Lesson 4 : Information geometry

Intro:

- last session on Information Theory: Fisher information
- fundamental applications to many domains:
  - optimization
  - modeling: priors on parameters
  - MDL: precision/encoding-cost of a parameter?
- end of the lesson: fast bandits, needed for projects

I - Fisher information

- Fisher information/metric/matrix
  - setting: Model(θ) −> $p_\theta(x)$
  - formula 1 : $J(\theta) = E_{x\sim p_\theta}[d^2(-\log p_\theta(x))/d\theta^2]$
  - average second derivative of the encoding cost, over all possible data under the model law
  - formula 2 : $E_{x\sim p_\theta}[d \log p_\theta(x)/d\theta^T d \log p_\theta(x)/d\theta]$
    - covariance of the gradient of the cost, under the model law
    - obtained by developping:
      - $d \log p/d\theta = 1/p\ dp/d\theta$
      - $d^2 \log p/d\theta^2 = -1/p^2\ dp/d\theta\ dp/d\theta + 1/p\ d^2p/d\theta^2$
      - $J(\theta) = \int_x [-...] = E[(d \log p/dt)^2] + 0$ as $\int_x d^2p/d\theta^2 dx = d^2/d\theta^2 \int_x p dx = 0$

Why information "geometry" ?

- norms in the space of probabilities (variations)
- examples of observations: $x1, x2 \ldots$
  or $(x1, y1), (x2, y2) \ldots$ to learn a function $y = f(x)$
  $\Longrightarrow$ finding the right θ = learning the function
- second order of $KL(p_{\theta+\delta\theta}||p_\theta) = \int_x p_{\theta+\delta\theta}(x) \log p_{\theta+\delta\theta}(x)/p_\theta(x)dx$
  - use $ln(1 + z) = z - 1/2z^2 + O(z3)$
  - use $dlnf(z)/dz = 1/z\ df/dz$
  - use $\int_x p_\theta(x) = 1$(for any θ) $\Longrightarrow$ $\int_x d^k p_\theta(x)/d\theta^k = 0$
  - From now on, all p mean $p_\theta(x)$

- $p_{\theta+d\theta}(x) = p(x) + \delta\theta dp/d\theta + 1/2\delta\theta\ d^2p/d\theta^2\ \delta\theta + O(\delta\theta^3)$
- $p_{\theta+d\theta}(x)/p_\theta(x) = 1 + \delta\theta\ 1/p\ dp/d\theta + 1/2\delta\theta\ 1/p\ d^2p/d\theta^2\ \delta\theta + O(\delta\theta^3)$
- $\log(p_{\theta+d\theta}(x)/p_\theta(x)) = \delta\theta\ 1/p\ dp/d\theta + 1/2\delta\theta\ 1/p\ d^2p/d\theta^2\ \delta\theta - 1/2\delta\theta\ 1/p^2\ dp/d\theta\ dp/d\theta^T\ \delta\theta + O(\delta\theta^3)$
- $KL(|) = E_{x\sim p_{\theta+\delta\theta}}[...] = E_{x\sim p}[...] + \delta\theta \int dp/d\theta[...] + ...$
  $= 0 + 0 - 1/2\ \delta\theta \int 1/p\ dp/d\theta\ dp/d\theta^2\ \delta\theta + 1 * idem + O(\delta\theta^3)$
  $= 1/2\ \delta\theta E_{x\sim p}[d\log p/d\theta\ d\log p/d\theta]\delta\theta + O(\delta\theta^3)$

–> metric associated to KL
–> curvature of relative entropy

// + differential entropy <–> Fisher information [Cover&Thomas p 672]

- In practice:
  - Fisher: $\int_{x\sim p_\theta}[...] \sim 1/n \sum_i [...]$
    $\implies$ = empirical covariance or average Hessian
  - sum makes sense as in the same space (tangent of parameter space at point θ)
- link with Cramer-Rao bound, reached (theorem, Amari) [Cover&Thomas: chapter 11.10, p 418 (392) : estimator bias, Cramer-Rao]
  - estimator: given observation x_i, guess the parameter θ so that $p_\theta(x)$ fits the best
  - estimator unbiased if $E_{X\sim p_\theta}[\theta_{estimated}(X)] = \theta$
  - question: mean of estimator ok, what about the variance? is $|\theta_{estimated} - \theta|$ typically big?
  - Theorem: Cramér-Rao inequality:
    variance(any unbiased estimator) $\geq 1/J(\theta)$ in dim 1
    covariance(estimator) $\geq J(\theta)^{-1}$ in higher dims
  - Proof in dim 1
    * unbiased estimator T
    * $p = p_\theta$
    * $V = d\log p/d\theta$
    * Cauchy-Schwartz : $E_{x\sim p}[VT]^2 \leqslant E[V^2]E[T^2]$
      $\leqslant J(\theta)var(T)$
    * $E[VT] = 1$:
    * $E[VT] = \int_x p\ d\log p/d\theta\ T = d/d\theta \int_x p\ T = d/d\theta\ \theta = 1$

II - Natural gradient [Bensadon MDL talk 6]

- desired properties

- invariance to reparameterization
    - importance of metrics
        - norm of a variation $d\theta$?
        - dependance to parameter representation

- natural gradient

- Approximations and related
    - $Hessian^-1$
    - diagonalized
    - Kalman
    - EM (expectation-minimization) = natural gradient step

- Newton
    - exponential family

III - Universal coding

[Bensadon p 30]

- model with parameters
- data arrives –> 4 ways to update models / encode parameters :

1. Explicit encoding of parameters

- read all data, estimate parameters
- encode parameters
- re-read all data and encode data given the model with those parameters –> Pb: encode a real value? infinite number of bits?
- k first binary digits of θ: $K(\theta) = k$, precision on $\theta = 2^{-k}$
- $K(\mu) - \log(\mu(x))$: complexity vs accuracy $\implies$ we'll see the optimal choice of k in next section (using Fisher information)

2. Parameters update, no encoding

- start from canonical parameters
- encode a bit of data
- update the parameters accordingly
- iterate

In 2: - no need to encode parameters! - gain: no encoded/hard-coded parameter - cost: the data is encoded with wrong parameters at the beginning, so its length is higher before parameters converge

3. Normalized Maximum Likelihood [Bensadon p 34]

- issue with 1:
  - given x, pick θ_1 and encode x with p_{θ_1} (encoding θ_1 first) or pick θ_2 and encode x with p_{θ_2} (encoding θ_2 first)
    $\implies$ these 2 codes are different but encode the same x
    $\implies$ redundancy $\implies$ not optimal code
- solution: for a given x, set $p_\theta(x) = 0$ for all θ for which $p_\theta(x) < p_{best\theta}(x)$; denote $\theta(x) =$ best θ for $x = argmax_\theta\ p_\theta(x)$
- and renormalize: $NML(x) = p_{\theta(x)}(x)/\sum_z p_{\theta(z)}(z)$
- issues: many, for instance $NML(x1, x2) \neq NML(x1)\ NML(x2|x1)$

4. Choose a prior q over parameters

- and integrate over it: $p_{model}(x) = \int_\theta q(\theta)\ proba_{model,\theta}(x) d\theta$
- now independent of θ
- as if replacing $\sum_\mu 2^{-\mu}[...]$ by $\int_\theta q(\theta)[...]$
- encode
- Note: choosing explicitely a parameter is less efficient: [Bensadon p 34]
  - $\max(p(\theta_1)\ p_{\theta_1}(x),\ p(\theta_2)\ p_{\theta_2}(x))$ vs $\sum_\theta p(\theta)\ p_\theta(x)$
  - $\implies$ longer codewords!

Cover&Thomas p 433-434

Example with Bernouilli(θ): binary sequence 001111000010100100

- code with θ known: entropy =
- count number of 1, encode it, select sequence among all sequences with that number of 1 = twice shorter
- uniform prior on $\theta \implies$ new $proba(x_1, ..., x_n) = \int_\theta p_\theta(x_1, ..., x_n) d\theta \implies$ idem
- Laplace estimate of θ on the fly: $p(x_{i+1}|x_{\leqslant i}) = \frac{\text{number of 1 so far+1}}{i+1+1}$
- from other prior on θ: Dirichlet$(1/2, 1/2)$ $(= \beta)$: $p(\theta) = 1/\pi\sqrt{\theta(1-\theta)}$ $\implies$ as if $p(x_{i+1}|x_{\leq i}) =$ (number of 1 so far $+1/2)/(i+1)$

IV - Parameter precision

[Bensadon p 30]

- precision when encoding / estimating parameters
- cf Cramer-Rao

V - Prior by default

- Jeffrey's prior [Bensadon p 34]
- Example: Krichevsky-Trofimov estimator
  - Bernouilli(θ)

VI - Examples / Miscellaneous

- justification of BIC? [Bensadon p 32]
- BIC : $K(\mu) = 1/2$ number of parameters * log(number of observations) Bayesian Information Criterion (BIC) [Schwartz, 1978].
- These 2 approaches leads to the same encoding cost:
  - two-part code prior (encode parameter then data):
    - cost of encoding parameters = optimal precision * nb parameters
  - regret using Jeffrey's prior

// + CTW [Bensadon p 37] // + ex: music partition generation with RNN ?

Note: - maximizing entropy can be good: + $KL(p_\theta || uniform) = \sum_x p_\theta(x) \log(p_\theta(x)|X|) = \log|X| - H(p_\theta) \implies$ if you have no information on the law, pick the parameter leading to highest entropy

---

VII - Conclusion of the Information Theory part

- Information geometry
