Analysis Seminar Talk Notes 11/22

Jared Marx-Kuo

Nov. 22nd, 2019

Talk Outline

1. Basics of Entropy
2. Defining CLT + Standardized Sum
3. Fisher Information
4. Variational Theorem
5. Main theorem
6. Mark what stuff to star/save throughout the talk. Maybe write it on the whiteboard

Basics of Entropy

1. As always, for random variables X with distribution $f : \mathbb{R} \to [0, \infty)$, we have

$$H(X) = -\int f(x) \log f(x) dx$$

2. **Theorem:** the Gaussian distribution is the maximizer when the mean and variance are fixed
Proof: A gaussian distribution with fixed mean and variance is given by density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-((x-\mu)/\sigma)^2/2}$$

So that the entropy is equal to

$$H = -\int_{\mathbb{R}} f(x) \log(f(x)) = -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} e^{-((x-\mu)/\sigma)^2/2} \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 - \log(\sqrt{2\pi}\sigma) \right] dx = \frac{1}{2}(1 + \log(2\pi\sigma^2))$$

For the rest of the proof, we need lemmas

3. **Lemma:** For $x \geq 0$, $y \geq 0$, we ahve

$$y - \log y \leq x - y \log x$$

with equality iff $x = y$.
Proof: follows because $\log t \leq t - 1$, setting $t = x/y$ and ruling out the case when $y = 0$. □

4. **Lemma:** Let $p(x)$, $q(x)$, continuous probability densities on $\mathbb{R}$, we have

$$-\int p \log p dx \leq -\int p \log q$$

if both integrals exist, with equality when $p \equiv q$.
Proof: By our previous lemma $p(x) - p(x) \log p(x) \leq q(x) - p(x) \log q(x)$, and so we integrate and use the fact that $\int p(x) = \int q(x) = 1$. Note that the equality case follows because equality yields

$$\int q - p \log q - p + p \log p = 0$$

and the integrand is non-negative everywhere by the lemma and so must in fact be zero everywhere. □

5. **Corollary:** For p and q as before, assume $q(x) > 0$ for all $x \in \mathbb{R}$. If we have

$$-\int p \log q = h(q)$$

then $h(p) \leq h(q)$ with equality iff $p \equiv q$.
Proof: This follows from the previous lemma □

6. Proof of the theorem: Let p be a probability density function on $\mathbb{R}$ with variance σ^2 and mean μ, then for q the Gaussian with the same mean, variance, we have

$$-\int p(x) \log q(x) = \frac{1}{2} \int p(x) \left(\log(2\pi\sigma^2) + \left(\frac{x - \mu}{\sigma} \right)^2 \right) dx = \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} = h(q)$$

by $\int p = 1$ and definition of variance of p. This tells us that $h(p) \leq h(q)$ with equality iff p is a Gaussian itself everywhere. □

7. Remark: When there are no constraints on the variance, the density function which maximizes entropy is the uniform distribution as in the discrete case

8. (SAVE) We define the normalized sum of n variables $\{X_i\}$, i.i.d. with variance 1 to be

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

9. Shannon first proved that $Ent(Y_2) \geq Ent(Y_1)$, which implies that $Ent(Y_{2^k}) \geq Y_{2^{k-1}}$

10. What about in general $Ent(Y_{n+1}) \geq Ent(Y_n)$? Turns out to be a much harder problem

CLT + Entropy

1. In this talk, we'll be proving the following theorem:
(SAVE) **Theorem:** (Monotonicity) For $\{X_i\}_{i=1}^{\infty}$ i.i.d. variables with finite variance, then

$$Ent\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}} \right) \leq Ent\left(\frac{X_1 + \cdots + X_{n+1}}{\sqrt{n+1}} \right)$$

2. **Remark**: We cited/showed that for fixed mean and variance, entropy has the unique maximizer of being a gaussian distribution, so assuming the X_i have variance 1, then if the entropy increases to 1, then there should be **some** convergence of our standardized sum distribution to a Gaussian distribution with variance 1.

3. **Remark**: I know what you're thinking: "Jared, non-strict inequality? Come on, no way we'll ever show that the entropy increases to 1." True but this is the best we'll get if we allow for any collection of i.i.d. variables, because we could very well get equality when all the X_i are normally distributed

4. For this, I'll refer you to: "Entropy and the Central Limit Theorem" by Andrew Barron for a proof of convergence

5. Nonetheless, the intuition is that the central limit theorem is driven by the second law of thermodynamics! And somehow, the Gaussian distribution is the most entropic/disorderly when associating it to a collection of rescaled variables (i.e. the standardized sum)

6. In order to prove theorem, we'll be proving some other lemmas/theorems

7. (SAVE) **Theorem:** : For $\{X_1, \ldots, X_{n+1}\}$ i.i.d. let $(a_1, \ldots, a_{n+1}) \in S^n$, then

$$Ent\left(\sum_{i=1}^{n+1} a_i X_i \right) \geq \sum_{i=1}^{n+1} \frac{1 - a_j^2}{n} Ent\left(\frac{1}{\sqrt{1 - a_j^2}} \sum_{i \neq j} a_i X_i \right)$$

8. **Remark:** Plug in $a_j = \frac{1}{\sqrt{n+1}}$ and we get the monotoncity theorem

Fisher Information

1. In order to prove the monotoncity theorem, we (re)introduce the **fisher information**. For a random var X with probability density f, we have

$$J(X) = \int_{\mathbb{R}} \frac{(f')^2}{f} = \int_{\mathbb{R}} \frac{[\partial_\theta f(x,\theta)]^2}{f(x,\theta)} dx$$

2. Morally, the fisher information comes from measuring how much a random variable X, depends on an unknown parameter θ.

3. We search for critical points of the conditional probability $f(x, \theta = \theta_0)$. The "best" guesses for a value of $\theta = \theta_0$ would be at a critical point of this function

4. The idea is as follows: suppose we have a family of probability distributions $\{p(\cdot,\theta)\}_{\theta\in\mathbb{R}}$. We fix $\theta = \theta_0$, but we don't know what it is. From $p(\cdot,\theta_0)$, we measure $X = x$.

5. Given our measurement of $X = x$, we want to figure out what θ is, so we now think of the function $g(\theta) = p(x,\theta)$, except really its

$$g(\theta) = \log p(x,\theta)$$

because statisticians care about the "log likelihood"

6. In particular, we're interested in critical points of g, because that's where g will either be the smallest (i.e. it's very unlikely that given $X = x$, we had parameter θ to start with) or the largest (i.e. its more likely). Critical points yields

$$\frac{\partial}{\partial\theta} g(\theta) = \frac{\partial_\theta p(x,\theta)}{p(x,\theta)}\Big|_{\theta=\theta_0} = 0 \implies \partial_\theta p(x,\theta)\Big|_{\theta=\theta_0} = 0$$

7. Moreover, at this critical point, we consider the curvature of the graph $(\theta, g(\theta))$, which we recall is given by

$$K = \frac{g''}{(1+g')^{3/2}}\Big|_{\theta_0} = g''(\theta_0)$$

for θ_0 a critical point. Intuitively, the curvature tells us about the likelihood that θ actually equals θ_0 - if the curvature is very high, then the likelihood dips very sharply around θ, so having observed x, you gain a lot of information about where θ probably is. Similarly, if the curvature is low, then in this neighborhood of θ_0, it's nearly equally likely that θ is one of these values, hence we glean little information

8. Fisher information then sums up all of the information we get about where θ actually is.

9. Ex: Let's do it for the Gaussian $X \sim N(\mu, \sigma^2)$ where μ is unknown but σ^2 is fixed, then $f(x,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ so

$$g(\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \implies g'(\mu) = \frac{x-\mu}{\sigma^2}, \qquad g''(\mu) = -\frac{1}{\sigma^2}$$

turns out, the Gaussian has the smallest fisher information over the collection of RV's with fixed variance.

10. (SAVE) Fisher regulation is not always finite, but when it is, it satisfies

$$Ent(G) - Ent(X) = \int_0^\infty (J(X^t) - 1)dt$$

11. Here X^t is the flow of X under some semigroup, but it doesn't matter what that flow is, it's just important to know that X^t is distributed as $\sqrt{e^{-2t}}X + \sqrt{1-e^{-2t}}G$ where G is a standard gaussian.

12. With the above, we'll replace X with Y_n, noting that

$$Y_n^t = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^t$$

the fact the flow commutes with taking the standardized sum is somewhat nontrivial.

13. Downshot: If we can show that fisher information decreases with n for all t, then we'll show that $Ent(Y_n)$ is increasing

Variational Theorem

1. In order to prove our entropy theorem, we use the **variational characterization of information**
2. (SAVE) **Theorem:** Let $w : \mathbb{R}^n \to (0, \infty)$ be a C^2 density on $\mathbb{R}^n$ with

   $$\int \frac{||\nabla w||^2}{w}, \quad \int ||Hess(w)|| < \infty$$

   For e a unit vector, and h, the marginal density in the direction of e defined by

   $$h(t) = \int_{te+e^\perp} w$$

   Then the fisher information of the marginal density, h, satisfies

   $$J(h) \le \int \left(\frac{\nabla \cdot (pw)}{w} \right)^2 w$$

   for any C^1 vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ such that $\langle p(x), e \rangle = 1$ for all x and p is integrable w.r.t. to w, i.e. $\int ||p||w < \infty$.
   Proof: Not very enlightening. After looking through it, it's a lot of holder inequalities and fubini theorem. □
3. **Remark:** The point of the theorem is that we want $\nabla_{e^\perp}(pw) - \partial_e w$ to integrate to zero on each hyperplane of the form $te + e^\perp$, and this is guaranteed by green's theorem if we have sufficient decay at infinity (draw picture)

Main theorem

1. For $a = (a_1, \ldots, a_{n+1}) \in S^n$, we now define

   $$w(x_1, \ldots, x_{n+1}) = f_1(x_1) \cdots f_{n+1}(x_{n+1})$$

   which is the marginal density of $Z = \sum_{i=1}^{n+1} a_i X_i$ in the direction $(a_1, \ldots, a_{n+1}) \in S^n$. This makes sense because $P(Z = z_0)$ like integrating the probability distribution over all $(x_1, \ldots, x_{n+1})$ such that $\sum_i a_i x_i = z_0$, which is integrating w over the hyperplane $z_0 \vec{a} + \vec{a}^\perp$.
2. Using our reduction of entropy to fisher information, the new goal is to prove

   $$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \le n \sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right)$$

   and then let $a_i = (n+1)^{-1/2}$ and $b_j = \frac{1}{n^2}\sqrt{1-a_j^2}$.
3. With the above, we have for $\{X_i\}$ a group of i.i.d. variables, that

   $$J(Y_n) \le n \cdot (n+1) \frac{1}{n^2} \frac{n}{n+1} J\left(\sqrt{\frac{n+1}{n}} \frac{1}{\sqrt{n+1}} \sum_{i \neq j} X_i\right) = J(Y_n)$$

   having noted that $J(c \sum_{i \neq j} X_i)$ is independent of j.
4. **Proof:** Let

   $$\hat{a}_j = \frac{1}{\sqrt{1-a_j^2}}(a_1, \ldots, a_{j-1}, 0, a_{j+1}, \ldots, a_n)$$

   which is a unit vector. Using our variational theorem, we look at the marginal density

   $$h_j = \int_{t\hat{a}_j + \hat{a}_j^\perp} w dx$$

and then use theorem 4 to define $p^j : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ to be a vector field such that

$$J\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right) = \int_{\mathbb{R}^n}\left(\frac{\nabla\cdot(wp^j)}{w}\right)^2 w$$

Moreover, we can choose p^j such that p^j has no jth coordinate, as we know that $\langle p^j, \hat{a}_j\rangle = 1$ for all x and so this condition nor the divergence will be affected by killing the e_j component of p^j.

Now look at $p = \sum_{j=1}^{n+1} b_j p^j$. Because $\sum_{j=1}^{n+1} b_j\sqrt{1-a_j^2} = 1$ by assumption, we get $\langle p, \hat{a}\rangle = 1$. Then by the variational theorem applied to p, we have

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq \int_{\mathbb{R}^n}\left(\frac{\nabla\cdot(wp)}{w}\right)^2 w = \int_{\mathbb{R}^n}\left(\sum_{j=1}^{n+1} b_j\frac{\div(wp^j)}{w}\right)^2 w$$

Now let $y_j = b_j\frac{\nabla\cdot(wp^j)}{w}$, then we want to show that in $L^2(w)$, i.e. the hilbert space with weight w, that

$$||y_1 + \cdots + y_{n+1}||^2 \leq n(||y_1||^2 + \cdots + ||y_{n+1}||^2)$$

Normally, cauchy schwartz gives us the above with n replaced by $n+1$, but because of the form of the y_i and the fact that we're considering this weighted L^2 space, it actually works out.

With this, we have

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq n\int\sum_{j=1}^{n+1}\left(\frac{\nabla\cdot(wp^j)}{w}\right)^2 w = n\sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right)$$

which is what we sought out to prove. □

References

https://sgfin.github.io/2017/03/16/Deriving-probability-distributions-using-the-Principle-of-Maximum-Entrop
https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf
https://en.wikipedia.org/wiki/Fisher_information#Definition
https://math.stackexchange.com/questions/265917/intuitive-explanation-of-a-definition-of-the-fisher-informa
https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf