Diagonalizing Dependencies

The essence of deep learning is to implicitly characterize a concept-to-be-learned by specifying how it should behave in composition with other concepts. Then to do gradient desceint. That is, *we build our tools by using them*. This Grothendieckian framework illuminates on many global architectural themes of deep learning: autoencoders, cycle consistency losses, attention layers, siamese networks, contrastive losses. A key example is self-supervised representation learning (e.g. Word2Vec), we learn to represent data by attempting to predict some of its "parts" or "aspects" from others.

In these notes we elaborate on this theme in the case of weak dependencies. We begin in the first of four sections by understanding dependency structures relevant to more than two variables.

Geometry of n-ary Mutual Information

Interactions, Distributivity, Local-Global

I believe that either Jupiter has life or it doesn't. But I neither believe that it does, nor do I believe that it doesn't.
— Five Thousand B.C., Raymond Smullyan

Inclusion-Exclusion — Why do we love entropy $\mathrm{H}(X)$ and mutual information $\mathrm{MI}(X;Y)$? One reason is that they are asymptotically invariant under data re-formatting: up to asymptotically negligible error, we can losslessly translate a long stream $X = (x_t : 0 \leq t < T)$ to a stream $X' = (x'_t : 0 \leq t < T)$ when the $\mathrm{H}(X) \leq \mathrm{H}(X')$. By *losslessly* I mean that X' should determine X. Likewise, we can super-losslesly translate (X, Y) to (X', Y') when $\mathrm{H}(X) \leq \mathrm{H}(X')$ and $\mathrm{H}(Y) \leq \mathrm{H}(Y')$ *and* $\mathrm{MI}(X;Y) \leq \mathrm{MI}(X';Y')$. By *super-losslesly* I mean that X' should determine X and Y' should determine Y. These entropic conditions (with equality rather than mere inequality) are necessary and sufficient (up to negligible error) for *bi-directional* super-lossless translation. We'll call such a two-way translation an *isomorphism* between same-sized tuples of random variables. For tuples of size 1 and 2, the "richness" functions $\mathrm{H}(\cdot)$ and $\mathrm{MI}(\cdot;\cdot)$ suffice to characterize isomorphism type.

Another operational characterization of mutual information is how much information Alice would have to send to Bob if Alice translates X to X' and then Bob translates Y to Y' and they want to minimize filesize.

← I think of water sloshing around in a balloon animal.

Q: What numbers characterize the isomorphism type of a tuple (X, Y, Z)? We will see that pairwise mutual information does not suffice and we hypothesize that even including triplet mutual information (in the sense of Ting 1962) does not suffice.

Venn Diagrams Mislead — I contend that 3-ary mutual information is slippery to define to our satisfaction due to a failure of additivity (precisely, distributivity) of information. An analogous situation occurs in the linear algebra: Note that if A, B, C are subspaces of a finite dimensional vector space, then

$$\dim(A + B) = \dim(A) + \dim(B) - \dim(A \cap B)$$

$$\dim(A \cap B) = \dim(A) + \dim(B) - \dim(A + B)$$

These reflect inclusion-exclusion in a venn diagram. Are the following also true?

$$\begin{aligned}\dim(A + B + C) = &\dim(A) + \dim(B) + \dim(C) \\ &- \dim(A \cap B) - \dim(B \cap C) - \dim(C \cap A) \\ &+ \dim(A \cap B \cap C)\end{aligned}$$

$$\begin{aligned}\dim(A\cap B\cap C) = &\dim(A)+\dim(B)+\dim(C)\\ &-\dim(A+B)-\dim(B+C)-\dim(C+A)\\ &+\dim(A+B+C)\end{aligned}$$

No! That's what we would expect from a Venn diagram, but it is wrong. Venn models MI badly for analogous reasons. I think of $\mathrm{MI}(X;Y;Z)$ not as a mere real number (like the area of a Venn lune) but as a geometric structure (much as a lattice of subspaces generated by sums and intersections of given subspaces carries more data than just the dimensions of the pure sums).

Here's a concrete linear algebra example. Let A,B,C be three coplanar, pairwise distinct 1-D subspaces of a three-dimensional space V. Let A',B',C' be three non-coplanar 1-D subspaces of a three-dimensional space V'. For either triplet, the intersection of any n of the subspaces has dimension $3,1,0,0$ for $n=0,1,2,3$. And yet the two situations are *not* isomorphic. So intersection dimensions — even all 2^3 dimensions together — do not determine isomorphism type.

Say $V;V'$ have bases $u,v,x;u',v',x'$ Let $\{u\},\{v\},\{u+v\}$ span A,B,C. Let $\{u'\},\{v'\},\{x'\}$ span A',B',C'.

Fixing inner products on V,V' and taking orthogonal complements of each subspace shows that sum-dimensions do not determine isomorphism type either.

Distributivity — Inspecting the above paragraphs' entropy and linear algebra examples, we see that key difference from the worlds of measure and of cardinality is a failure of distributivity. In linear algebra, even when $A\cap B=0$ we can have

$$A\cap C+B\cap C\subseteq(A+B)\cap C$$

but often not the other inclusion. And for A,B independent we have

$$\mathrm{MI}(A\,;\,C)+\mathrm{MI}(B\,;\,C)\leq\mathrm{MI}(A,B\,;\,C)$$

Without independence, it can go the other way: think of bits with $A=B=C$.

but often not the other inequality (think of bits with $A+B+C=0$). Indeed:

$$\begin{aligned}[(A)+(B)+(C)-(ABC)]-[(A)+(C)-(AC)]-[(B)+(C)-(BC)]\\ &=(AC)+(BC)-(C)-(ABC)\\ &=[(C|A)+(A)]+[(C|B)+(B)]-(C)-[(C|AB)-(A)-(B)]\\ &=(C|A)+(C|B)-(C|AB)-(C|)\\ &\geq(C|A)+(C|B)-(C|)-(C|)\geq 0\end{aligned}$$

Can we use the variational formulation instead? We'd take $f_{A;C},f_{B;C}$ witnesses (up to ϵ error) to MI$(A;C)$, MI$(B;C)$'s values. We'd set $f_{A,B;C}((a,b),c)$ to be their sum or something to get a lower bound on MI$(A,B;C)$:

Above, $(\cdots|\cdots)$ is shorthand for the conditional entropy between the tuples $\cdots$ and $\cdots$ of variables; $(\cdots)=(\cdots|)$, for the unconditional entropy.

Cohomology — Cohomology is a scary word for the study of how local data patches together to make global data. A classical example is that we can make a sphere from two hemispheres, either of which doesn't have a "hole"; but the sphere has a hole. We summarize the (2015 theory) of Pierre Baudot and Daniel Bennequin.

Information Lattice Geometry

In what sense is the left figure a better featurization of our data than the right?

Universal Features

One beginning and one ending for a book was a thing I did not agree with. A good book may have three openings entirely dissimilar and inter-related only in the prescience of the author, or for that matter one hundred times as many endings
— At Swim-Two-Birds, Brian O'Nolan

We summarize the (2019 theory) of Shao-Lun Huang, Anuran Makur, Gregory W. Wornell, and Lizhong Zheng.

Independence is a heady draft...
— Mom & Me & Mom, Maya Angelou

MI between nearly joint-uniform variables

Suppose X, Z have uniform marginals and are nearly independent —

$$p(x', z') = \frac{\exp(\epsilon\pi(x', z'))}{|X \times Z|}$$

— in that their mutual information $\mathrm{MI}(X; Z)$ is $\ll \epsilon^2$. Here, $\pi(\cdot, \cdot) \ll 1$. Write $M = |X \times Z|$.

Bird's Eye Story — We wish to featurize X, Z into k-dimensional variables X', Z' so as to preserve as much shared information as possible:

$$X' - X - Z - Z' \qquad \mathrm{MI}'(X'; Z') \lesssim \mathrm{MI}(X; Z)$$

Now, our notion MI' of shared information ought not literally be mutual information; otherwise, for positive k almost all embeddings of X, Z into $\mathbb{R}^k$ would trivially preserve information. Still, there is a sense in which some embeddings of X, Z are "more aligned" or "more co-informative" than others. We'll focus on the notion of *accessible co-informativity*: that is, on a measure of how related variables appear despite bounded computation to compare them. Variational bounds provide a convenient way to model bounded computation:

$$\mathrm{MI}(X; Z) = \sup_c \mathbb{E}_{x,z \sim \mu_{X \times Z}} [c(x, z)] - \log \mathbb{E}_{x,z \sim \mu_X \times \mu_Z} [\exp c(x, z)]$$

with the supremum achieved at $c_{x,z} = \log(p_{x,z}/(p_x p_z)) = \epsilon\pi_{x,z}$. Let's define MI' to be the same thing except the function c over which we vary must be **bilinear**.

What does that MI objective look for c near the unconstrained optimum $c = \epsilon\pi$? Well, for $c = \epsilon(\gamma + \pi)$ we have to second order in ϵ and terms constant in γ:

$$\begin{aligned}
\cdots &= \sum_{x,z} \epsilon\gamma_{x,z} \cdot \exp(\epsilon\pi_{x,z})/M - \log\left(1 + \sum_{x,z} (\exp(\epsilon\gamma_{x,z}) - 1) \cdot \exp(\epsilon\pi_{x,z})/M\right) \\
&\in C - \frac{1}{2}\sum_{x,z} (\epsilon\gamma_{x,z})^2 \cdot \exp(\epsilon\pi_{x,z})/M + \frac{1}{2}\left(\sum_{x,z} \epsilon\gamma_{x,z} \cdot \exp(\epsilon\pi_{x,z})/M\right)^2 + o(\epsilon^2) \\
&= C - \frac{\epsilon^2}{2} \mathrm{Var}_{x,z \sim \mu_{X \times Z}} [\gamma_{x,z}] + o(\epsilon^2)
\end{aligned}$$

Neither the true nor approximate MI objectives care whether we translate γ vertically. So let's take γ to have mean zero with respect to the true joint. The above says that for constrained c (so long as the feasible set gets very close to the true optimum), we want a constrained c that minimizes a least-squares error $\mathrm{Var}_{x,z \sim \mu_{X \times Z}} [\epsilon\pi - c]$ — wow!

Let's consider c bilinear in x', z' (the latter featurized per f, g):

$$c_{x,z} = \epsilon \sum_k f_x^k g_z^k$$

Momentarily neglecting centering and the difference between variance with respect to joint vs independent marginals, we see that the SVD minimizes that least-squares error!
Desired Result: *To leading order in ϵ, the optimal features f, g are proportional to the top k non-trivial singular vectors of* $\exp(\epsilon\pi)$ *(when SVD and optimization are uniquely soluble).*

TECHNICAL DETAILS —

This result justifies the truncation developed in HMWZ 2019 (§IV.C).

COMPARISON TO HMWZ — That paper characterizes its truncation as a solution to five optimization problems (§III.B; §V.C thru §V.F), each of which : (TODO: verify!)

	§III.B	§V.C	§V.D	§V.E	§V.F	ours
norm	Ky-fan				nuclear	information
constraint				feature noise		computation
	Ky-fan					

- relies on a norm introduced manually rather than derived from KL (Ky-fan, nuclear,) or
- supposes, besides near-independence of X, Z, an asymptotically extreme feature noise

By contrast, in our analysis the universal features fall out from a linearity constraint (having to do with pure linearity rather than norms).

adaptation to non-uniform marginals

He had bought a large map representing the sea,
Without the least vestige of land:
And the crew were much pleased when they found it to be
A map they could all understand.
— The Hunting of the Snark, Charles Dodgson

principal components We review principal component analysis.

PYTHAGOREAS AND GAUSS —

SIDETRACK: CAUSAL GRASSMANIANS — Pearl's do-calculus –> fit a nilpotent matrix to data?! PCA:correlation :: ?:causation

```
(I + T) nxn
T^n = 0
 FED
(1...
( 1..
(  1.
(   1)
```

TODO: in nearly markov case, notion of "orthogonality" is different... offdiagonals ... perhaps a triangularization?? what does this look like?

—

FISCHER METRIC —

greg's analysis of weak pairwise dependencies

sample complexity

Nearly Markov

> *... it invariably delivered a cupful of liquid that was almost, but not quite, entirely unlike tea.*
> — Hitchhiker's Guide to the Galaxy, Douglas Adams

```
x1'    x2'    y'
 |      |      |
x1-----x2-----y
       |
      x2'
```

We're studying a perturbation of a system exactly described by a markov structure (shoe horizontal edges far from nearly-independent). One interesting task in this setting is to predict y based on x2, then fine-tune based on x1. This relates to self-supervised learning for representation learning; pretraining. We want to perturb in "p(y | x2) vs p(y | x2,x1)"; perhaps the right analogue to the original story is for us to assume access to two-point but not three-point marginals.

Start with understanding for one simple SVD in nearly markov how to interpret that SVD.

Now, *what are analogues of the original story's singular vectors?* The bounded-computation framework suggests that we maximize something like

$$\mathrm{MI}'(x1'; y' \mid x2) \qquad \text{or maybe} \qquad \mathrm{MI}'(x1'; y' \mid x2')$$

The thing to do, though, is to define MI′! For instance, might we condition on a "mediating" variable just by taking an expectation over the MI′s of its conditionals, as follows?

$$\mathrm{MI}'(X; Z \mid M) = \mathbb{E}_{m \sim \mu_M} \sup_{c \text{ bilinear}} \mathbb{E}_{x,z \sim \mu_{X \times Z|M}(\cdot|m)} [c(x,z)] - \log \mathbb{E}_{x,z \sim \mu_{X|M}(\cdot|m) \times \mu_{Z|M}(\cdot|m)} [\exp c(x,z)]$$

Or should c instead somehow know about m as well?

Relatedly, if we condition on $x2'$ then is there a way we would like to incorporate $x2'$'s linear structure? And if we condition instead on $x2$, is there a decomposition theorem reducing to old work.

Self-Supervised Learning