Diagonalization of Dependencies

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

In deep learning, we implicit characterize a concept-to-be-learned by specifying how it should behave in composition with other concepts. That is, *we build our tools by using them*. This Grothendieckian framework sheds light 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, wherein we learn to representing 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.

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

Universal Features

Independence is a heady draft...
— Maya Angelou

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

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:

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

	§III.B	§V.C	§V.D	§V.E	§V.F	ours
norm	Ky-fan				nuclear	none
noise	Ky-fan			featurization		
norm	Ky-fan					

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!)

- 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

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

Nearly Markov

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

In nearly markov structure, predict y based on x2; and fine-tune based on x1 — relate to self-supervised learning for representation learning; pretrainig. (Assume have two-point but not three-point marginals). PRIORIZZE THIS PARAGRAPH perturb in "p(y—x2) vs p(y—x2,x1)" — what is analogue of singular vectors/values here?

Generalized Self-Supervised

Sample Complexity