

Things I wish I understood better about learning theory

David Corfield

Energy or probability Ia

- Gibbs distribution is used to assign a probability distribution to a collection of states and their energy levels.
- $P(x) = \frac{1}{Z} e^{-\beta E(x)}$, where $Z = \frac{\int d^d x \mu(x) e^{-\beta E(x)}}{\int d^d x \mu(x)}$.
- $\beta = 1/T$, the inverse of the temperature.
- So we can pass in both directions between relative probabilities and energies. This involves changing the semiring.
- $\mathbb{R}^+ = \{[0, +\infty), +, 0, \times, 1\} \rightarrow \mathbb{R}^T = \{\mathbb{R} \cup \{+\infty\}, \oplus, +\infty, +, 0\}, x \mapsto -T \ln x$

Energy or probability Ib

- $\mathbb{R}^T = \{\mathbb{R} \cup \{+\infty\}, \oplus, +\infty, +, 0\}$
- Taking the low temperature limit, $T \rightarrow 0$, \mathbb{R}^T tends to $\mathbb{R}^{min} = \{\mathbb{R} \cup \{+\infty\}, min, +\infty, +, 0\}$
- At zero temperature, system falls into state of minimum energy
- Laplace transform becomes Legendre transform

Energy or probability II

- $\mathbb{R}^T = \{\mathbb{R} \cup \{+\infty\}, \oplus, +\infty, +, 0\}$
- If T is replaced by $i\hbar$, we can relate statistical mechanics and quantum theory.
- Laplace transform becomes Fourier transform.

Energy or probability III

- Yann LeCun (see references) promotes greater flexibility of unnormalised energies. If you are not going to treat the whole space of states, but rather just concentrate on the state of least energy and its neighbours, then no need to bear computational burden. Also some unnormalisable energy functions are very useful.
- However, Lemm: “Statistical field theories, which encompass quantum mechanics and quantum field theory in their Euclidean formulation, are technically similar to a nonparametric Bayesian approach.” So lose insight from physics.
- And, miss out on geometry, perhaps.

Maximum Likelihood

- Consider distributions over the four different bit strings of length 2, $p_{00}, p_{01}, p_{10}, p_{11}$.
- Form a three-dimensional simplex as the sum of the p s is 1.
- Distributions which share the same mean number of 1s form two-dimensional subspaces.
- Say we observe 6 00s and 3 11s, empirical distribution $(2/3, 0, 0, 1/3)$.
- Let whiteboard represent distributions for which
$$2.p_{00} + p_{01} + p_{10} : p_{01} + p_{10} + 2.p_{11} = 2$$

Maximum Likelihood II

- Try to model data by assuming independence in bit strings, i.e, $p_{ij} = \text{prob}(i).\text{prob}(j)$
- This forms a one dimensional family, (q^2, pq, pq, p^2)
- Family meets board at
 $p_{00} = 4/9, p_{01} = p_{10} = 2/9, p_{11} = 1/9$
- This is the Maximum likelihood distribution, i.e., the member of the family which finds the evidence most likely.

Information Geometry

Two stories:

- (1) Member of this family to which the distance from the empirical distribution is least.
- (2) Distribution on board from which distance to uniform distribution is least.
- Board is affine for mixtures: line joining $(2/3, 0, 0, 1/3)$ and $(4/9, 2/9, 2/9, 1/9)$ stays in the board.
- Binomial family not affine for mixtures: e.g. $(4/9, 2/9, 2/9, 1/9)$ and $(1/4, 1/4, 1/4, 1/4)$, but affine if you take logarithms, take point on line between them, and normalise.

Information Geometry II

- Binomials form an exponential family, as do normal, gamma, chi-square, beta, Dirichlet, Bernoulli, multinomial, Poisson, negative binomial, geometric...
- These are maximum entropy (Shannon) distributions given constraints.
- Expressible as $P(x|\theta) = \exp(\langle \theta, \phi(x) \rangle - g(\theta))$, where $g(\theta) = \ln \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x)$.
- $\phi(x)$ represents sufficient statistics.

Approximate moment matching

- Exponential family model has as many parameters as moments matched.
- If many moments, family will come close to empirical distribution. This will become a huge problem when we look at infinite-dimensional models.
- Instead we may require that moments are matched only approximately.
- E.g., moments of estimate need to be in a box or sphere centred on empirical moments.

Maximum a posteriori

Match moments approximately = Placing prior on family.

- (1) Where in this family is the sum of the distance from empirical and negative log prior probability least?
- (2) Which distribution in thickened neighbourhood of board is nearest to reference?
- Fenchel duality explains what's going on.
- Box corresponds to Laplace prior, sphere to Gaussian prior.

Different geometries

- Why was it distance TO estimate FROM empirical, but distance FROM estimate TO reference?
- KL divergence FROM empirical
- KL divergence TO reference = reverse KL divergence FROM reference
- Distances $D_\delta(p, q) = - \int \frac{p^\delta q^{1-\delta}}{\delta(1-\delta)}$, for $\delta \neq 0, 1$
- $D_1(p, q) = D_0(q, p) = \int p \log \frac{p}{q}$
- Could use any pair of dual geometries D_δ and $D_{1-\delta}$
- Distances extend to all finite positive measures.

Fisher information metric

Taylor expansion of δ -deviations

- $D_\delta(\theta + \epsilon\nu, \theta) = \frac{1}{2}g_{ij}\nu^i\nu^j\epsilon^2 + \dots$, all δ
- $g_{ij} = \int \frac{\partial \log p(x, \theta)}{\partial \theta_i} \frac{\partial \log p(x, \theta)}{\partial \theta_j} p(x, \theta) dx$
- “The distance between two points on a statistical differential manifold is the amount of information between them, i.e. the informational difference between them.”
- Jeffrey prior $\propto \sqrt{\det J(\theta)}$, where $J = (g_{ij})$

Information geometry

Zippering about on geodesics makes sense of:

- ICA, PCA, mean field, EM, boosting,...
- Belief propagation in neural networks.
- But, not yet Bayesian

Bayesian IG - Zhu and Rohwer

Let τ be estimator. Given sample z , $\tau(z)$ is a probability distribution.

- Define the generalization error:

$$E(\tau) = \int_p P(p) \int_z P(z|p) D(p, \tau(z))$$

- Show $\hat{p}(z) = \tau(z) = \arg \min_q \int_p P(p|z) D(p, q)$
- $\hat{p}(z) = \tau(z) = \int p P(p|z)$, posterior mean.
- Different geometries would require different coordinates for p .

Bayesian IG - Snoussi

- Choice of prior as weighted mixture of proximity to uniform (Jeffrey) distribution over distributions and proximity of mean to a specific distribution.
- Different spaces, different geometries
- Captures entropic prior, conjugate prior, Gaussian, Jeffrey.
- Question: can we tell two stories for Bayesian IG?
- Question: can we make sense of Bayesian ICA, PCA, etc.?

Conditional spaces

- Guy Lebanon studied the sensible metrics for conditional distributions with finite X and Y
- Product of simplices
- More or less - Fisher information.
- Continuous spaces?

Infinite-dimensional families

- Manifolds with charts to subspaces of an RKHS - Fukumizu
- Maximum likelihood leads to problem with moment matching, need regularisation
- MAP + RKHS conditional exponential family capture gaussian process classification and regression - Smola and Altun
- Bayesian information geometry + RKHS conditional exponential family = full Bayesian GP?

Model Comparison

There has been a long history of trying to compare statistical models. It has been known for many centuries that one wants to balance accuracy with simplicity. Several criteria have been suggested, including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These and the one discussed next involve the dimension of the model. My question is whether the techniques we shall now consider, although devised for finite-dimensional case, can still help us when we treat infinite-dimensional models.

Model Comparison - Balasubramian

- $AIC : -NL_N(\hat{\theta}) + d$
- $BIC : -NL_N(\hat{\theta}) + \frac{d}{2} \ln N$
- $BAL : -NL_N(\hat{\theta}) + \frac{d}{2} \ln \frac{N}{2\pi} + \ln \int d\theta \sqrt{\det J(\theta)} + \frac{1}{2} \ln \left(\frac{\det I(\hat{\theta})}{\det J(\hat{\theta})} \right) + O(1/N)$
- $I_{\mu\nu} = (-1/N) \nabla_{\theta_\mu} \nabla_{\theta_\nu} \ln Pr(E|\theta)|_{\hat{\theta}}$ is the empirical Fisher information.
- Terms 2, 3 and 4 are $\ln \left(\frac{V(A)}{V_c(A)} \right)$, where $V_c(A)$ is essentially the volume of a small ellipsoid around $\hat{\theta}$ within which $Pr(E|\theta)$ is appreciable.

Model Comparison - Balasubramian

- $Pr(A|E) = \frac{Pr(A)}{Pr(E)} \int d^d\Theta w(\Theta) Pr(E|\Theta)$, Jeffrey prior for $w(\Theta)$.
- Focusing only on the integral:
- $P_{A|E} = \frac{\int d^d\Theta \sqrt{\det J} Pr(E|\Theta)}{\int d^d\Theta \sqrt{\det J}}$
- $P_{A|E} = \frac{\int d^d\Theta \sqrt{\det J} \exp[-N(\frac{-\ln Pr(E|\Theta)}{N})]}{\int d^d\Theta \sqrt{\det J}}$

Model Comparison - Balasubramian

- $$P_{A|E} = \frac{\int d^d \Theta \sqrt{\det J} \exp[-N(\frac{-\ln Pr(E|\Theta)}{N})]}{\int d^d \Theta \sqrt{\det J}}$$
- Compare $Z = \frac{\int d^d x \mu(x) e^{-\beta E(x)}}{\int d^d x \mu(x)}$, partition function from statistical mechanics.
- N is inverse temperature, and energy of a state is $(-1/N)\ln Pr(E|\Theta)$ which converges almost certainly by strong law of large numbers to $D(t, \Theta) + h(t)$, low temperature expansion.
- Rodriguez looks at $O(1/N)$ terms, curvature of manifold, and for different geometries δ .

Model Comparison - Nonparametric?

- BAL is detecting good feature of model that an appreciable volume of parameter space is not much further from truth than maximum likelihood.
- Can we use Snoussi's priors and expand around MAP rather than ML?
- Any chance of using this to think about RKHS spaces?
- Perhaps the ellipsoid in many dimensions is very large, leaving only effective dimensions?

Model Comparison - Nonparametric?

- Gaussian process regression can be treated analytically, but Gaussian process classification requires approximation. Rasmussen and Kuss compare two approximations, which try to fit a Gaussian, with Monte Carlo simulation.
- The Laplace approximation centres the Gaussian on the MAP distribution, and chooses as covariance matrix the inverse Hessian of the log posterior density there. This resembles Balasubramanian's set up.
- The EP approximation chooses mean and covariance to match approximate marginal moments for each data point.
- The EP approx. matches MCMC method very closely. Can we use it to think about effective dimension?

References and websites

- LeCun et al., ‘A Tutorial on Energy-Based Learning’, <http://yann.lecun.com/exdb/publis/index.html#lecun-06>.
- Jörg Lemm, Bayesian Field Theory, <http://pauli.uni-muenster.de/lemm/>.
- Papers by Zhu and Rohwer, Snoussi and Djafari, Balasubramanian, and Rodriguez, ‘A Geometric Theory of Ignorance’, <http://omega.albany.edu:8008/ignorance/>.
- M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679-1704, 2005. www.jmlr.org/papers/volume6/kuss05a/kuss05a.pdf
- Also see <http://www.dcorfield.pwp.blueyonder.co.uk/MaxEntInfGeom.html>.