10-708

10/06/2020

L11 - Variational inference - loopy belief prop.

(*) continue L10 - derivation of KF for SSMs

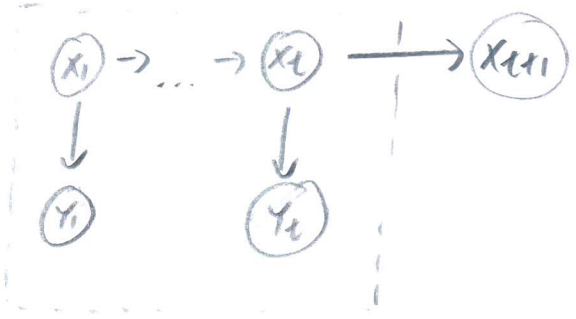
(**): solidify analogy/mental model to crystallise → true understanding.

Kalman filtering derivation

- recall filtering problem
- inference of $p(x_t|y_{1:t})$ (i.e. cond. distn of last hidden state given all observations up to now)
- use diagram as reference point for visualising queries.
- we require:-

  $p(x_t|y_{1:t})$ from a recursive procedure involving $p(x_{t+1}|y_{1:t+1})$, y_{t+1}

- Break up problem into 2 steps:-



1) time update:

- compute $p(x_{t+1}|y_{1:t})$

  from prior belief $p(x_t|y_{1:t})$ and

  dyn. model $p(x_{t+1}|x_t)$

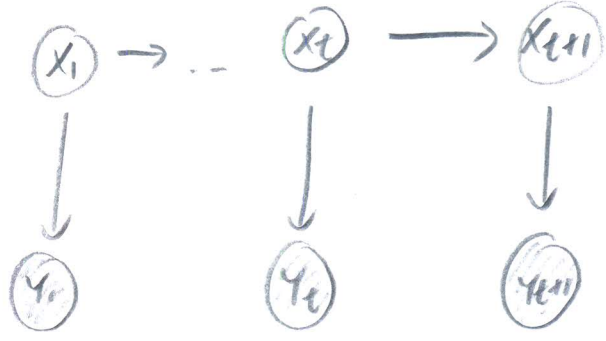
2) measurement update:

- compute new belief $p(x_{t+1}|y_{1:t+1})$

  from prediction $p(x_{t+1}|y_{1:t})$,

  observation y_{t+1}

  obs. model $p(y_{t+1}|x_{t+1})$



(*) Make use of Gaussian properties everywhere

- inference problem → dealing with new means and covariances

(**): same technique as in F.A.

(*) Schematic of the strategy:-

M.I.L.

(Gaussian manip.)

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$z_1 \rightarrow z_1 z_2 \longrightarrow z_1 | z_2$

$\uparrow$

$z_2 | z_1 \quad (\mu_1, \mu_2)$ (*) $\qquad \mu_2 | \mu_1$ (*)

corresp.

$z_1 \sim x_t | y_{1:t} \qquad x_{t+1} = Ax_t + Gw_t = f(x_t) \longrightarrow p(x_{t+1}, y_{t+1} | y_{1:t}) \longrightarrow p(x_{t+1} | y_{1:t}, y_{t+1})$

$\uparrow$

$p(y_{t+1} | x_{t+1})$

$\uparrow$

$y_{t+1} = Cx_{t+1} + v_{t+1}$

$p(x_{t+1} | y_{1:t})$

(*) μ_1

(A1) - tidy this up (makes some sense, but want full clarity)

- Ex: conditioning is not scary; just some constants in eq.; focus on LHS of cond.

(*) Compute means and covariances within the schematic (A2) review

---

(*) Predict step

- dynamical model $x_{t+1} = Ax_t + Gw_t \quad w_t \sim N(0, Q)$

$$\hat{x}_{t+1|t} = \mathbb{E}[x_{t+1} | y_1, \ldots, y_t] = A\hat{x}_{t|t}$$

$$P_{t+1|t} = \mathbb{E}\left[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_1, \ldots, y_t \right]$$

$$= \mathbb{E}\left[(Ax_t + Gw_t - \hat{x}_{t+1|t})(Ax_t + Gw_t - \hat{x}_{t+1|t})^T | y_1, \ldots, y_t \right]$$

(A1)

$$= AP_{t|t}A + GQG^T$$

- $\hat{x}_{t+1|t}$ and $P_{t+1|t}$ are conditional means and covariances (*)

(A3) - reflect on

observation model $y_{t+1} = C x_{t+1} + v_t \quad v_t \sim N(0, R)$

$$\mathbb{E}[y_{t+1} | y_1, \dots, y_t] = \mathbb{E}[C x_{t+1} + v_{t+1} | y_1, \dots, y_t] = C \hat{x}_{t+1|t}$$

$$\mathbb{E}[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})^T | y_1, \dots, y_t] = C P_{t+1|t} C^T + R$$

(*) update step

- From previous slide:-

we have joint $p(x_{t+1}, y_{t+1} | y_{1:t}) = N(m_{t+1}, V_{t+1})$ where:-

$$m_{t+1} = \begin{pmatrix} \hat{x}_{t+1|t} \\ C \hat{x}_{t+1|t} \end{pmatrix} \qquad V_{t+1} = \begin{pmatrix} P_{t+1|t} & P_{t+1|t} C^T \\ C P_{t+1|t} & C P_{t+1|t} C^T + R \end{pmatrix}$$

- Now use conditional Gaussian formulae for partitioned matrices to get $p(x_{t+1} | y_{t+1}, y_{1:t})$

(*) Kalman Filter

- Yields closed form measurement and time updates:-

measurement update :- $\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C \hat{x}_{t+1|t})$

$$P_{t+1|t+1} = P_{t+1|t} - K C P_{t+1|t} \qquad (?)$$

- K_{t+1} - Kalman gain matrix

Time updates: $\hat{x}_{t+1|t} = A \hat{x}_{t|t}$

$$P_{t+1|t} = A P_{t|t} A + G Q G^T$$

$K_{t+1} = P_{t+1|t} C^T (C P_{t+1|t} C^T + R)^{-1}$

(A3) - Review notation and intuition

Ex. Be sensitive to conditioning in terms of info being used at each point.

(*) Example of KF in 1D

- Just apply KF equations

(PQ) - intuition behind these

(*) KF intuition (ON)

- KF update of mean:- $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - C\hat{x}_{t|t-1})$ - innov.

$$= \frac{(\sigma_z + \sigma_x) z_t + \sigma_z \hat{x}_{t|t-1}}{\sigma_z + \sigma_x + \sigma_z}$$

(ON) new belief is convex combo of updates from prior and observation, weighted by Kalman gain matrix

$$K_t = P_{t|t-1} C^T (C P_{t|t-1} C^T + R)^{-1}$$

- observation unreliable, σ_z (i.e. R) large so K_t is small → more att. to prediction

- old prior is unreliable, large σ_x, or process is unpredictable, large σ_x, we pay more attention to observation.

(Q): where do A, G, C matrices come from?

- A parameter estimation (as opposed to filtering, which is an inference problem)
  - inference problem: i.e. compute p.d. of latent var, given data in this context.
- inference assumes learning/ param est. has already occurred.

(*) Complexity of one KF step., Inf prob 2, RTS smoother, RTS deriv.

- Review not in our time

(*) Learning SSMs

(*) use EM. Inference from KF or RTS filtering forms E-step!

(*) Form complete log-like.

(*) E-step → compute suffi. statistics using inference results to give you a posterior

M-step → MLE

EX: You can & should derive in your spare time.

(*) Key principle → inference can be treated as a subroutine for learning in a partially obs. setting

(*) Nonlinear systems

- use derivatives, Taylor exp. approxim. / linearization.

(*) C.S thinking → you can approximate functions to the order you can afford

(*) EX A lot of seemingly difficult, obtuse algebra can be viewed/reduced through an understanding of principles motivating them (often recur).

10-708 10/06/2020

11 - Approximate Inference and Topic Models (mean field and loopy BP) S.2019 11b

- (*) Looks like lecture 11a. VI and loopy B.P. not covered
- (*) Approximate inference methods started as 'tricks', ones which worked then proved.

EX: Maintain 'art of modelling' trajectory, present an example that exposes the concrete need for approx. inference

---

(*) Probabilistic Topic Models

- started as a class project.

(*) How to get started for a new modelling task?

(BM): There is gold here: a way of thinking about the art of modelling

EX: Start with a concrete task/problem you want to solve.
- methods are invoked in service of the task!

e.g. Bird's eye view of 1 million documents.

- Task → clustering; give back cluster label.
  - ↳ embedding → visualisation of cluste labels
- Representation of data → e.g. continuous, binary, counts? (design choice)

(*) Inside each element one at a time

(*) Tasks - document embedding.
- How each document embedded in a space.

(*) Summarising data using topics
- Want the embedding to be meaningful.

(*) See how data changes over time
- Representation of topics may evolve.

(#) user interest topic modelling
- secondary task

(#) Representation
- Bag of words rep.
- convert article → BOW.

(*) For each document; count no. of words, order does not matter.

(*) Each document is a vector in word-space

Ex: Benefit and costs? (of representation)
- Benefits → storage, hash tables (makes data simple)
- comparing documents of different lengths (1 x 100 word vs novel
- Probability based on word-orderings → longe docs give smaller likelihood (due to product)
  - cannot compare.
- allows comparability.
- costs → ordering may be important for semantics

(*) BOW is a baseline representation.

(#) How to model semantics?
- brilliant using orderings

Ex: mental analogy → certain keywords give a niche prob. of a document coming from a topic.
- match a topic with keywords @ intuition.

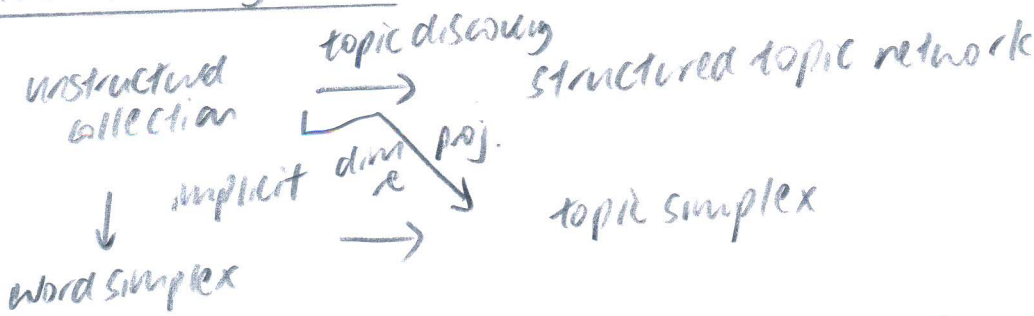
(*) A topic is a vector of words (w given all vocab.)
A document contains topics in a p (mixing proportion)

(*) ~~#40~~ Note info. compression: a document gets compressed into a weighted sum of topics (low-dimensional with some semantic meaning)

(*) Can then compare similarity of documents

actually a vector of probabilities indicating topic mixing propor

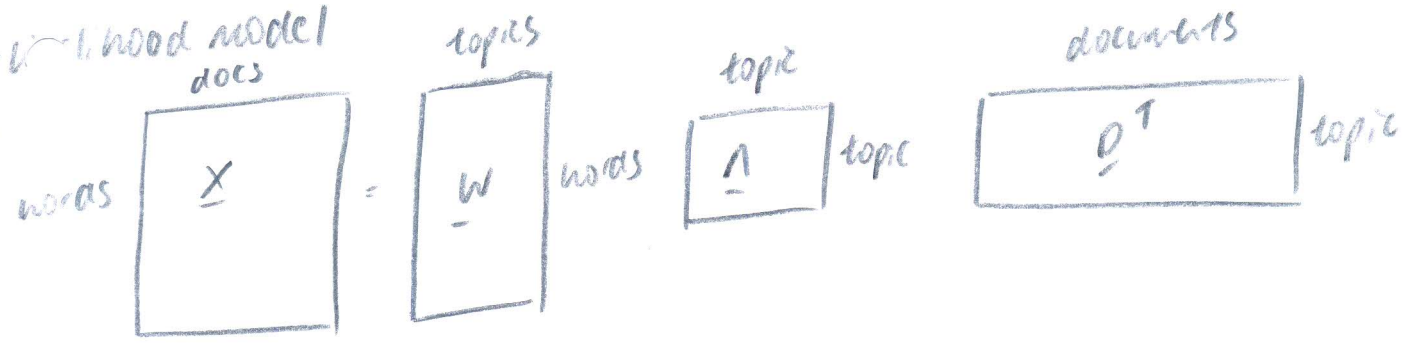
(*) Topic Models - Big Picture

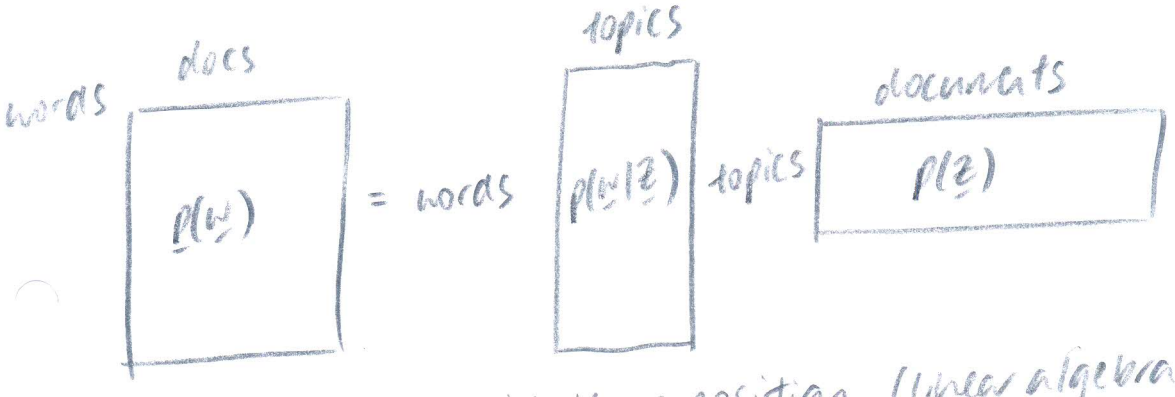


(*) A topic corresponds to a point on the word simplex.
(*) A document —"— on the topic simplex

(*) LSI vs Topic Model (probLSI)

- likelihood model





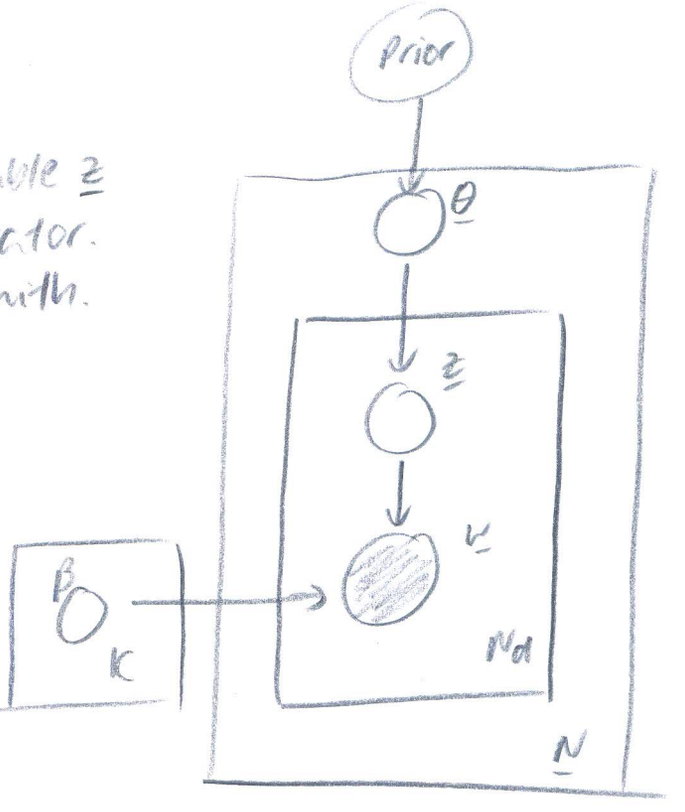
- LSI → word-document matrix decomposition (linear algebraic)
- topic models → conceptually similar (probabilistic inference)
- matrix decomposition techniques (e.g LSI) make it difficult to bypass computationally expensive matrix inversions (especially for large matrices) (batch)
(*) Probabilistic methods allow ways of dealing with this issue in a piece-wise iterative manner (e.g. LASSO vs batch LR)
i.e. local operations to

(*) Admixture Models

- skip

(*) Topic Models

- generative process for a document.
  - for every word there is a latent variable $\underline{z}$ indicating its topic assignment/indicator.
  - i.e. what topic a word is affiliated with.
  - topic indicator $\underline{z}$ comes from a weight vector $\underline{\theta}$
  - every document is a vector of topical weights
  - prior: distn of different topics in corpus.



generating: $\underline{\theta}$

1) draw a topical weight from a prior distribution vector/document.

- Given this, n^{th} out of for every N_d words:
  - draw a topic indicator $z_n \sim$ multinomial $(\underline{\theta})$
  - conditioning on the topic indicator of the n^{th} word z_n, sample the word (using a coll. of word-frequency distns)
    - draw $w_n | z_n, \{\beta_k\}$ from multinom (β_{z_n})

(N): Presentation is not the clearest

(*) Choices of priors

- Dirichlet (LDA), Blei et al. 2003
- Logistic normal, Blei & Lafferty (200x), Ahmed & Xing (2006).

(see the facets of these and modelling intuition)

(*) Generative semantics of LDA/TAM

- captures intuition of topics being highly correlated
e.g. sports and politics
- use a covariance matrix?
  - only in MVGs (not simplest).
- OK, so use MVGs ⟶ to yield $\underline{\chi}$

ISSUE: we cannot use $\underline{\chi}$ to sample $\underline{z}$ → $\underline{\chi}$ is Gaussian vector with -ve components; NOT multinomial

- Apply a transformation, apply exponential, then normalise to get appropriate prior.
  $\underline{\chi} \to \underline{\theta}$ (logistic normal)

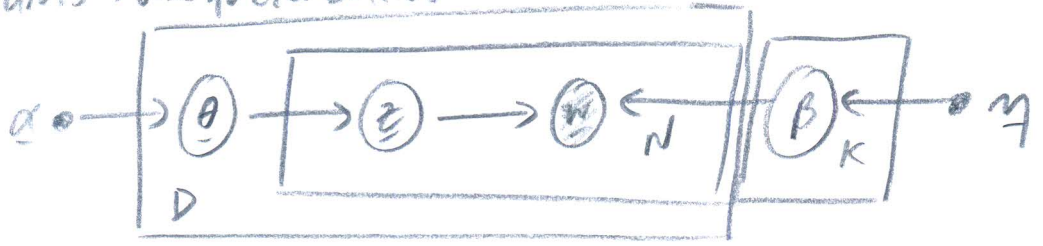
Ex: Model design can be arbitrary, flexible, to a degree.

(*) Posterior inference results

- Familiar → recap (A1) - review.
- (A2) see allowed slides

(*) Joint likelihood (A2) - check dim.

- Plates → use factorisation law!



$$p(\underline{\beta}, \underline{\theta}, \underline{z}, \underline{w}) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn}, \beta)$$

(*) we want posterior of any of latent variables given observed words $\underline{w}$ and likelihood of the word.

(*) Inference and learning both intractable.

(AB) - Review the intractability to get a real sense of how powerful approx inference.

e.g. $p(\theta_n|\underline{D})$ and $p(\underline{D})$

- No known technique for performing these techniques exactly

(*) Approximate Inference

i) variational inference
- turns solution of an inference problem → sol. of an optimisation problem.

ii) MCMC

(*) variational inference (cont)

(AH) - review logic

- Consider generative model $p_\theta(\underline{x}|\underline{z})$, prior $p(\underline{z})$
- Joint distri:- $p_\theta(\underline{x},\underline{z}) = p_\theta(\underline{x}|\underline{z})p(\underline{z})$
- Assume variational distri $q_\phi(\underline{z}|\underline{x})$
- objective: maximise lower bound for log-likelihood:-

$$\log p(\underline{x}) = KL(q_\phi(\underline{z}|\underline{x}) \| p_\theta(\underline{z}|\underline{x})) + \int_{\underline{z}} q_\phi(\underline{z}|\underline{x}) \log \frac{p_\theta(\underline{x},\underline{z})}{q_\phi(\underline{z}|\underline{x})} \, d\underline{z} \quad ?$$

$$\geq \int_{\underline{z}} q_\phi(\underline{z}|\underline{x}) \log \frac{p_\theta(\underline{x},\underline{z})}{q_\phi(\underline{z}|\underline{x})}$$

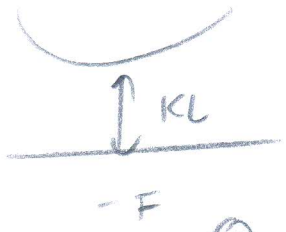
$$:= \mathcal{L}(\theta;\phi;\underline{x})$$

Equivalently; minimise free-energy (upper bound on log-like.)
(surrogate on target)

$$F(\theta;\phi;\underline{x}) = -\log p(\underline{x}) + KL(q_\phi(\underline{z}|\underline{x}) \| p_\theta(\underline{z}|\underline{x}))$$

(*) Distance between free energy and ... (-F)

(*) $q_\phi(\underline{z}|\underline{x}) = p_\theta(\underline{z}|\underline{x}) \Rightarrow KL(q||p) = 0$



(*) Variational inference

(AS): review

(*) Maximise variational lower bound:-

$$\mathcal{L}(\theta;\phi;\underline{x}) = \mathbb{E}_{q_\phi(\underline{z}|\underline{x})}\left[\log p_\theta(\underline{x}|\underline{z})\right] + KL\left(q_\phi(\underline{z}|\underline{x}) \| p(\underline{z})\right)$$

$$= \log p(\underline{x}) - KL\left(q_\phi(\underline{z}|\underline{x}) \| p_\theta(\underline{z}|\underline{x})\right)$$

(W): A little abstract.

(*) E-step: maximise $\mathcal{L}$ wrt ϕ with θ fixed:-

$$\max_\phi \mathcal{L}(\theta;\phi;\underline{x})$$

If closed form sol. exist:-

$$q_\phi^*(\underline{z}|\underline{x}) \propto \exp\left[\log p_\theta(\underline{x},\underline{z})\right]$$ (do not set this to $\neq p(\underline{z}|\underline{x})$ make V.I.)

(*) M-step: maximise $\mathcal{L}$ wrt θ; with ϕ fixed.

$$\max_\theta \mathcal{L}(\theta;\phi;\underline{x})$$

Ex use $q_\phi(\underline{z}|\underline{x})$ i.e. an inference step on $\underline{z}$ (latent) given data $\underline{x}$

- make sure q_ϕ is 'good' in sense of being 'close' to $p_\theta(\underline{z}|\underline{x})$ using KL-divergence as a measure of closeness
- Why not just set $q_\theta(\underline{z}|\underline{x}) = p_\theta(\underline{z}|\underline{x})$? → p_θ intractable

---

(*) Mean-field assumption (in topic models)

- True posterior:- $p(\beta,\theta,\underline{z}|\underline{w}) = \dfrac{p(\beta,\theta,\underline{z},\underline{w})}{p(\underline{w})}$

(*) Break dependency using fully factorised distri:-

$$q(\beta, \theta, \underset{\sim}{z}) = \prod_{k} q(\beta_k) \prod_{d} q(\theta_d) \prod_{n} q(z_{dn})$$

- Product of non-coupled, individual distri (marginals)

EX: Why does this make it things easier?

W:- Can optimise individually by breaking into subproblems

(A6): Review - not entirely sure of i) logic
ii) dimensionality

(*) Mean-Field Approx

(A7): Pick up during review. (EX gone through quickly)

- Read/glance/skim papers

(*) Co-ordinate ascent algorithm for LDA

- Get an iterative program.
- (A8): Review pseudocode.

EX: A lot of material to digest

- Next lectures ⟶ more examples of approx. inference
- Give a 'grand theory' to unify/better understood.