Supplement for Mathematics Texts

Stefan Lukits

October 29, 2018

Contents

1 Legendre Duality and Convex Conjugates

I will provide an original new proof for this theorem (there are more conventional proofs in Savage, 1971, 788; and Selten, 1998, section 4). My proof gives the reader a brief introduction to convex conjugates, which may in many other respects be a fruitful mathematical tool dealing with scoring rules, entropy functions, and divergence functions. Let $\mathcal{P}$ be the set of probability distributions over a trichotomy-type $n+1$-dimensional outcome space. $x \in \mathcal{P}$ can be represented by the probabilities $x_0, \ldots, x_n$ or by the vector $x = (x_1, \ldots, x_n)^\intercal \in \mathbb{R}^n$ with the usual restrictions on probabilities. If the latter is the case, which will be true for the rest of this subsection, then $x_0 = 1 - \sum x_i$.

This differs in notation from the rest of the paper; and so does my assumption for the rest of the subsection that S is a reward (rather than a loss) function. The reason for the inconsistency of convention is that I will use a fair amount of convex analysis in this subsection. The entropy functions for loss scores are concave. All theorems of convex analysis are valid for concave functions just as much as for convex functions, but in convex analysis, for obvious reasons (to align convex sets with convex functions), the convention is to use convex functions.

Definition 1.1. A function $f : T \to \mathbb{R}$, where T is a convex subset of $\mathbb{R}^n$, is convex if its epigraph is a convex set.

The epigraph $\mathrm{epi}(f)$ is defined as $\{(x, \mu) \in T \times \mathbb{R} \,|\, \mu \geq f(x)\}$. The more conventional definition of convexity is that for $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \text{ for all } x, y \in T. \tag{1}$$

Theorem 4.1 in Rockafellar, 1997, 25, shows these two definitions of convex functions to be equivalent.

The reward function $S_{\text{LS}} : \mathbb{R}^n \to \mathbb{R}^{n+1}$ for the Log score is

$$S_{\text{LS}}(x) = \left(\ln \left(1 - \sum x_i \right), \ln x_1, \ldots, \ln x_n \right)^\intercal \tag{2}$$

and the entropy function is

$$H_{\text{LS}}(x) = -\sum_{i=1}^{n} x_i \ln x_i - \left(1 - \sum_{i=1}^{n} x_i \right) \ln \left(1 - \sum_{i=1}^{n} x_i \right). \tag{3}$$

I will use equations (2) and (3) to illustrate convex conjugates.

Let the entropy function H for the scoring rules under consideration be a twice differentiable, strictly convex function on $\mathcal{P}$. For the remainder of this subsection, I will assume commutativity of the inner product and leave aside technicalities about the inner product between dual spaces.

Definition 1.2. The convex conjugate $H^* : \mathcal{P}^* \to \mathbb{R}$ is defined on

$$\mathcal{P}^* = \{x^* \in \mathbb{R}^n \,|\, \nabla H(x) = x^*, x \in \mathcal{P}\} \tag{4}$$

by

$$H^*(x^*) = \sup_{x \in \mathcal{P}} \{\langle x, x^* \rangle - H(x)\}. \tag{5}$$

Based on the strict convexity and differentiability of H, $\mathcal{P}$ and $\mathcal{P}^*$ are dual spaces and

$$H^*(x^*) = \langle x, x^* \rangle - H(x) \tag{6}$$

where x is the dual element corresponding to x^*, i.e. $\nabla H(x) = x^*$ and the supremum in (5) is a maximum, if it exists.

Convex conjugates induce the so-called Legendre-Fenchel duality between convex differentiable functions on $\mathcal{P}$ and $\mathcal{P}^*$.

Lemmas 1.1, 1.2, and 1.3 are adapted from section 12 of R. Tyrrell Rockafellar's book *Convex Analysis*.

Lemma 1.1. $H^{**} = H$.

Proof. First, I will show that H^* is convex, otherwise duality cannot be maintained because x would no longer have a unique dual element x^*. The definition of H^* immediately suggests that

$$H^*(x^*) = \sup_{(x,\mu) \in \mathrm{epi}(H)} \{\langle x, x^* \rangle - \mu\} \tag{7}$$

Therefore,

$$\mathrm{epi}(H^*) = \{(x^*, \nu) \,|\, \nu \geq \sup_{(x,\mu) \in \mathrm{epi}(H)} \{\langle x, x^* \rangle - \mu\} \tag{8}$$

Let (x_1^*, ν_1) and (x_2^*, ν_2) be elements of $\text{epi}(H^*)$. Then $(\lambda x_1^* + (1-\lambda)x_2^*, \lambda\nu_1 + (1-\lambda)\nu_2)$ is also an element of $\text{epi}(H^*)$ because for fixed $x_1^*, x_2^*, \nu_1, \nu_2$,and $0 < \lambda < 1$

$$\lambda\nu_1 + (1-\lambda)\nu_2 \geq \lambda \sup_{(x,\mu)\in\text{epi}(H)} \{\langle x, x_1^*\rangle - \mu\} + (1-\lambda) \sup_{(x,\mu)\in\text{epi}(H)} \{\langle x, x_2^*\rangle - \mu\} \geq$$
$$\sup_{(x,\mu)\in\text{epi}(H)} \{\lambda(\langle x, x_1^*\rangle - \mu) + (1-\lambda)(\langle x, x_2^*\rangle - \mu)\} =$$
$$\sup_{(x,\mu)\in\text{epi}(H)} \{\langle x, \lambda x_1^* + (1-\lambda)x_2^*\rangle - \mu\} \tag{9}$$

This establishes convexity for H^*. Consider

$$\sup_{(x^*,\nu)\in\text{epi}(H^*)} \{\langle x, x^*\rangle - \nu\} \tag{10}$$

On the one hand, (10) is $H^{**}(x)$ by definition. On the other hand, it is

$$\sup_{x^*\in\mathcal{P}^*} \{\langle x, x^*\rangle - H^*(x^*)\} = \sup_{x^*\in\mathcal{P}^*} \left\{\langle x, x^*\rangle - \sup_{\xi\in\mathcal{P}}\{\langle \xi, x^*\rangle - H(\xi)\}\right\} \tag{11}$$

The inside supremum is achieved where $\nabla H(\xi) = x^*$ because H is differentiable. Differentiate $\langle x-\xi, \nabla H(\xi)\rangle + H(\xi)$, using the product rule for the inner product, to see that the outside supremum is achieved where $\nabla H(x) = x^*$. The convexity of H demands that $x = \xi$, and it is therefore established that (10) is not only $H^{**}(x)$, but also $H(x)$, so $H^{**} = H$. □

Lemma 1.2. Taking conjugates reverses functional inequalities: $f_1 \leq f_2$ implies $f_1^* \geq f_2^*$.

Proof. A real-valued function f_2 is greater or equal to another real-valued function f_1 on a common domain if on that common domain $f_1(x) \leq f_2(x)$. Let $f_1 \leq f_2$. Then

$$f_2^*(x^*) = \sup_{x\in\mathcal{P}}\{\langle x, x^*\rangle - f_2(x)\} \leq \sup_{x\in\mathcal{P}}\{\langle x, x^*\rangle - f_1(x)\} = f_1^*(x^*) \tag{12}$$

for all $x^* \in \mathcal{P}^*$. □

Lemma 1.3. $\langle x, x^*\rangle \leq f(x) + f^*(x^*)$ for all $x \in \mathcal{P}, x^* \in \mathcal{P}^*$ (not necessarily dual elements as above), and any differentiable convex function f and its conjugate f^*.

Proof. Consider for fixed x, x^*

$$f^*(x^*) = \sup_{\xi \in \mathcal{P}} \{\langle \xi, x^* \rangle - f(\xi)\} \geq \langle x, x^* \rangle - f(x) \tag{13}$$

$\langle x, x^* \rangle \leq f(x) + f^*(x^*)$ follows. □

The inequality in lemma 1.3 is called Fenchel's inequality.

Lemma 1.4. $(\nabla H)^{-1} = \nabla (H^*)$.

Proof. H is a differentiable and strictly convex function on $\mathcal{P}$, therefore ∇H exists and is bijective from $\mathcal{P}$ to $\mathcal{P}^*$. Let $(\nabla H)^{-1}$ be the inverse function of ∇H. According to (6), the supremum in (5) is reached at $\nabla H(x) = x^*$, therefore

$$H^*(\nabla H(x)) = \langle x, \nabla H(x) \rangle - H(x) \tag{14}$$

Differentiating (14) yields

$$\nabla H^*(\nabla H(x)) \cdot \nabla^2 H(x) = \nabla H(x) + x \cdot \nabla^2 H(x) - \nabla H(x) \tag{15}$$

This can only be true if $\nabla H^* \circ \nabla H = \mathrm{id}$, which establishes lemma 1.4. My proof is based on Boissonnat et al., 2010, 287. Rockafellar treats the topic with greater generality, where H need not be differentiable, see his Theorem 23.5 on page 218 and his Theorem 26.5 on page 258. □

To keep the Bregman divergence corresponding to the entropy function positive, it will in this subsection also be adjusted to rewards rather than losses and defined in accordance with `def:logahpoh`

$$D_H(x \| y) = H(x) - H(y) + \langle y - x, \nabla H(y) \rangle. \tag{16}$$

Lemma 1.5. $D_H(x \| y) = D_{H^*}(y^* \| x^*)$.

Proof. Recall that $H(x) + H^*(x^*) = \langle x, x^* \rangle$ for $\nabla H(x) = x^*$ according to (6). Together with lemma 1.4, the bilinearity and commutativity of the inner

product, and `eq:thiepooz` this yields

$$
\begin{aligned}
&D_H(x\|y) - D_{H^*}(y^*\|x^*) = \\
&H(x) - H(y) - \langle x - y, \nabla H(y)\rangle - H^*(y^*) + H^*(x^*) + \langle y^* - x^*, \nabla H^*(x^*)\rangle = \\
&H(x) - H(y) - \langle x, \nabla H(y)\rangle + \langle y, y^*\rangle - H^*(y^*) + \\
&H^*(x^*) + \langle \nabla H(y), \nabla H^*(\nabla H(x))\rangle - \langle x^*, \nabla H^*(\nabla H(x))\rangle = \\
&H(x) - H(y) + \langle y, y^*\rangle - H^*(y^*) + H^*(x^*) - \langle x^*, x\rangle = 0
\end{aligned}
\tag{17}
$$

and establishes the lemma. The proof is based on Boissonnat et al., 2010, 287. □

My illustration for convex conjugates is the Log score (the Brier score would make a poor illustration, because, as we will find out, the Brier score is uniquely self-dual). For a given $x \in \mathcal{P}$, differentiate (3) to define x^*

$$
x_i^* = \frac{\partial}{\partial x_i} H_{\text{LS}}(x) = \ln\left(1 - \sum_{j=1}^{n} x_j\right) - \ln x_i. \tag{18}
$$

To find ∇H_{LS}^*, which is equal to $(\nabla H_{\text{LS}})^{-1}$ according to lemma 1.4, solve the system of equations

$$
\sum_{j=1}^{n} \left(e^{x_i^*} + \delta_{ij}\right) x_j = 1. \tag{19}
$$

I found this solution by playing around with traces, inverses, and determinants on octave. Schmierbuch page 2510 has a summary of the conjecture I made on the basis of this playing around. You'd have to do some linear algebra to find a proof for this conjecture.

$$
x_i = \frac{1 + \left(\sum_{i=1}^{n} e^{x_i^*}\right) - n e^{x_i^*}}{1 + \left(\sum_{i=1}^{n} e^{x_i^*}\right)}. \tag{20}
$$

Because of lemma 1.4, just as $\nabla H_{\text{LS}}(x)$ corresponds to the right-hand side of (18), $\nabla H_{\text{LS}}^*(x^*)$ corresponds to the right-hand side of (20). H_{LS}^* is an antiderivative of the partial derivatives in (20) with the constant of integration fixed by definition 1.2.

Lemma 1.6. The only self-dual convex differentiable function on $\mathcal{P}$ is $F(x) = \frac{1}{2}\langle x, x\rangle$.

Proof. I am adapting a proof in Rockafellar, 1997, 106. That F is self-dual follows from definition 1.2 and lemma 1.4. Now let G be a self-dual convex differentiable function. By lemma 1.3 (Fenchel's inequality),

$$\langle x, x\rangle \leq G(x) + G^*(x) = 2G(x). \tag{21}$$

This means that $G \geq F$. By lemma 1.2, $G^* \leq F^*$. Together with the self-duality of both F and G, this implies $F = G$. □

I will use the preceding lemmas to prove `thm:iechohng` stated in the beginning of the subsection, showing that only the Brier score and its close relatives are symmetric.

Proof. Recall that in the context of convex functions, I am using reward functions rather than loss functions. With this adjustment, the reward function and entropy function for the Brier score according to (39) and (40) are

$$S(\xi_i, x) = \frac{2x_i}{\sum_k x_k} - 1 - \sum_j \left(\frac{x_j}{\sum_k x_k}\right)^2 \tag{22}$$

$$H(x) = \sum_i x_i \left(\frac{2x_i}{\sum_k x_k} - 1 - \sum_j \left(\frac{x_j}{\sum_k x_k}\right)^2\right). \tag{23}$$

Since I assume probabilism, the sum of the probabilities is one. The above equations simplify to

$$S(\xi_i, x) = 2x_i - 1 - \sum_j x_j^2 \tag{24}$$

$$H(x) = \sum_i x_i \left(2x_i - 1 - \sum_j x_j^2\right). \tag{25}$$

Note that according to McCarthy's theorem `thm:lahpoodu` the i-th component of the vector $\nabla H(x)$ equals $S(\xi_i, x)$ and therefore (see `cor:quiphaef`)

$$H(x) = \sum_{i=1}^{n} x_i S(\xi_i, x) \tag{26}$$

For a close relative of the Brier score, therefore, reward function and entropy function are

$$S(\xi_i, x) = m\left(2x_i - 1 - \sum_j x_j^2\right) + b \tag{27}$$

$$H(x) = m\left(\sum_i x_i\left(2x_i - 1 - \sum_j x_j^2\right)\right) + b \tag{28}$$

for some $m \in \mathbb{R}^+, b \in \mathbb{R}$. The following calculation shows that the Brier score and its close relatives are symmetric.

$$\begin{aligned}
&D(x\|y) - D(y\|x) = H(x) - H(y) + \langle y, \nabla H(y)\rangle - H(y) + H(x) - \langle x, \nabla H(x)\rangle = \\
&2\left(m\left(\sum_i x_i\left(2x_i - 1 - \sum_j x_j^2\right)\right) + b\right) - \\
&\sum_i x_i\left(m\left(2x_i - 1 - \sum_j x_j^2\right) + b\right) - \sum_i x_i\left(m\left(2y_i - 1 - \sum_j y_j^2\right) + b\right) - \\
&2\left(m\left(\sum_i y_i\left(2y_i - 1 - \sum_j y_j^2\right)\right) + b\right) + \\
&\sum_i y_i\left(m\left(2y_i - 1 - \sum_j y_j^2\right) + b\right) + \sum_i y_i\left(m\left(2x_i - 1 - \sum_j x_j^2\right) + b\right) = 0
\end{aligned} \tag{29}$$

□

2 Hector and Paris

2.1 Setup

Hector and Paris have one dollar to invest in bets on a Bernoulli coin toss. A Bernoulli coin toss models whether or not an event will take place. Am I HIV-positive or not? Will I inherit a large sum of money in the next three months? Will this physical coin toss land heads or tails? Is the next die roll

going to be a five? Will this stock increase by a certain amount in the next two weeks? Betting on a Bernoulli coin toss is related to the partial belief that an agent has about whether the coin toss lands heads. Investing in the stock market or shorting a stock is in some ways like betting on a Bernoulli coin toss; so is paying a certain price for a certain good.

Hector and Paris's betting strategy is represented by a function $R : [0, 1] \to [0, 1]$. If w represents the bets offered to them, $R(w)$ is their response. The offer works as follows. Hector and Paris can invest in an H-bet at w or a T-bet at $1 - w$. I will call their investments the investment schema; I will call the two bets offered to them (H-bet at w and T-bet at $1 - w$) the bets associated with w.

Let us say one of them invests v_H dollars in an H-bet at w. Then his prize money is zero dollars if T is tossed; v_H/w if H is tossed. The gain is $-v_H$ for T; $(v_H - v_H w)/w$ for H. The investment of v_T in a T-bet at $1 - w$ works likewise. Consequently, for investment schema

$$(v_H, v_T) \in \{(v_H, v_T) \in \mathbb{R}^2 | 0 \leq v_H, 0 \leq v_T, v_H + v_T \leq 1\} \tag{30}$$

the gain is

$$\frac{v_H - v_T w - v_H w}{w} \qquad \text{if } H \text{ is tossed} \tag{31}$$

and

$$\frac{-v_H + v_T w + v_H w}{1 - w} \qquad \text{if } T \text{ is tossed} \tag{32}$$

Proposition 2.1. For any investment schema (v_H, v_T) there is an investment schema (v'_H, v'_T) with the same outcome as (v_H, v_T) and $v'_H + v'_T = 1$.

Proof. Define $v'_H = v_H + w - w(v_H + v_T)$ and $v'_T = 1 - v'_H$. Plugging v'_H and v'_T into equations 31 and 32 leaves the resulting gain unchanged, but now $v'_H + v'_T = 1$. □

As a consequence, the investment schemas (v'_H, v'_T) where $v'_H + v'_T = 1$ are representative for all other possible investment schemas. We shall identify the

investment schema (v'_H, v'_T) by $x = v'_H$ (v'_T is uniquely determined by $1 - x$). The function R assigns to each set of bets associated with w the investment schema $R(w) = x$.

If a betting agent wants to bail (not bet), they simply invest $(w, 1-w)$. This strategy guarantees them a gain of zero, no matter what the result of the toss is. For a betting agent who always wants to bail, $R(w) = w$ for all $w \in [0, 1]$.

R is in some way indicative of the agent's partial belief in the outcome of the coin toss. It may be indicative of other things as well, for example the risk tolerance of the agent or how much information the agent believes to have about the circumstances of the coin toss. For the purposes of this paper, the agent models the coin toss as the outcome of a random generator. This means that at least in the agent's model there is an objective chance c that the coin toss will land heads. The agent usually does not know this number and might have partial beliefs about it.

Let the cumulative credence function F for the objective chance c be

$$\mathrm{Cr}(c \leq y) = F(y) \tag{33}$$

where Cr is the credence (in this case that the objective chance c is smaller than or equal to y). Furthermore, assume that F is absolutely continuous so that there is a Lebesgue-integrable density function f with

$$F(y) = \int_{-\infty}^{y} f(u)\, du \tag{34}$$

Let $G(x, w)$ be the gain of investing x in an H-bet at w and $1 - x$ in a T-bet at $1 - w$. There are two values for $G(x, w)$, one if H is tossed, another if T is tossed.

$$G(x, w) = \left(\frac{x - w}{w}, -\frac{x - w}{1 - w} \right) \tag{35}$$

Proposition 2.2. The expected gain for an agent holding Cr and following the investment schema x upon being offered the set of bets associated with w is

$$E(x, w) = \frac{x - w}{w(1 - w)} \left(\int y f(y)\, dy - w \right) \tag{36}$$

Proof. Simplify

$$E(x,w) = \int f(y) \left(y \left(\frac{x}{w} - 1 \right) + (1-y) \left(-\frac{x-w}{1-w} \right) \right) dy \tag{37}$$

using $\int f(y)\, dy = 1$. □

$\int y f(y)\, dy$ is the agent's sharp credence in H. Clearly, $E(x,w)$ is maximized with $x = 1$ when the sharp credence is strictly greater than w; and with $x = 0$ when the sharp credence is strictly less than w.

3 Zillner Map for Revising the Geometry of Reason

Start off with Clinton/Trump and propriety. Restrict everything to finite. Introduce tetrahedron. Outline Pettigrew's symmetry argument. Cast doubt on symmetry and geometry of reason. (i) Reductio using Levinstein et al. (ii) The oddity of Pettigrew's use of Bronfman. Build up the case for the logarithmic measure. Summarize Shannon's and Kullback-Leibler's arguments. Show how the logarithmic measure generalizes Bayes and Jeffrey. But then there are oddities with the logarithmic measure as well. Solution: divorce yourself even more from geometry.

Make a list of desiderata, everything from the literature, in particular

symmetry see Savage for equivalence Brier score and symmetry; Ovcharov references Boissonnat for this fact—it would be interesting to see if it could be proved using Legendre duality

propriety uncontroversial

entropy have an entropy function that matches Shannon's requirements

locality make reward only depend on p_i (Goodness, more detailed description in Dawid, Lauritzen, Parry, page 594; Savage, page 794)

horizon horizon

reductio-resistance especially Levinstein etc.

Bronfman Pettigrew's argument against diversity

additivity a Pettigrew argument; both BS and LS fulfill it

Perhaps you can make a case that Brier and Log are the last two candidates standing. Pettigrew's case is that Brier is the only reasonable symmetric SR, whereas there is more diversity on the asymmetric side and therefore vulnerability to Bronfman.

Open questions:

- Can the Shannon entropy be extended to non-standard probabilities? Answer: No luck on google.
- In what sense is the logarithmic score a procedure that leads to MLE (Dawid et al, 594)? For an answer to this question see Gneiting and Raftery, pages 372ff.
- How can you fix things so that D_{KL} is the divergence function of the logarithmic score? Answer: McCarthy's theorem.
- The above procedure shows how you can move from scoring rules to entropy and divergence. What is the reverse procedure using gradients? Answer: McCarthy's theorem.
- Does BdF's argument go through if the distance function is a divergence? The answer to this question is in Theorem I.D.5 in `richard-pettigrew-chapter-08.p` page 13 (page 92 in the book); and the answer is yes. BdF's argument is valid for any additive Bregman divergence.
- Where is the symmetry argument? Can it be rephrased in terms of Legendre duality? The first question appears to be answered in Gneiting and Raftery, page 361: the symmetry argument is original to Savage 1971.
- What is the connection between the H of HB and entropy? Here is a comment in my annotation on Gneiting and Raftery, page 361: "G measures what I will score if Q is nature's chance function. I score best

when I predict Q and the SR is proper. The entropy function indicates my maximum reward if I correctly guess at Q. The more uncertainty the more reward. Why would this be true? Why would I get a higher reward for correctly guessing a more uncertain distribution?"

4 Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock: Bregman Voronoi Diagrams

This paper is inspirational, especially with respect to convex conjugates. For example, it seems to me you should be able to prove that symmetry and Brier score go together on the basis of convex conjugates. First, let me try to find the convex conjugates for BS and LS (Brier score and Log score). I have verified the following in SB pages 2441ff. For LS, the loss function is

$$\text{(LS) } S(\xi_i, x) = \ln x_i - \ln \sum x_k \tag{38}$$

on the positive orthant (convex cone containing $\mathcal{P}$). For BS, it is

$$\text{(BS) } S(\xi_i, x) = \frac{2x_i}{\sum x_k} - 1 - \sum_j \left(\frac{x_j}{\sum x_k}\right)^2 \tag{39}$$

The corresponding entropy functions are

$$\text{(LS) } H(x) = \sum_i x_i \ln \frac{x_i}{\sum x_k} \tag{40}$$

$$\text{(BS) } H(x) = \sum_i x_i \left(\frac{2x_i}{\sum x_k} - 1 - \sum_j \left(\frac{x_j}{\sum x_k}\right)^2\right) \tag{41}$$

I have verified using sage that for both LS and BS it is true that

$$\nabla H(x) = S(\xi, x) \tag{42}$$

where $S(\xi, x) = (S(\xi_1, x), \ldots, S(\xi_n, x))^\intercal$ and

$$\nabla H(x) = \left(\frac{\partial H}{\partial x_1}, \ldots, \frac{\partial H}{\partial x_n}\right)^\intercal \tag{43}$$

The scoring rule is the derivative of the entropy function. The entropy function generates a divergence via (see Boissonnat, 241)

$$D(p\|q) = H(p) - H(q) - \langle p - q, \nabla H(q)\rangle \tag{44}$$

where $\langle ., . \rangle$ is the dot product (or matrix multiplication if you are using dual spaces). In the case of LS, the divergence is the Kullback-Leibler divergence (verified in SB 2440f),

$$D_{\text{LS}}(p\|q) = \sum p_i \ln \frac{p_i}{\sum p_k} - \sum p_i \ln \frac{q_i}{\sum q_k} \tag{45}$$

In the case of BS, it is

$$D_{\text{BS}}(p\|q) = \sum p_i \left[\sum_j \left(\frac{q_j}{\sum q_k} - \delta_{ij} \right)^2 - \sum_j \left(\frac{p_j}{\sum p_k} - \delta_{ij} \right)^2 \right] \tag{46}$$

where δ_{ij} is the Kronecker delta. In either case, for LS or BS, the divergence is the difference between the expected loss of reported q when p is the true distribution and the entropy of p, which is the expected loss of reported p when p is the true distribution.

Define

$$F^*(\hat{x}) = \sup_{x \in \mathcal{X}} \{\langle \hat{x}, x\rangle - F(x)\} \tag{47}$$

Let $G(x) = \langle \hat{x}, x\rangle - F(x)$. Then the supremum is reached where the gradient of $G(x)$ vanishes.

$$\frac{\partial G}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\sum x_i \frac{\partial F}{\partial x_i} - F(x) \right) \tag{48}$$

5 Leonard Savage: Elicitation of Personal Probabilities and Expectations

The quarrel between Brier score and Log score roughly maps onto the quarrel between difference and ratio in the debate about evidence.

- If loss is a function of difference discrepancy $x - r$, then this condition is equivalent to symmetry and only linear variations of the Brier score fulfill it.
- If loss is a function of ratio discrepancy r/x, then this condition is equivalent to horizon and only linear variations of the Log score fulfill it.
- If, more generally, loss is a function of $g(r) - g(x)$, then only the Log score or the Brier score fulfill it (789).

Savage helpfully uses the terminology of the trichotomy (win, loss, tie). I put screen shots of Trump/Clinton predictions by nytimes/538 in the folder `barney/learning/dissertation/chapters/GeometryOfReason`.

6 Tilmann Gneiting and Matthias Katzfuss: Probabilistic Forecasting

Nice diagram of a probabilistic forecast on page 127 (Bank of England). The goodness of a probabilistic forecast is measured by the extent to which realizations are distinguishable from from random draws from predictive distributions (127). Propriety is equivalent to Bregman form is attributed to Savage 1971 on page 136.

7 Claude Shannon: Mathematical Theory of Communication

Go straight to section 6 on page 10. The proof is in appendix 2 on page 28.

1. H should be continuous in the p_i
2. If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Here is an example for the last requirement,

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right) = H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right) \tag{49}$$

The only H satisfying the three above assumptions is of the form

$$H = -K\sum_{i=1}^{n} p_i \log p_i \tag{50}$$

8 John McCarthy: Measures of the Value of Information

McCarthy uses the expression "keep the forecaster honest" (654).

Theorem 1. A payoff rule keeps the forecaster honest if and only if

$$f_i(q) = \frac{\partial}{\partial q_i} f(q) \tag{51}$$

is a convex function of q which is homogeneous of the first degree. $f_i q$ is the reward for the occurrence of the i-th event. This is terrible notation. $f_i(p)$ is $S(x_i, P)$ and $f(p)$ is $H(P)$. McCarthy omits the proof. The expectation of an honest forecaster is then

$$\sum p_i f_i(p) = f(p) \tag{52}$$

I.J. Good considered the problem of paying the forecaster with the restriction that $f_i(q) = F(q_i$, i.e. the payoff depends only on the probability assigned to the event which actually occurred. He showed that putting

$$F(x) = A \ln x + B \tag{53}$$

keeps the forecaster honest, and Gleason (unpublished) showed that this is the only $F(x)$ which does. (This is McCarthy verbatim, 654.)

I played this through for the Brier Score and the Log Score in Schmierbuch 2388f. Indeed,

$$\sum_j q_j \frac{\partial S(x_j, Q)}{\partial q_i} = 0 \tag{54}$$

for the Brier Score $S(x_i, Q) = 1 - 2q_i + \sum_j q_j^2$. However, for the Log Score $S(x_i, Q) = \ln q_i$ I get

$$\sum_j q_j \frac{\partial S(x_j, Q)}{\partial q_i} = 1 \tag{55}$$

McCarthy's theorem says (55) implies that the Log Score is not proper. What gives? Answer: Hendrickson and Buehler identify this problem on page 1918. The derivative needs to be taken with respect to D (the convex cone of $\mathcal{P}$), not $\mathcal{P}$.

Note that

$$\sum_j q_j \frac{\partial S(x_j, Q)}{\partial q_i} = 0 \tag{56}$$

looks suspiciously like the LHS in Euler's Homogeneous Function Theorem in

`http://mathworld.wolfram.com/EulersHomogeneousFunctionTheorem.html`

The intuitive content of the convexity restriction is that it is always a good idea to look at the outcome of an experiment if it is free (655).

9 Arlo D. Hendrickson and Robert J. Buehler: Proper Scores for Probability Forecasters

9.1 Inner Products

HB have an interesting suggestion. $p = (p_1, \ldots, p_n)$ (the credence function) is as usual an n-dimensional vector, but now so is $f(q) = (S(x_1, q), \ldots, S(x_n, q))$,

the score of q. This move turns the expectation of q (the prediction) with respect to true world p into an inner product $p \cdot f(q)$. Propriety demands

$$H(p) = p \cdot f(p) \geq p \cdot f(q) \tag{57}$$

for all p, q in order to discourage hedging. See the comedy of errors in section 11.1.

9.2 Definition 2.1 of Subgradients

A subgradient, in my understanding, is a linear real-valued function L which, if a real-valued function f is convex on a neighbourhood U of $x \in \mathbb{R}^n$, is less than f, i.e. $L(y) \leq f(y)$ on U. HB define as follows:

$$f(y) \geq (y - x)L(y) + f(x) \tag{58}$$

If all of this is one-dimensional, it's straightforward enough.

$$\frac{f(y) - f(x)}{y - x} \tag{59}$$

is the slope of the secant line, and for a differentiable, convex function $f : \mathbb{R} \to \mathbb{R}$ the only way (58) can be true is if $L(y) = f'(y)$. When $n \geq 2$, $L(y)$ is presumably the vector of partial derivatives and the product is the inner product. So, if f is convex and differentiable,

$$f(y) - f(x) \geq \sum (y_i - x_i) \frac{\partial f}{\partial y_i}(y) \tag{60}$$

I've always been puzzled that the slopes along the coordinate axes are simply added, but this appears to be true also for the product rule in higher dimensions, see (117). If f is differentiable, then the subgradient uniquely equals the gradient (see Rockafellar, theorem 25.1).

9.3 Euler's Theorem

Theorem 2.1 tells us that if a function f has a subgradient everywhere on a convex set D, then f is convex on D. When HB refer to Euler's Theorem, they mean Euler's Homogeneous Function Theorem, see

`http://mathworld.wolfram.com/EulersHomogeneousFunctionTheorem.html`

Theorem 2.2 is basically Euler's Theorem, but HB give an elegant proof. For the proof, note that

$$\frac{\lambda^r - 1}{\lambda - 1} = \lambda^{r-1} + \ldots + \lambda^1 + \lambda^0 \tag{61}$$

As $\lambda \to 1$, (61) goes to r because there are r summands.

9.4 Brier Score and Log Score

HB give a nice explanation why a superficial differentiation of the entropy function for the log score doesn't give you the log score, see page 1918.

I checked the following using pen and paper first, sage second. The Brier score $f : \mathbb{R}^n \to \mathbb{R}^n$ is for $I = \{1, \ldots, n\}$

$$f(x) = \left(2x_i - 1 - \sum_{k=1}^{n} x_k^2 \right)_{i \in I} \tag{62}$$

where $x \in \mathcal{P}$. It is interesting that in HB the scoring rule is restricted to the set of probability vectors $\mathcal{P}$. The entropy function $H : \mathbb{R}^n \to \mathbb{R}$ is defined on the convex cone $\mathcal{D}$. It is

$$H(y) = H(\lambda x) = \lambda H(x) = \lambda x \cdot f(x) \tag{63}$$

for $y \in \mathcal{D}, x \in \mathcal{P}$, and $\lambda = \sum y_i$. So defined, it is indeed true that

$$\frac{\partial H(y)}{\partial y_i} = f_i(y) = 2x_i - 1 - \sum_{k=1}^{n} x_k^2 \tag{64}$$

when $y \in \mathcal{P}$, but the derivative of H must be taken with respect to $\mathcal{D}$. For the log score, the corresponding functions are

$$f(x) = (\log x_i)_{i \in I} \tag{65}$$

and, indeed

$$\frac{\partial H(y)}{\partial y_i} = f_i(y) = \log y_i \tag{66}$$

There is some sage code commented in the source code of `supplement.tex`.

10 Gneiting and Raftery: Strictly Proper Scoring Rules, Rediction, and Estimation

Generally nice summary. Bregman divergences come into play once the sample space is finite and the entropy function is sufficiently smooth (361). More details, I believe, are in Predd et al (2009).

Note the use of hedging strategies on page 362.

11 Evgeni Ovcharov: Proper Scoring Rules and Bregman Divergences

11.1 A Comedy of Errors

It seems to me that there is a mistake in Ovcharov. Definition 2.2 says that a scoring rule S is called proper if it maximizes its expected score at the true density,

$$p \cdot S(p) = \max_{q \in \mathcal{P}} p \cdot S(q) \tag{67}$$

However, it seems to me that this should say

$$p \cdot S(p) = \max_{q \in \mathcal{P}} q \cdot S(p) \tag{68}$$

The truth needs to vary, not the prediction!

This is crazy. Ovcharov (5) and Hendrickson/Buehler (1916) seem to agree; but Dawid et al and Gneiting/Raftery have it my way. Let a chance-world be such that if it is known the chances are known but not the outcomes.

Ovcharov/Hendrickson/Buehler There is one true chance-world. I want to avoid that a forecaster will choose a forecast that is different from what she knows to be true. This is not realistic. Forecasters don't know which chance-world is true.

Dawid/Gneiting/Raftery There are many possible chance-worlds, an unknown one of which is true. I want to avoid that a forecaster will choose a forecast that does not reflect her best estimate. For example, a TV weather forecaster may deflate POP in order to boost TV ratings. In this case, the reward $p \cdot f(q)$ (p is the forecaster's best estimate, q is the deflated estimate) is greater than the reward $p \cdot f(p)$; which is what we don't want. But that's what Ovcharov/Hendrickson/Buehler are saying.

This is all a comedy of errors. The only person who's had this wrong is I. Both Ovcharov/Hendrickson/Buehler and Dawid/Gneiting/Raftery clearly state that the true world stays fixed and the prediction varies.

12 Leszek Wronski: Belief Update Methods and Rules

13 Dawid, Lauritzen, Parry: Proper Local Scoring Rules on Discrete Sample Spaces

Let $P = (p_1, \ldots, p_n)$ be a probability distribution over n mutually exclusive and collectively exhaustive events x_i. Let Q be the true (probability) distribution. Then $S(x_i, P)$ is the loss suffered by the agent espousing P when x_i is true. Define

$$S(Q, P) = \sum_{i=1}^{n} q_i S(x_i, P) \tag{69}$$

which is the expected loss of P with respect to Q. Dawid et al have nature first, forecaster second. Proper scoring rules require for $H(P) = \inf_{Q \in \mathcal{P}} S(Q, P)$

$$H(P) = S(P, P) \tag{70}$$

H is the entropy function. (70) encourages honesty. It prohibits the following scenario (which has been called 'hedging'). A forecaster's best guess at the

true probability distribution Q is P' (with $P' \neq P$), but she knows that $S(P', P) < S(P', P')$, so instead of asserting P' she asserts P.

The divergence function is

$$d(P, Q) = S(P, Q) - S(Q, Q) \tag{71}$$

Notice that strict propriety requires that $d(P, Q) > 0$ iff $P \neq Q$ (i.e. that d is positive-definite).

The Brier score is

$$S_B(x_i, P) = 1 - 2p_i + \sum_{j=1}^{n} p_j^2 \tag{72}$$

The summand 1 is sometimes omitted (as in Gneiting and Raftery, example 2.1, but not in Pettigrew). Consequently,

$$S_B(P, Q) = \sum_{i=1}^{n} q_i \left(1 - 2p_i + \sum_{j=1}^{n} p_j^2\right) = 1 - 2(P \cdot Q) + \sum_{i=1}^{n} p_i^2 \tag{73}$$

where $P \cdot Q$ is the dot product. Note that

$$\frac{\partial}{\partial p_k} = 2p_k - 2q_k \tag{74}$$

so that, indeed, $\inf_{Q \in \mathcal{P}} S(Q, P) = S(P, P)$.

This means that

$$H_B(P) = S_B(P, P) = 1 - \sum_{i=1}^{n} p_i^2 \tag{75}$$

Again, the summand 1 is omitted accordingly. The divergence function is the squared Euclidean distance

$$d_B(P, Q) = \sum_{i=1}^{n} (p_i - q_i)^2 \tag{76}$$

Now let's run through this for the logarithmic score (see Schmierbuch 2415).

$$S(x_i, P) = -\ln p_i \tag{77}$$

The entropy function is the Shannon entropy

$$H(P) = -\sum_{i=1}^{n} p_i \ln p_i \tag{78}$$

The divergence function is

$$d(P,Q) = S(P,Q) - S(Q,Q) = \sum_{i=1}^{n} (q_i - p_i) \ln q_i \tag{79}$$

which, disappointingly, is not the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(Q,P) = S(Q,P) - S(Q,Q) \tag{80}$$

Always remember that by convention the second argument of the Kullback-Leibler divergence is the prior probability.

The logarithmic score is local in the sense that $S(x_i, P)$ depends only on p_i. The Brier score is not local in this sense (see Dawid et al, 597f; I.J. Good addresses this as well).

14 Richard Pettigrew: Accuracy and the Laws of Credence

14.1 Measuring Accuracy: A New Account

14.1.1 Using Lagrange Multipliers to Find a Dominant Probability Distribution

Let $\mathcal{F}$ be an algebra of propositions. There may be other algebras of propositions, but it's easy to construct one using events that are logically unrelated in the sense that they do not entail one another, although they may be probabilistically dependent on each other. Let these events be $E_1, \ldots, E_n$. Then the algebra contains 2^{2^n} elements, 2^n atomic events of the form

$$\omega_i = (\neg)E_1 \cap \ldots \cap (\neg)E_n \tag{81}$$

for $i = 1, \ldots, 2^n$ and then combinations of the ω_i such as

$$r_j = \omega_3 \cup \omega_6 \cup \omega_{12} \tag{82}$$

for $j = 1, \ldots, 2^{2^n}$. If there is only one event A, the algebra is

$$\mathcal{F} = \{\emptyset, A, \neg A, \Omega\} \tag{83}$$

where $\emptyset$ is the conjunction of all other propositions in the algebra and Ω is the disjunction of all other propositions in the algebra.

My concern is that Joyce and Pettigrew only evaluate inaccuracy on atomic events. Here is my proof that for atomic events, there is indeed always a probability function which will dominate a credence function unless that credence function is itself a probability function. Let $m = 2^n$ be the number of atomic events and q a credence function defined on them with $\sum q_i = Q$. Then (beginning on page 2398 in Schmierbuch) I want to find a p with $\sum p_i = 1$ that dominates q, meaning that p's inaccuracy measured by the Brier score is less than q's inaccuracy, no matter which one of the ω_i is true. To find such a p, I look for the one that is closest to q. Define the Lagrangian

$$L(p) = \sum_{i=1}^{m} (p_i - q_i)^2 + \lambda \left(\left(\sum_{i=1}^{m} p_i \right) - 1 \right) \tag{84}$$

Differentiate with respect to each p_i and resolve the constraint for the following result

$$p_k = q_k - \frac{1}{m}(Q - 1) \tag{85}$$

It turns out that the difference $I(q, w) - I(p, w)$ is a function $J_{(m,w)}(Q)$ of the sum of the q_i's, but not the individual q_i's themselves. Let $w_i = 1$ iff ω_i is true and $w_i = 0$ iff ω_i is false. Recall that

$$I(q, w) = \sum_{i=1}^{m} (q_i - w_i)^2 \tag{86}$$

(although this is a simplification of the algebra that I will cast doubt on later). Then

$$J_{(m,w)}(Q) = \frac{1}{m}(Q - 1)^2 \tag{87}$$

This result depends on

$$\sum_{i=1}^{m} w_i = 1 \tag{88}$$

which is true because the atomic events are pairwise disjoint and their disjunction is Ω. (87) tells us that a dominating p can be found for each q that is not itself a probability function, i.e. $Q = 1$. It will be interesting to compare this proof to Pettigrew's proof. On page 80, he refers to theorems I.A.2, I.D.5, and I.D.7, which together establish theorem 4.3.4, and theorem 4.3.4 makes the above dominance claim.

14.1.2 Tetrahedrons and n-Simplices

Indeed, a couple of days later, it was very interesting! De Finetti uses convex hulls to show that probability functions and *only* probability functions are convex combinations of worlds. Consider two events A, B.

$$\begin{array}{rcl} \omega_0 & = & \neg A \cap \neg B \\ \omega_1 & = & \neg A \cap B \\ \omega_2 & = & A \cap \neg B \\ \omega_3 & = & A \cap B \end{array} \tag{89}$$

Let the r_j be as in (82), e.g. $r_0 = \emptyset, r_{14} = \neg(\neg A \cap \neg B), r_{15} = \Omega$. (I have been using $\neg$ for the set complement, which is a bit of notational mishmash between events and propositions.) Then the r_j are an exhaustive list of the propositions that you can form using A, B. There are four possible worlds, one for each ω_i to be true. I will write the vectors in matrix form to save space. Here are the four possible worlds, indicating whether r_j is false or true.

$$W_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{90}$$

$$W_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{91}$$

$$W_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{92}$$

$$W_3 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{93}$$

Zähneknirschend räume ich es ein, that de Finetti succeeds in showing that all and only probability functions are convex combinations of these four worlds,

$$p = \lambda W_3 + \mu W_2 + \nu W_1 + (1 - \lambda - \mu - \nu) W_0 \tag{94}$$

The probability distributions are a three-dimensional shape in a sixteen-dimensional space. More generally, for n events, the probability distributions form a $(2^n - 1)$-dimensional simplex in a (2^{2^n})-dimensional space. For two events, it is a tetrahedron in $\mathbb{R}^{16}$; for three events, it is a 7-simplex in $\mathbb{R}^{256}$. So, I need to get off the triangle/simplex bandwagon that I was on when I wrote the Geometry of Reason and get on the tetrahedron/n-simplex bandwagon!

BdF claims that for a convex set and an arbitrary point x, we can always find a point y that is an element of the convex set which is at least as close to all the other elements of the convex set as x is. Is this true when we use the logarithmic divergence measure?

14.1.3 Differential Equations for Bregman Divergences

Let D be a divergence with $D : [0,1]^n \times [0,1]^n \to [0,\infty]$ and $D(x,y) \geq 0$, equality iff $x = y$. Then D is an additive Bregman divergence iff the following conditions are met:

$$D(x,y) = \sum_{i=1}^{n} d(x_i, y_i) \tag{95}$$

where d is a one-dimensional divergence and generated by a smooth, bounded, strictly convex function f such that for all $a, b \in [0,1]$

$$d(a,b) = f(a) - f(b) - f'(b)(a-b) \tag{96}$$

$d(a,b)$ is effectively the difference at a between f and the first Taylor expansion of f with respect to b. Finding f for a given divergence d is a differential equation problem which I have posed on Math Stackexchange at `https://math.stackexchange.com/questions/2586331/ordinary-differential-equation-with-deviating-distributed-retarded-arguments`

Note that there is good material in Predd et. al. (2009) about this.

14.2 Bronfman Objection

This is where Pettigrew makes an argument for symmetry. It would be a good entry point for my criticism. It is an epistemic argument for symmetry in the sense that its validity depends on our current possibly deficient epistemic state.

Reading notes:

page 2 RP says No Drop is "if agent has opinion set $\{A, B\}$ and A entails B then rationality requires that $c(A) \leq c(B)$." However, rationality does not require that the agent knows that A entails B. Is the agent logically omniscient?

page 4 RP defines the squared Euclidean distance in terms of atomic propositions, but keeps saying all propositions. Which one is it?

page 26 The Brier score is proper; is the Kullback-Leibler Divergence proper?

page 17 Theorem 1.0.2 can't be right the way it stands because it privileges normalized probability functions.

ibidem High-level criticism, not just of Pettigrew, but also of Joyce and de Finetti. When they prove de Finetti's theorem, do they take into account that a credence function is completely free in choosing any credence for any proposition in the algebra, not just the atomic propositions? If there is one event E, then the algebra consists of 4 elements, $\{\emptyset, A, not - A, \Omega\}$. If there are two events A and B, then there are four atomic propositions which determine any probability distribution: A and B, A and not-B, not-A and B, not-A and not-B. The algebra consists of any combination of these atomic propositions connected by OR, for how to put this in code see definetti.jl. If n is the number of events under consideration, 2^n is the number of atomic propositions and $2^(2^n)$ is the number of propositions in the algebra. Probabilities are only free over the atomic propositions; credences are free over the whole algebra. There are several things I want to know: are there credences that are not Brier-score-dominated by a single probability distribution given this more complicated state of affairs? Does de Finetti's proof address this more complicated state of affairs? Are there credences (which must be positive to be measurable by DKL) that are not DKL-dominated by a single probability distribution at each possible world?

section 4.1 Pettigrew claims that there are no irreducible global features of inaccuracy. I agree. This, however, does not mean that global inaccuracy must be the sum of local inaccuracies. It only means that global inaccuracy is a function of local inaccuracies, and there are many other functions beside addition that would give us the local veritism that Wormwood wants.

page 51 Here is a problem with DKL and local inaccuracy. Let my credence that a coin toss is H be 0.52. The world w1 at which we measure my credence's local inaccuracy with respect to this proposition is such that H is true. Then a local version of inaccuracy for DKL is (?) $1 \cdot \log(1/0.52)$ or $-\log(0.52)$. So far so good, except that it subscribes to regularity in the sense that my inaccuracy is infinite if I have an extreme credence of zero. However, if at w2 not-H is true, then my

local inaccuracy is $0 \cdot \log(0/0.52)$, which is conventionally taken to be zero. I wouldn't have been inaccurate at all! Furthermore, my local inaccuracy for the logically equivalent proposition not-H, in which, say, I have credence 0.48, is completely different! Do these problems go away if I model credences as manifolds and apply the Fisher metric rather than modeling them using Cartesian coordinates and applying DKL? What Wormwood calls a one-dimensional divergence does not appear to be straightforward for logarithmic measures.

page 83 Why is $\mathcal{B}_{\mathcal{F}}$ restricted to credence functions that take values between 0 and 1? This may help me because DKL can't apply to negative numbers or zero. Pettigrew is consistent here with his definition of credence functions on page 16, Definition 1.0.1.

page 83 What is $\mathcal{W}_{\mathcal{F}}$? I am assuming it is a set of worlds which corresponds to whether the propositions in $\mathcal{F}$ are true or not. $\mathcal{F}$, according to Pettigrew, is not necessarily an algebra. A credence function c, however, is only a probability function if it can be extended to a credence function that is a probability function on an algebra. There is some circularity in P's definition of what a probability function is in Definition 1.0.1.

ibidem The meat is all in theorem I.A.2. Let's see if we can poke any holes. Here is the only place I can think of: if p and p' are probability functions, then so is $\lambda p + (1 - \lambda)p'$.

15 Bernhard Schutz: Geometrical Methods of Mathematical Physics

15.1 Some Basic Mathematics

On page 15, show that $d''(x, y)$ is not a norm.

15.2 Differentiable Manifolds and Tensors

On page 28, cover the interior of the annulus by a single coordinate patch.

15.2.1 Vectors and Vector Fields

Explain (2.2) on page 32. Schutz' λ is confusing to me. It appears to be a function from a closed interval $[a, b] \subset \mathbb{R}$ into the manifold M (a curve). f is $C^\infty(M)$, so $f : U \to \mathbb{R}$. g is just a coordinate map, what Jeff Lee calls x, and goes from U to $\mathbb{R}^n$. $g : U \to \mathbb{R}$. This makes no sense to me. What the hell is

$$\frac{dg}{d\lambda} \tag{97}$$

if g is defined on U and goes into $\mathbb{R}^n$?

Answer: you are mistaking figure 2.9 to apply to section 2.7 instead of section 2.6. g is not a coordinate map in section 2.7. It looks like the x^i determine the curve. In Jeff Lee's notation, let $\gamma : (a, b) \to V$ ((V, y) is the usual open neighbourhood in M plus coordinate map pair). Then Schutz'

$$(x^1(\lambda), \ldots, x^n(\lambda)) \tag{98}$$

corresponds to Lee's $(y \circ \gamma)(\lambda)$ in $\mathbb{R}^n$. Schutz' $f : \mathbb{R}^n \to \mathbb{R}$ corresponds to

$$\hat{f} : \xi \in \mathbb{R}^n \to f(y^{-1}(\xi)) \tag{99}$$

in Lee. I derived (2.2) in Schutz with a complicated translation table between Lee and Schutz on page 2256 in Schmierbuch. Write g in Schutz instead as $\hat{f} \circ x$, where $\hat{f}$ is as defined in (99) and x is as defined in (98), unfortunately mixing Schutz and Lee, but (2.2) pretty handily follows.

Here is an example for the chain rule and multivariable differentiation. Let

$$\begin{aligned} f(x) &= \left(x^2, \tfrac{x}{2}\right) \\ g(w, z) &= 3w - z^2 \end{aligned} \tag{100}$$

Then $(g \circ f)(x) = \frac{11}{4}x^2$ and therefore $(g \circ f)'(x) = \frac{11}{2}x$. According to the chain rule

$$\frac{d(g \circ f)}{dx} = \sum \frac{df^i}{dx} \frac{\partial (g \circ f)}{\partial f^i} = 2x \cdot 3 + \frac{1}{2} \cdot \left(-2 \cdot \frac{x}{2} \right) = \frac{11}{2} x \tag{101}$$

This is what is behind Schutz' equation (2.2) on page 32.

15.2.2 Exercise 2.1

On page 44.

16 Michael K. Murray and John W. Rice: Differential Geometry and Statistics

Note that there is an informative article on exponential families and categorical distributions in Wikipedia. It suggests as parameters for the categorical distribution

$$\left(\ln \frac{p_1}{p_k}, \ldots, \ln \frac{p_{k-1}}{p_k}, 0 \right)^{\intercal}. \tag{102}$$

I wonder if the last parameter is necessary, since it's always zero.

Let P be the parametric family of normal distributions. Let $q = \mathcal{N}(0, 1)$ (the origin) and $p = \mathcal{N}(7, 9)$. The density for the normal distribution is

$$p(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(\frac{-(x - \mu)^2}{2\sigma^2} \right) \tag{103}$$

Expressed as an exponential family, this is

$$p(\theta^1, \theta^2) = \exp \left(\theta^2 x^2 + \theta^1 x^1 - K(\theta) \right) \tag{104}$$

x^2 is a second random variable (using differential geometry's habit of indexing vector components on the top right), but in the case of the normal distribution it is the first random variable squared, which makes for a hell of notational confusion.

For q and p as above, the parameters are $(-0.5, 0)$ and $(-1/18, 7/9)$, respectively. The origin q is not at $(0, 0)$, which also creates confusion. The vector $\vec{w}$ associated with p, for example, so that

$$q \oplus \vec{w} = p \tag{105}$$

where $q \oplus \vec{w} = exp(\vec{w}x)q(x)$ (x is (x^1, x^2)), is

$$\vec{w^{\intercal}} = \left(\frac{4}{9}, \frac{7}{9}\right). \tag{106}$$

However, $q \oplus \vec{w} = p^*$ such that $[p^*] = [p]$, where $[p]$ is the equivalence class of measures up to scale. In other words, p^* is p once it is normalized (2153).

For the normal distribution,

$$\theta^1 = -\frac{1}{2\sigma^2}, \qquad \theta^2 = \frac{\mu}{\sigma^2}, \qquad K(\theta) = \frac{1}{2}\ln\left(-\frac{\pi}{\theta^1}\right) - \frac{(\theta^2)^2}{4\theta^1} \tag{107}$$

On (2155f) I wonder if probability distributions/densities generally wouldn't be better expressed by equivalence classes of measures rather than normalized measures. The following appear to be privileged once you use the exponential parameters rather than simply the probabilities of the categories:

- sufficient statistics
- conjugate priors
- maximum entropy derivation (see Wikipedia)

The log-likelihood of $\mathcal{N}(0,1)$ is 0 because

$$q = q \oplus 0 = e^0 q \tag{108}$$

The log-likelihood of $\mathcal{N}(7,9)$ is

$$\ln \frac{1}{3} + \frac{1}{18}(8x^2 + 14x - 49) \tag{109}$$

Murray and Rice appear to get this wrong, assuming that the coordinates of the origin q are $(0,0)$ so that the log-likelihood of p would be $\theta_p x - K(\theta_p)$, when really it is $\alpha_p x - K(\theta_p)$ with $\theta_q + \alpha_p = \theta_p$ (2160).

Murray and Rice make a momentous claim on page 16:

> Hence P is an exponential family if and only if $\tilde{P}$ is an affine subspace of $\mathcal{M}$.

I could not track with this claim and started reading Jeff Lee's book in order to understand this better (beginning of August, 2017; notes for Murray and Rice end on (2171) and begin for Lee on (2172)). I should note (January 2018) that Boissonnat, Nielsen, and Nock make the following claim: Banerjee et al have shown that there is a bijection between the regular exponential families in statistics and a subset of the Bregman divergences called regular Bregman divergences (285).

17 Jeffrey M. Lee: Manifolds and Differential Geometry

17.1 The Tangent Structure (Chapter 2)

17.1.1 A Multivariable Calculus Exercise

Here is a nice exercise for my multivariable calculus students. Consider the vector

$$\vec{OP} = (0.2, 0.3, \sqrt{0.87})^\intercal \tag{110}$$

P is on the unit sphere. Provide the equation of the tangent plane. There are two ways to do this (2179):

- use partial derivatives m_x, m_y and the plane equation $z = z_0 + m_x(x - x_0) + m_y(y - y_0)$
- use the dot product and the fact that $\vec{OP}$ is orthogonal to the tangent plane

17.1.2 Lemma 2.4

Showing $(f \circ \gamma_v)'(0) = (f \circ c)'(0)$ on page 59 was a bit of a challenge. Interestingly, understanding this is important for solving exercise 2.16 on page 67 later on. On (2178) I am still carrying on with the assumption on Lee, page 56, that this is the special case of a submanifold of $\mathbb{R}^N$, but given Lee's talk about the "general setting" on the bottom of page 57, this must be incorrect. Show that

$$\left.\frac{d}{dt}\right|_{t=0} (f \circ x_\alpha^{-1} \circ h)(t) = D(f \circ x_\alpha^{-1})(x_\alpha(p)) \cdot \vec{v} \tag{111}$$

with $x_\alpha(p) = \vec{q} \in \mathbb{R}^n$, $p \in \mathcal{M}$, and $h(t) = x_\alpha(p) + t \cdot \vec{v}$, $\vec{v} \in \mathbb{R}^n$. f is a smooth real-valued function defined on $\mathcal{M}$, which has a chart (U_α, x_α). Use the chain rule,

$$D(f \circ x_\alpha^{-1} \circ h)(0) = D(f \circ x_\alpha^{-1})(h(0))D(h)(0) = D(f \circ x_\alpha^{-1})(x_\alpha(p)) \cdot \vec{v} \tag{112}$$

and also

$$(f \circ c)'(0) = (f \circ x_\alpha^{-1} \circ x_\alpha \circ c)'(0) = D(f \circ x_\alpha^{-1})(x_\alpha(p))(x_\alpha \circ c)'(0) \tag{113}$$

(see (2207)).

17.1.3 The Leibniz Law

For the use of the Leibniz Law on page 61 note that $C^\infty(\mathcal{M})$ is the set of real-valued, smooth functions on $\mathcal{M}$. Think of tangent spaces as a straightening of the manifold at a point. $T_p\mathcal{M}$ (kin) is the straightforward geometric interpretation, $T_p\mathcal{M}$ (phys) uses transformation laws, and $T_p\mathcal{M}$ (alg) uses linear functions which fulfill the Leibniz Law. Either way it's conversion therapy for manifolds: straighten them out at a point p by using the linear functions provided by differentiation.

17.1.4 Theorem 2.10

There is a proof on page 63 in Lee that

$$\left(\left.\frac{\partial}{\partial x^1}\right|_p, \ldots, \left.\frac{\partial}{\partial x^n}\right|_p \right) \tag{114}$$

is a basis for $T_p\mathcal{M}$ (alg). First, note that g needs to be defined all along u from 0 to u, thus the remark about the convex set. Next, let's have a look at the claim that

$$g = g(0) + \sum g_i u^i \tag{115}$$

based on the FTC. Let $g : x(U) \to \mathbb{R}$ be smooth. Let $h : \mathbb{R} \to x(U)$ be defined by $h(t) = t\vec{u}$. Then

$$h'(t) = \begin{pmatrix} u^1 \\ \vdots \\ u^n \end{pmatrix}^\intercal \tag{116}$$

and

$$(g \circ h)'(t) = g'(h(t)) \cdot h'(t) \tag{117}$$

using matrix multiplication. Thus

$$(g \circ h)'(t) = \sum_{i=1}^{n} \frac{\partial g}{\partial u^i}(h(t))u^i \tag{118}$$

and by the FTC

$$\int_0^1 (g \circ h)'(t)\, dt = g(u) - g(0). \tag{119}$$

It follows that

$$\int_0^1 (g \circ h)'(t) dt = \int_0^1 \sum_{i=1}^{n} \frac{\partial g}{\partial u^i}(h(t))u^i\, dt =$$

$$\sum_{i=1}^{n} \left(\int_0^1 \frac{\partial g}{\partial u^i}(t\vec{u})\, dt \right) u^i = \sum_{i=1}^{n} g_i(u)u^i = g(u) - g(0). \tag{120}$$

17.1.5 Exercise 2.16

Show that $T_p f(\beta w_p) = \beta T_p f(w_p)$. Let $[c] = w_p$. Let (U_α, x_α) be an arbitrary chart of $\mathcal{M}$ and $(\tilde{U}_\alpha, \tilde{x}_\alpha)$ an arbitrary chart of $\mathcal{N}$. By definition (see Lee, page 58f),

$$T_p f(\beta w_p) = [f \circ \gamma_1] \text{ and } \beta T_p f(w_p) = [\gamma_2] \tag{121}$$

with $\gamma_1(t) = (x_\alpha^{-1} \circ \phi_1)(t)$ and $\gamma_2(t) = (\tilde{x}_\alpha^{-1} \circ \phi_2)(t)$.

$$\phi_1(t) = x_\alpha(p) + t\beta(x_\alpha \circ c)'(0) \tag{122}$$

$$\phi_2(t) = \tilde{x}_\alpha(q) + t\beta(\tilde{x}_\alpha \circ f \circ c)'(0) \tag{123}$$

(see box A on (2201) and box B on (2203)). Show that for all $g \in C^\infty(\mathcal{N})$

$$(g \circ f \circ \gamma_1)'(0) = (g \circ \gamma_2)'(0) \tag{124}$$

Use the chain rule to separate ϕ_1 and ϕ_2, so the claim reduces to LHS=RHS for

$$\text{LHS} = D(g \circ f \circ x_\alpha^{-1})(x_\alpha(p)) \circ D(\phi_1)(0) \tag{125}$$

$$\text{RHS} = D(g \circ \tilde{x}_\alpha^{-1})(x_\alpha(f(p))) \circ D(\phi_2)(0) \tag{126}$$

with

$$D(\phi_1)(0) = \beta D(x_\alpha \circ c)(0) \tag{127}$$

and

$$D(\phi_2)(0) = \beta D(\tilde{x}_\alpha \circ f \circ c)(0). \tag{128}$$

Now reconstitute (undo the chain rule) for

$$\text{LHS} = \beta D(g \circ f \circ x_\alpha^{-1} \circ x_\alpha \circ c)(0) \tag{129}$$

$$\text{RHS} = \beta D(g \circ \tilde{x}_\alpha^{-1} \circ \tilde{x}_\alpha \circ f \circ c)(0) \tag{130}$$

which are obviously equal to each other (see (2211)). To show that $T_p f(v_p + w_p) = T_p f(v_p) + T_p f(w_p)$ use the same procedure. Let $v_p = [c_v]$ and $w_p = [c_w]$. Then you need to show that

$$(g \circ f \circ \gamma_1)'(0) = (g \circ \gamma_2)'(0) \tag{131}$$

where

$$\gamma_1(t) = x_\alpha^{-1}(x_\alpha(p) + t((x_\alpha \circ c_v)'(0) + (x_\alpha \circ c_w)'(0))) \tag{132}$$

and

$$\gamma_2(t) = \tilde{x}_\alpha^{-1}(\tilde{x}_\alpha(f(p)) + t((\tilde{x}_\alpha \circ f \circ c_v)'(0) + (\tilde{x}_\alpha \circ f \circ c_w)'(0))) \tag{133}$$

Use the distributive law of matrix multiplication to show (131) (see (2213)). Now show that $T_p f$ is a linear isomorphism if f is a diffeomorphism. For surjectivity, let $[\hat{c}] \in T_q\mathcal{N}$. Define $c : t \to (f^{-1} \circ \hat{c})(t)$. Then $[f \circ c] = [\hat{c}]$. For injectivity, let c_1, c_2 be such that $[f \circ c_1] = [f \circ c_2]$ and let $g : \mathcal{M} \to \mathbb{R}$ be smooth. Then $g \circ f^{-1}$ is a smooth function from $\mathcal{N} \to \mathbb{R}$ and therefore

$$(g \circ c_1)'(0) = (g \circ f^{-1} \circ f \circ c_1)'(0) = (g \circ f^{-1} \circ f \circ c_2)'(0) = (g \circ c_2)'(0) \tag{134}$$

which establishes $[c_1] = [c_2]$ (for this proof see (2215)). □

17.1.6 Definition 2.18

On page 68, Lee claims that the definition of $T_p\mathcal{M}(v_p)$ is independent of the representative of v_p. Let $(p, v_1, (U, x_1)), (p, v_2, (U, x_2))$ be two representatives of v_p. Let (V, y) be an arbitrary chart for $\mathcal{N}$ (I suppose we should also show independence of this chart for $\mathcal{N}$). Then $T_p\mathcal{M}(v_p)$ is defined to be represented by $(q, w, (V, y))$ with

$$w = D(y \circ f \circ x^{-1})\big|_{x(p)} \cdot v_p \tag{135}$$

(this is a correction of an omission on Lee's part). The two representatives corresponding to the two charts for $\mathcal{M}$ are $(q, w_1, (V, y))$ and $(q, w_2, (V, y))$ with

$$w_1 = D(y \circ f \circ x_1^{-1}) \cdot v_1 \tag{136}$$

$$w_2 = D(y \circ f \circ x_2^{-1}) \cdot v_2 \tag{137}$$

We need to show that

$$w_2 = D(y \circ y^{-1})\big|_{y(q)} \cdot w_1. \tag{138}$$

You show this by taking $D(y \circ y^{-1} \circ y \circ f \circ x_1^{-1} \circ x_1 \circ x_2^{-1})\big|_{x_2(p)} \cdot v_2$ and using (136) and (137) on the one hand, the chain rule on the other hand (see (2216f)).

17.1.7 Definition 2.20

A word about the definition of differentials on page 69. Using derivations, a tangent vector v_p is a linear map from $C^\infty(\mathcal{M})$ into the real numbers obeying

Leibniz's Law. The space of these linear maps is $T_p\mathcal{M}$. Consider a function $f \in C^\infty(\mathcal{M})$. Now define a linear map df assigning to each v_p the real number $v_p \cdot f$. df is a function from $T_p\mathcal{M}$ to the real numbers and depends on f and p.

Consider the dual space (the cotangent space) of $T_p\mathcal{M}$. It is defined as the vector space spanned by $(v_1^*, \ldots, v_n^*)$, where $(v_1, \ldots, v_n)$ is a basis for $T_p\mathcal{M}$ and v_i^* is a function from $T_p\mathcal{M}$ to the real numbers with $v_i^*(v_j) = \delta_{ij}$. Lee's claim is that $df \in T_p^*\mathcal{M}$.

Now look at this in basic calculus. The differential dy is defined as a function assigning to each dx the real number $f'(x)dx$. Think of it as a linear function from the tangent space of f at point a into the real numbers. dx is usually identified with Δx, but now you can see the subtle difference between Δx and dx. If $f : \mathbb{R}^2 \to \mathbb{R}$, then dx is a vector in $\mathbb{R}^2$ and gets mapped to

$$dy = df(p)(dx) = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 \tag{139}$$

where dx_1 is the x-component of dx and dx_2 is the y-component. The z-component of dx is dy by definition (see the diagram on (2223)). For (139), see (2222). The upshot here is that dx is a vector in $T_p\mathcal{M}$, while Δx is a vector in the xy-plane, and we can generalize

$$dy = f'(x)dx \tag{140}$$

to

$$dy = df(p)(dx) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i \tag{141}$$

where the last expression is only defined for $f : \mathbb{R}^n \to \mathbb{R}$. Now let's see if Lee is correct with his claim that $df \in T_p^*\mathcal{M}$ and that the generalization in (141) holds.

We have a vector space $T_p\mathcal{M}$. Therefore, there is a dual vector space $T_p^*\mathcal{M}$. We know (theorem 2.10) that $T_p\mathcal{M}$ has the basis

$$\left(\frac{\partial}{\partial x^1}\bigg|_p, \ldots, \frac{\partial}{\partial x^n}\bigg|_p \right) \tag{142}$$

The basis for $T_p^*\mathcal{M}$ is $(v_1^*, \ldots, v_n^*)$ such that

$$v_i^* \left(\frac{\partial}{\partial x^j}\bigg|_p \right) = \delta^{ij} \tag{143}$$

If (U, x) is a chart, then the differentials of the coordinate functions $x^1, \ldots, x^n$ fulfill (143) because

$$dx^i\big|_p \left(\frac{\partial}{\partial x^j}\bigg|_p \right) = \frac{\partial x^i}{\partial x^j}(p) = \delta^{ij} \tag{144}$$

From now on, I will follow Lee and write M and N for manifolds, rather than $\mathcal{M}$ and $\mathcal{N}$.

17.1.8 Another Remark on Definition 2.21

Let T_pf be a tangent map from $T_pM \to T_{f(p)}\mathbb{R}$ and $df(p)$ be a differential of f at p. Lee's claim is that T_pf and $df(p)$ basically do the same thing once you identify $T_{f(p)}\mathbb{R}$ with $\mathbb{R}$ using the natural isomorphism on page 66. Therefore, a differential is just a special tangent map into $T_{f(p)}\mathbb{R}$ (identified with $\mathbb{R}$).

To see why this is so note that $T_pf : T_pM \to T_{f(p)}\mathbb{R}$ is the map that sends v_p to $T_pf(v_p)$, a linear map which sends any real-valued function g to the real number

$$(T_pf(v_p))(g) = v_p(g \circ f) \tag{145}$$

Note also that $df(p)$ is a real-valued map sending v_p to $v_p f$. This means we have succeeded once $T_p f(v_p)$ is identified with $v_p f$ via the natural isomorphism $\jmath$. Consider the real number $v_p f$.

$$\jmath(v_p f) = [\hat{c}] \tag{146}$$

where

$$\hat{c}(t) = f(p) + t \cdot (v_p f) \tag{147}$$

Once we show that $[\hat{c}] = T_p f(v_p)$ we are done. Note that $[\hat{c}] \in T_{f(p)}\mathbb{R}$.

Let g be a real-valued function in $C^\infty(M)$ and $[c] = v_p$. Use the interpretation on page 65, the chain rule, and (147) to show that

$$(T_p f(v_p))(g) = v_p(g \circ f) = \left.\frac{d}{dt}\right|_{t=0} (g \circ f \circ c) =$$

$$g'(f(p))f'(p)c'(0) = g'(f(p))(v_p f) = \left.\frac{d}{dt}\right|_{t=0} (g \circ \hat{c}) \tag{148}$$

17.1.9 Definition 2.21

Consider Lee's comment on the bottom of page 69. Let γ be a smooth curve from a real neighbourhood of 0 to $[a, b] = I$ such that

$$[\gamma] = \left.\frac{\partial}{\partial u}\right|_{t_0} \tag{149}$$

According to the interpretation of page 65 (identifying kinematic and algebraic tangent spaces),

$$\left.\frac{d}{dt}\right|_{t=0} f \circ \gamma = \left(\left.\frac{\partial}{\partial u}\right|_{t_0}\right)(f) \tag{150}$$

Let $f = \mathrm{id}$ for

$$\left.\frac{d}{dt}\right|_{t=0} \gamma = \left.\frac{\partial \mathrm{id}}{\partial u}\right|_{t_0} = 1 \tag{151}$$

Now use the same interpretation in reverse for

$$\dot{c}(t_0) \cdot f = \left(T_{t_0} c \left(\left.\frac{\partial}{\partial u}\right|_{t_0} \right) \right) (f) = \left.\frac{d}{dt}\right|_{t=0} f \circ (c \circ \gamma) =$$

$$\left.\frac{d}{dt}\right|_{t=t_0} f \circ c \cdot \left.\frac{d}{dt}\right|_{t=0} \gamma = \left.\frac{d}{dt}\right|_{t=t_0} f \circ c \tag{152}$$

as claimed by Lee. This is a nice example how a mixed use of algebraic and kinematic tangent spaces can be reconciled using the interpretations of page 65. I suppose an alternative way of doing this is to use the algebraic definition of a tangent map $T_p f$ instead (definition 2.19).

17.1.10 Exercise 2.23

$T_p f = 0$ for all $p \in M$ means that for any smooth $g : N \to \mathbb{R}$ and for any $v_p = [c]$ it is true that

$$v_p(g \circ f) = 0 \tag{153}$$

Let p_1 and p_2 be two arbitrary connected points in M such that a curve $c : I \to M$ connects them with $c(a) = p_1, c(0) = p, c(b) = p_2, a < 0 < b$. Then (153) is true for v_p and therefore

$$v_p(g \circ f) = \left.\frac{d}{dt}\right|_{t=0} (g \circ f \circ c) = 0 \tag{154}$$

Define $h(t) = (g \circ f \circ c)(t)$ for an arbitrary g. Claim: $h'(t) = 0$ for all t in the open interval (a, b). Assume that $h'(t_0) \neq 0$. Then define

$$\phi(t) = r_2 t^2 + r_1 t + t_0 \tag{155}$$

with

$$r_1 = \frac{b^2(a - t_0) - a^2(b - t_0)}{ab(b - a)} \tag{156}$$

$$r_2 = \frac{a(b - t_0) - b(a - t_0)}{ab(b - a)}. \tag{157}$$

Then $\phi(a) = a, \phi(0) = t_0, \phi(b) = b$. Also,

$$\phi'(0) = r_1 \neq 0 \tag{158}$$

because $b^2(a - t_0) \neq a^2(b - t_0)$ ($a - t_0$ is the only negative term). Let $\hat{c} = c \circ \phi$ and

$$(g \circ f \circ \hat{c})'(0) = 0 \tag{159}$$

for the same reason as (154). However

$$(g \circ f \circ \hat{c})'(0) = (g \circ f \circ c \circ \phi)'(0) = (g \circ f \circ c)'(\phi(0)) \cdot \phi'(0) \tag{160}$$

and by assumption and (158) both of these factors are not equal to zero. The contradiction gives us the claim that $h'(t) = 0$ for all $t \in (a, b)$. According to the Fundamental Theorem of Calculus, there is a $\hat{s}$ with $a \leq \hat{s} \leq b$ giving us (together with our claim)

$$h(b) - h(a) = h'(\hat{s})(b - a) = 0 \tag{161}$$

Therefore $(g \circ f)(p_1) = (g \circ f)(p_2)$ for arbitrary g. This is only possible if $f(p_1) = f(p_2)$. Therefore, f is locally constant (2229–2237). □

17.1.11 Theorem 2.25

Lee claims that under the assumptions of Theorem 2.25 M and N have the same dimension. Remember that

$$\left(\left.\frac{\partial}{\partial x^1}\right|_p, \ldots, \left.\frac{\partial}{\partial x^n}\right|_p \right) \tag{162}$$

is a basis for T_pM. Therefore, $\dim(M) = \dim(T_pM)$ and $\dim(N) = \dim(T_qN)$. If there is a linear isomorphism between two vector spaces V and W, then $\dim(V) = \dim(W)$. For a proof by S.F. Ellermeyer, see

`http://ksuweb.kennesaw.edu/~sellerme/sfehtml/classes/math3260/isomorphicvectorspaces.pdf`

Since T_pf is such a linear isomorphism between T_pM and T_qN, it follows that $\dim(M) = \dim(N)$. □

Notice how Exercise 2.16 and Theorem 2.25 are related. I needed Exercise 2.16 for the claim that T_pf is a *linear* isomorphism. $D(y \circ f \circ x^{-1})(x(p))$ is a linear isomorphism because it is one of the definitions (Definition 2.18) of a tangent map T_pf, and T_pf is a linear isomorphism by assumption. If $y \circ f \circ x^{-1}$ is injective, then f is also injective. Let $\pi_i = x^{-1}(p_i)$ and $f(p_1) = f(p_2)$. Then $\pi_1 = \pi_2$ and therefore $p_1 = p_2$. Restricting the codomain of f to $f(O)$ means that $f|_O$ is a diffeomorphism on O.

17.1.12 Lemma 2.28 Partials Lemma

My work is on (2329–2332). (v, w) is prima facie not a tangent vector. It becomes one by identification with $(T\imath^q + T\imath_p)(v, w)$, see Lee, page 73. $T\imath^q$

is not the inverse of $T_{(p,q)}\text{pr}_1$, as the diagram near the middle of page 73 in Lee suggests. It is the tangent map with respect to the insertion function $\imath^q$. Insertion $\imath^q$ and stripping $T_{(p,q)}\text{pr}_1$ are not inverse to each other. The validity of the lemma is now easily demonstrated.

$$(T_{(p,q)}f(v,w))(g) \overset{(1)}{=} (v,w)(g\circ f) \overset{(2)}{=} v(g\circ f\circ \imath^q) + w(g\circ f\circ \imath_p) \overset{(3)}{=}$$

$$(\partial_1 f_{(p,q)}v)(g) + (\partial_2 f_{(p,q)}w)(g) \tag{163}$$

(1) is true based on the definition of tangent maps (definition 2.19 on page 68) with respect to f. (2) is true based on the identification referred to above and the definition, again, of tangent maps, this time with respect to $\imath^q$ and $\imath_p$. (3) is true based on the definition of partial tangent maps on page 72. □

17.1.13 Definition 2.30

I needed help from math stackexchange to understand why a countable union of zero measure sets, under the given description, has measure zero. See Anthony Peter's answer to questin 1156907 in math stackexchange. The idea is to pick

$$\epsilon_i = \frac{\epsilon}{2^i} \tag{164}$$

for

$$A = \bigcup_{i\in\mathbb{N}} A_i. \tag{165}$$

Then the sum of the cube volumes will not exceed ϵ, even though there are countably many of these sums.

17.1.14 Morse Lemma 2.5.1

Lee makes the claim that $df|_{p_e} = T_{p_e}f = 0$. Definition 2.20 tells us that $df|_{p_e}$ is a function $T_{p_e}M \to \mathbb{R}$ which maps v_{p_e} to $v_{p_e}f$. To make sense of Lee's claim, $v_{p_e}f$ (which is in $\mathbb{R}$) needs to be identified with $u_{f(p_e)}$ in $T_{f(p_e)}\mathbb{R}$.

We use the identification procedure on page 66. Let $g \in C^\infty U$ defined on a neighbourhood $U \subset \mathbb{R}$ containing $f(p_e)$. Then

$$u_{f(p_e)}(g) = \left.\frac{d}{dt}\right|_{t=0} g(f(p_e) + t(v_{p_e} f)) = g'(f(p_e))(v_{p_e} f) \tag{166}$$

Note that

$$g'(f(p_e))(v_{p_e} f) = \left.\frac{d}{dt}\right|_{t=0} g \circ \gamma \tag{167}$$

where $[\gamma] = u_{f(p_e)}$ and (167) is true because of the kinematic interpretation we can give to the algebraic version of a tangent vector (see page 65). Since $T_{p_e} f$ sends v_{p_e} to $[f \circ c]$, where $v_{p_e} = [c]$, $T_{p_e} f$ and $df|_{p_e}$ are equal via the identification $\mathbb{R} \leftrightarrow T_{f(p_e)}\mathbb{R}$ if

$$(g \circ \gamma)'(0) = (g \circ f \circ c)'(0) \tag{168}$$

for all smooth $g : U \to \mathbb{R}$. (168) is equivalent to

$$g'(f(p_e))(v_{p_e} f) = g'(f(p_e))(f \circ c)'(0) \tag{169}$$

which is true if $v_{p_e} f = (f \circ c)'(0)$, but that is just the formula for setting up the interpretation between the algebraic and kinematic version of the tangent space $T_{p_e} M$ (see page 65). Now we need to show that $T_{p_e} f = 0$ (again via the identification $\mathbb{R} \leftrightarrow T_{f(p_e)}\mathbb{R}$). 0 corresponds to $[\hat{\gamma}]$ with

$$\hat{\gamma}(t) = f(p_e) + t \cdot 0 = f(p_e) \tag{170}$$

a constant function for which $\hat{\gamma}'(0) = 0$. The claim $T_{p_e} f = 0$ is true if and only if

$$(g \circ f \circ c)'(0) = (g \circ \hat{\gamma})'(0) = 0 \tag{171}$$

Since we also have

$$(g \circ f \circ c)'(0) = g'(f(p_e)) \cdot (f \circ c)'(0) \tag{172}$$

by the chain rule, (171) is true if

$$(f \circ c)'(0) = 0 \tag{173}$$

This holds because $f \circ c$ has an extremum (oBdA a maximum) at 0 and

$$0 \leq \lim_{h \to 0^-} \frac{(f \circ c)(h) - (f \circ c)(0)}{h} = \lim_{h \to 0} \frac{(f \circ c)(h) - (f \circ c)(0)}{h} =$$
$$(f \circ c)'(0) = \lim_{h \to 0^+} \frac{(f \circ c)(h) - (f \circ c)(0)}{h} \leq 0 \tag{174}$$

17.1.15 Exercise 2.37

This really needs elucidation from Problem 13 (which, by the way, is at the end of chapter 2 in the exercises section) and section 2.8, vector fields. Lie derivatives should be very helpful in showing $X_p(Yf) = Y_p(Xf)$.

The first problem is to show that the Hessian $[H_{f,p}]_x$ is invertible if and only if $[H_{f,p}]_{\hat{x}}$ is invertible, where $\hat{x}$ is an alternative coordinate system on M. I needed to learn a lot about the multivariable second-derivative test and the associated linear algebra before I could tackle this question. Williamson and Trotter in *Multivariable Mathematics* give a nice overview of why the second derivative provides conditions for extrema, especially

$$g(x) = g(0) + g'(0)x + \int_0^x (x-t)g''(t)\,dt \tag{175}$$

for $g(x) = f(x_0 + xu)$, where u is an arbitrary unit vector. Use partial integration to see that (175) is true; substitute f in (175) to see that f has a minimum where $\partial^2 f/\partial u^2$ is positive. (It is unclear to me how Theorem 4.4 in WT relates to Theorem 4.5 and 4.6, this would be worth pursuing.)

The linear algebra bit is in R.J. Goult, *Applied Linear Algebra*, a nondescript little brown book I have on my bookshelf. Section 3.5 about quadratic forms provides the connection between multivariable second derivatives and matrices, sections 5.3 and 5.4 provide background with respect to eigenvalues and eigenvectors of real symmetric matrices, which Hessians are.

Let $f : M \to \mathbb{R}$ and x, y be two coordinate functions for a neighbourhood of $p_e \in M$. p_e is a critical point for f, so df at p_e is zero. Now let the Hessian $[H_{f,p_e}]_x$ be singular. Show that $[H_{f,p_e}]_y$ is also singular. This establishes that nondegeneracy is well-defined according to the first definition (definition 2.36).

To show this, I am interested in $f \circ y^{-1}$ and whether it is degenerate, given that $f \circ x^{-1}$ is degenerate. Since

$$f \circ y^{-1} = f \circ x^{-1} \circ x \circ y^{-1} \tag{176}$$

the question reduces to the following: given degenerate $g = f \circ x^{-1}$ and

differentiable $z = x \circ y^{-1}$ show that $g \circ z$ is degenerate. Note that $g : \mathbb{R}^n \to \mathbb{R}$ and $z : \mathbb{R}^n \to \mathbb{R}^n$.

Let me show that $x(p_e) = t_0$ is a critical point of $f \circ x^{-1}$. Define $c_i(t) = x^{-1} \circ b_i$ with $b_i(t) = t_0 + t \cdot u_i$, where $b_i : \mathbb{R} \to \mathbb{R}^n$ and u_i is the usual i-th basis vector of $\mathbb{R}^n$. Because $c_i(0) = p_e$ there exists a $v_{p_e}^{(i)}$ such that $[c_i] = v_{p_e}^{(i)}$. On the one hand, for fixed but arbitrary i,

$$\left.\frac{d}{dt}\right|_{t=0} (f \circ c_i) = v_{p_e}^{(i)} \cdot f = df(p_e) \cdot v_{p_e}^{(i)} = 0 \tag{177}$$

because p_e is a critical point. On the other hand

$$\left.\frac{d}{dt}\right|_{t=0} (f \circ c_i) = \left.\frac{d}{dt}\right|_{t=0} (f \circ x^{-1} \circ b_i) = \nabla f \circ x^{-1}\big|_{b_i(0)} \cdot b_i'(0) = \big(\nabla(f \circ x^{-1})(t_0)\big) \cdot u_i \tag{178}$$

This means that the i-th component of $\nabla(f \circ x^{-1})(t_0)$ is zero, but since i was arbitrary, t_0 is a critical point of $f \circ x^{-1}$.

Now let $y(p_e) = s_0$. I will show that s_0 is a critical point of $g \circ z$.

$$\nabla(g \circ z)(s_0) = \nabla g|_{z(s_0)} \cdot z'(s_0) = 0 \tag{179}$$

because, as shown in the previous paragraph, $\nabla(f \circ x^{-1})(t_0) = 0$.

Here is Faà di Bruno's formula for the second derivative of a multivariable function.

$$\frac{\partial^2 y}{\partial x_i \partial x_j} = \sum_k \frac{\partial y}{\partial u_k} \frac{\partial^2 u_k}{\partial x_i \partial x_j} + \sum_{k,l} \frac{\partial^2 y}{\partial u_k \partial u_l} \frac{\partial u_k}{\partial x_i} \frac{\partial u_l}{\partial x_j} \tag{180}$$

Applying this to $g \circ z$ at s_0, the first sum immediately cancels because $t_0 = z(s_0)$ is a critical point of g. The second sum is a quadratic form (see wikipedia entry for quadratic form as well as Goult, page 75). We know nothing about the latter two factors of the second sum, but the first factor of the second sum is simply the entries of $[H_{f,p_e}]_x$ at t_0. The LHS of (180) is the entries of $[H_{f,p_e}]_y$ at s_0. Let us call $\hat{M} = [H_{f,p_e}]_x$ and $M = [H_{f,p_e}]_x$. Then the entries m_{ij} of M are

$$m_{ij} = u_i^\intercal \hat{M} u_j \tag{181}$$

for some column vectors u_k of an $n \times n$ matrix U. $\hat{M}$, being the Hessian of a well-behaved function, is real and symmetric. It is also singular by assumption. We are trying to show that M is singular as well. (181) implies immediately that M is real and symmetric (it better be, since it is the Hessian of a well-behaved function). I put a 4×4 example in `schmierbuch-2018-10-28.ipynb` and, indeed, if $\hat{M}$ is real, symmetric, and singular, so is M.

Now for the kill. Let $\hat{M} = P^\intercal B P$, i.e. $\hat{M}$ is congruent to $B = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. That B exists and reflects the eigenvalues of $\hat{M}$ is a well-known property of real, symmetric matrices, see Goult, page 142f. This means that

$$m_{ij} = (Pu_i)^\intercal B (Pu_j) \tag{182}$$

or, in other words,

$$m_{ij} = w_i^\intercal B w_j \tag{183}$$

for some column vectors w_k of an $n \times n$ matrix W. Because B is a diagonal matrix, this just means that

$$M = W^\intercal B W \tag{184}$$

M is therefore congruent to B, which is congruent to $\hat{M}$, and as matrix congruence is an equivalence relation, M is congruent to $\hat{M}$. Sylvester's law of inertia states that two congruent symmetric matrices with real entries have the same numbers of positive, negative, and zero eigenvalues. If $\hat{M}$ is singular, it mus have a zero eigenvalue. Its congruence with M means that M also has a zero eigenvalue and that means that M is singular.

Here is a nugget from my investigations at the end of October 2018. Around a putative critical point x_0 you can approximate

$$f(x) - f(x_0) \approx f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \tag{185}$$

If $f'(x_0)$ were not zero, it would beat out the other term and at x_0 there would be a sign change for $f(x) - f(x_0)$, therefore there could not be an extremum. Furthermore, if $f'(x_0) = 0$, it is clear that a positive $f''(x_0)$ signifies a minimum and a negative $f''(x_0)$ a maximum for f at x_0.

References

Boissonnat, Jean-Daniel, Frank Nielsen, and Richard Nock. "Bregman Voronoi Diagrams." *Discrete and Computational Geometry* 44, 2: (2010) 281–307.

Rockafellar, Ralph. *Convex Analysis*. Princeton, N.J: Princeton University, 1997.

Savage, Leonard. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66, 336: (1971) 783–801.

Selten, Reinhard. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics* 1, 1: (1998) 43–62.