

then

$$\theta_2^{k+1} = \theta_2^{k} + A_{2,1}^{-1}\, \dot{l}_{2,1} \longrightarrow (3)$$

however $\dot{l}_{2,1} = \dot{l}_2$ at $\theta_1 = \hat{\theta}_1(\theta_2)$

since $\dot{l}_{2,1} = \dot{l}_2 - A_{21} A_{11}^{-1} \underbrace{\dot{l}_1}_{\downarrow\, 0}$

at $\theta_1 = \hat{\theta}_1(\theta_2)$

(3) evaluated at

$\theta_1^{k} = \hat{\theta}_1(\theta_2^{k})$

$\theta_2 = \theta_2^{k}$

Guass- Newton :

$$\theta^{k+1} = \theta^{k} + (\dot{\mu}^T \mu)\, \dot{\mu}^T (y - \mu)$$

where $E(y) = \mu$

Partitioned Gauss- Newton

$$\theta_2^{k+1} = \theta_2^{k} + \left[\dot{\mu}_2^T (I - P)\, \dot{\mu}_2\right]^{-1} \dot{\mu}_2^T (y - \mu)$$

$$\theta_1 = \hat{\theta}_1(\theta_2)$$

If θ_1 ~~are~~ is linear, then, the above is called seperable least-squares ...

$\rightarrow$ if $\mu = X(\theta_2)\,\theta_1$, then

$$\hat{\theta}_1 = (X^T X)^{-1} X^T y \quad \text{and}$$

$$\dot{\mu}_1 = X$$

$$P = \dot{\mu}\left(\dot{\mu}_i^T \dot{\mu}\right)^{-1} \dot{\mu}^T$$

projection matrix of $\dot{\mu}$

→ Separable least-squares is clearly a Fisher-scoring if nested

$$y \sim N(x(\theta_2)\theta_1, \sigma^2 I)$$

since if $y \sim N(\mu(\theta), C(\theta))$

$$I(\theta) = \dot{\mu}^T C^{-1} \dot{\mu} + \frac{1}{2} tr\left[\left(C^{-1}(\theta)\frac{\partial C}{\partial \theta}\right)^2\right]$$

stability: ill-conditioned matrices

$$\theta^{k+1} = \theta^{k} + A^{-1} S$$

① A is ill-conditioned

then $A_{u} = (\alpha I + A)$

α : small +ve constant

②

$$\theta^{k+1} = \theta^{k} + \underline{\alpha} \cdot A^{-1} S$$

α : damping factor

Partitioning in EM algorithm:

standard EM: in I_c is over-informative

rate-of-convergence : $\propto$ largest eigen vector of $I_c^{-1} I_m$

I_m: missing information matrix

observed incomplete IM

$$I_c = I_{IC} + I_m$$

choose "missing" information s.t

I_c is as small as possible.

trick: partition:

$\theta \rightarrow (\theta_1, \theta_2)$

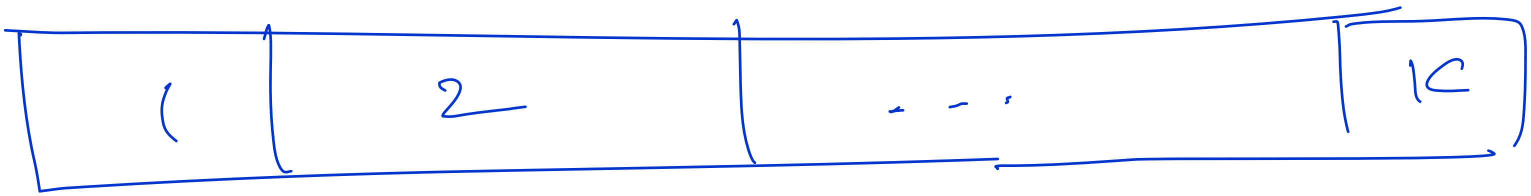
$\theta_1 \rightarrow$ seperate hidden-space

$\theta_2 \rightarrow$ seperate hidden-space

SAGE

space-alternating generalized EM algorithm

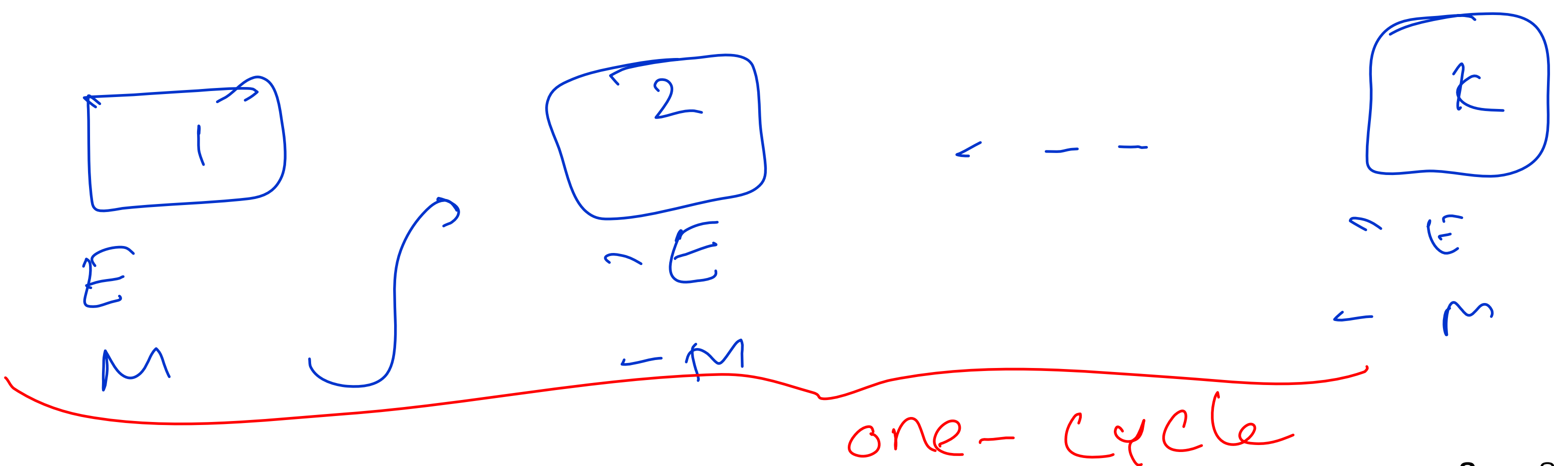
Partition the data



k - blocks of data:

Partial E - step

Partial M - step



useful in distributed MC estimation

x_1 θ or t_1 $\rightarrow$ x_2 or t_2 $\rightarrow$ x_3 or t_3 ...

$$f(x) = \sum_i f_i(x)$$

sub-objective functions

$$\theta^{k+1} = \theta^{k} - \underbrace{\alpha^{k}}_{\text{step-size}} \cdot \underbrace{\nabla f_i(\theta^{K})}_{\text{gradient}}$$

References : $\boxed{\text{Google these key words}}$

1) Krishnan & McCullan: "EM algorithm & its extensions" wiley & sons

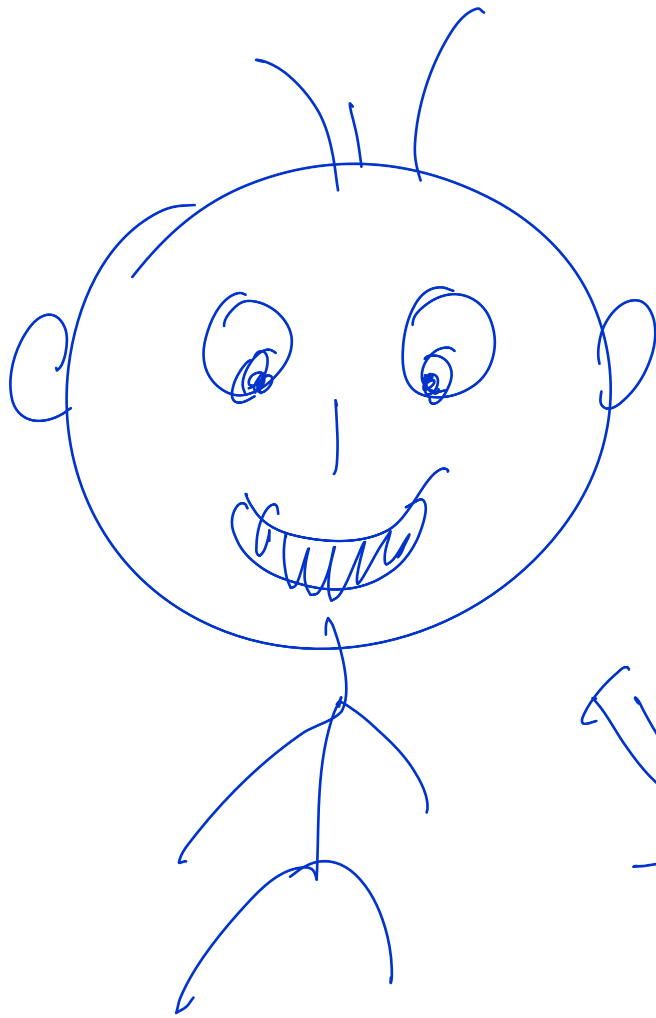
2) Fessler and A.O. Hero, "space alternating generalized EM algoritm" SAGE

3) Neal and Hilton, " A view of EM algorithm that justifies incremental view point" Incremental EM algoritm

4) Dimitris (stanford university), " Incremental LMS/ steepest descent algoritmns", distributed optimization

5) Robert Nowak, " Distributed EM algoritm

6) Golube, " Review of variable projection seperable least-squares



Thank you

– Soma