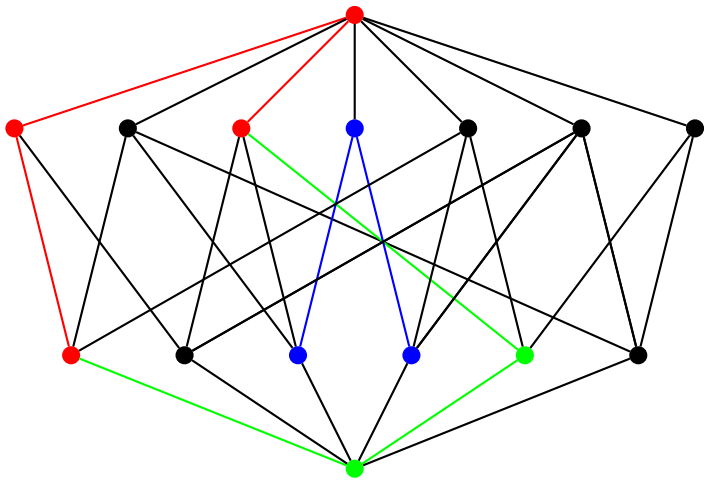
Meta-embeddings: a probabilistic generalization of embeddings in machine learning

Niko Brümmer[1] , Lukáš Burget, Paola Garcia,
Oldřich Plchot, Johan Rohdin, Daniel Romero,
David Snyder, Themos Stafylakis, Jesús Villalba

JHU HLTCOE 2017 SCALE Workshop



[1]Corresponding author: niko.brummer@gmail.com. The other authors are in alphabetical order.

Contents

List of Figures

Chapter 1

Introduction

Embeddings are familiar in modern machine learning. Neural nets to extract word embeddings[1] were already proposed in 2000 by Bengio [1]. Now embeddings are used more generally, for example in state-of-the-art face recognition, e.g. Facenet [2].

Embeddings are becoming popular also in speech and speaker recognition. In Interspeech 2017, eighteen papers had the word 'embedding' in the title.[2] In speaker recognition and spoken language recognition, we have been using i-vectors, a precursor to embeddings, for almost a decade [3, 4, 5]. More general embeddings are now appearing in speaker recognition, see for example the Voxceleb paper [6] and David Snyder's work [7, 8]. A new embedding method for language recognition is proposed in [9].

Input patterns (sequences of acoustic feature vectors, images, text, ...) live in large, complex spaces, where probability distributions and geometrical concepts such as distance are difficult to formulate. Embeddings extracted from the input patterns live in simpler spaces, e.g. $\mathbb{R}^d$ (multidimensional Euclidean space), where distance is naturally defined and can be put to work to compare patterns.

At the Johns Hopkins HLTCOE SCALE 2017 Workshop[3] one of the topics of interest was embeddings for speaker recognition. During the workshop, the idea to generalize to *meta-embeddings* was conceived. At the time of the writing of this document, we are still developing the theory and no code has been written, nor have any experiments been performed. This document describes the theory.

We envisage that meta-embeddings could be widely applicable to a variety of machine learning or pattern recognition problems, but we shall make

[1] en.wikipedia.org/wiki/Word_embedding

[2] www.interspeech2017.org/program/technical-program/.

[3] http://hltcoe.jhu.edu/research/scale/scale-2017

our discussion concrete by using the language of automatic speaker verification/recognition. To interpret everything (say) as face recognition, just do *recording* $\mapsto$ *image* and *speaker* $\mapsto$ *face*, and so on ....

1.1 Meta-embeddings

An important concept used throughout this document is that of the *hidden identity variable*—a hypothetical, multidimensional variable that ideally contains all of the information that distinguishes one speaker/face/individual from others. In speaker recognition, the hidden *speaker identity variable* is well known from the work of Patrick Kenny in JFA [10] and PLDA [11]. Similar identity variables appeared in earlier face recognition applications of PLDA [12, 13, 14].

We argue that the way embeddings are currently treated, essentially makes them *point estimates of hidden variables*. We propose instead to let the identity variables—and therefore the embeddings—remain hidden and to instead extract *meta-embeddings*, which are probabilistic representations of what the values of the hidden variables might be. For example, if the embedding lives in $\mathbb{R}^d$, then the meta-embedding could be in the form of a multivariate normal distribution, where the mean could act as point estimate, but where there is also a covariance matrix that quantifies the uncertainty around the mean.

Quantifying the uncertainty is very important if the recognizer is to be applicable to variable and sometimes challenging conditions. In speaker recognition, a short, noisy, narrow-band recording should leave much more uncertainty about the speaker than a long, clean, wideband recording. In face recognition, compare a well-lit, high resolution, full-frontal face image to a grainy, low resolution, partly occluded face. In fingerprint recognition, compare a clean, high-resolution ten-print, to a single, distorted, smudged fingermark retrieved from a crime scene.

The idea to represent the uncertainty is not novel. Indeed, it is *obvious* that to do things properly, we must take the uncertainty into account. The problem is that it turns out to be difficult to do in practice.

- Part of the problem is computational. For example, a point estimate for an identity variable in $\mathbb{R}^d$ can be represented as a d-dimensional vector. But for a multivariate normal meta-embedding, the covariance is a d-by-d matrix, which is more expensive to store and manipulate.

- Then there is the added complication that covariance matrices have to be positive definite. If we use a neural net to compute our uncertainty

for us, we have to specially structure the neural net to respect this requirement.

- But the problem has another facet—if a neural net manufactures meta-embeddings for us, how do we know that the uncertainty thus represented is reasonable? Given meta-embeddings extracted from some supervised database, what criterion can we apply to measure the goodness of the information and the uncertainty they contain? Such criteria are important, because we need them for discriminative training of our neural nets.

This document will provide some suggestions for tackling the above problems.

The main purpose of the document is however to show that the meta-embeddings are themselves embeddings in the sense that they live in a vector space, equipped with geometric and algebraic structure that can be employed to compute probabilistic recognition scores (likelihood ratios). The hope is that the theoretical part of this document will help to provide the structure to guide further research into finding better solutions for the practical problems of working with this kind of uncertainty.

1.2 Prior work

In [15], Gaussian embeddings are proposed to represent uncertainty. To be expanded, ...

Mention uncertainty propagation in i-vectors[16, 17, 18, 19], ...

Chapter 2

Distillation of meta-embeddings from independence assumptions

In this chapter we use probability theory and some basic independence assumptions to motivate and derive the concept of meta-embeddings.

We use speaker recognition terminology, but everything is applicable more widely (images, etc. ...). Our inputs are *voice recordings*, each assumed to contain a single speaker. Recordings can have various representations, i-vectors, sequences of feature vectors, the raw speech samples, etc. .... At the writing of this document, recordings represented as feature vector sequences are of primary interest.

In general, we will be interested in *sets* of recordings, ranging from the set of all the recordings in a training database, down to just a pair of recordings in a verification trial. We represent a set of n recordings as:

$$\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$$

We shall work with hypotheses about how such sets of recordings are partitioned w.r.t. speaker and we shall make use of independence assumptions conditioned on these hypotheses.

2.1 Independence assumptions

We base everything that follows on a pair of simple independence assumptions:

- Multiple recordings of the same speaker are *exchangeable*.
- Recordings of different speakers are *independent*.

Almost all current probabilistic solutions in speaker recognition—also PLDA—make use of these assumptions. While these assumptions are mathematically convenient, in the real world they are not exactly true:

- The speech of any speaker will change over a time span of years and recordings made over such a period will not be exchangeable.
- The independence assumption between different speakers can only apply if our recognizer is sufficiently well trained: The recognizer has to have been exposed to so much speech, that only the speech of a new speaker itself can provide information about that speaker.

As long as we understand the limitations imposed by our assumptions, we can proceed.

In what follows, we use the notation H_1 for the hypothesis that the set of recordings of interest, $\mathcal{R} = \{r_1, \ldots, r_n\}$, were all spoken by the same (one) speaker. For the hypothesis that they were spoken by m different speakers, we use H_m. Exchangeability means:

$$P(\mathcal{R} \mid H_1) = P(r_1, r_2 \ldots, r_n \mid H_1) = P(r_2, r_1, \ldots, r_n \mid H_1) = \cdots$$

where the joint probability does not depend on the order of the recordings. According to De Finetti's exchangeability theorem,[1] the only way to obtain exchangeability is to introduce some hidden variable, say $\mathbf{z} \in \mathcal{Z}$, such that the recordings are *conditionally independent* given $\mathbf{z}$. Marginalizing over $\mathbf{z}$ w.r.t. some prior distribution $\pi(\mathbf{z})$, gives the exchangeable joint distribution [20]:

$$P(\mathcal{R} \mid H_1) = \Big\langle \prod_{i=1}^{n} P(r_i \mid \mathbf{z}) \Big\rangle_{\mathbf{z}\sim\pi} \tag{2.1}$$

where the angle brackets denote expectation—for continuous $\mathbf{z}$, this could be written as an integral and for discrete $\mathbf{z}$ as a summation. By this marginalization, we are inducing dependence between recordings of the same speaker, while retaining exchangeability. *We do need this dependence, otherwise speaker recognition would be impossible!*

Let $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_m$ be a number of non-overlapping sets of recordings, spoken respectively by m different speakers. The between-speaker independence

[1] en.wikipedia.org/wiki/Exchangeable_random_variables

assumption gives:

$$P(\mathcal{R}_1, \ldots, \mathcal{R}_m \mid m \text{ speakers}) = \prod_{i=1}^{m} P(\mathcal{R}_i \mid H_1) = \prod_{i=1}^{m} \Big\langle \prod_{r \in \mathcal{R}_i} P(r \mid \mathbf{z}) \Big\rangle_{\mathbf{z} \sim \pi} \tag{2.2}$$

where we have expanded $P(\mathcal{R}_i \mid H_1)$ using (2.1).

According to this model, the hidden variable, $\mathbf{z} \in \mathcal{Z}$, contains everything that is to be known about recordings of a speaker. Patrick Kenny calls $\mathbf{z}$ the *speaker identity variable*. At this stage, we are not committing ourselves to the nature of the hidden variables and therefore we leave $\mathcal{Z}$ undefined—after all, $\mathbf{z}$ is hidden so there is much freedom in choosing its nature. (If you need something concrete to shape your thoughts, $\mathcal{Z} = \mathbb{R}^d$ is a good choice.)

Think of the identity variable, $\mathbf{z}$, as the ideal embedding that an oracle could extract. If we were given $\mathbf{z}$, the problem would be instantly solved. Unfortunately, there is noise, distortion, occlusion and all kinds of other mechanisms that induce uncertainty, so we can never exactly extract $\mathbf{z}$. Let's not pretend that we can do this. Let us not extract some point-estimate for $\mathbf{z}$ and call that our embedding. Let's instead extract a *meta-embedding*, which is *information about* the ideal embedding, $\mathbf{z}$, where the information has a probabilistic form. We derive this form below.

2.2 Speaker identity likelihood functions

Given a pair of recordings, $\mathcal{R} = \{r, r'\}$, we can answer the question of whether they belong to the same speaker or not in terms of the *likelihood-ratio (LR)* [21]:

$$\begin{aligned} \frac{P(r, r' \mid H_1)}{P(r, r' \mid H_2)} &= \frac{\big\langle P(r \mid \mathbf{z})\ P(r' \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}}{\big\langle P(r \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi} \big\langle P(r' \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}} \\ &= \frac{\big\langle kP(r \mid \mathbf{z})\ k'P(r' \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}}{\big\langle kP(r \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi} \big\langle k'P(r' \mid \mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}} \\ &= \frac{\big\langle f_r(\mathbf{z})\ f_{r'}(\mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}}{\big\langle f_r(\mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi} \big\langle f_{r'}(\mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi}} \end{aligned} \tag{2.3}$$

where we have defined the *speaker identity likelihood functions*:

$$f_r(\mathbf{z}) = kP(r \mid \mathbf{z}) \qquad \text{and} \qquad f_{r'}(\mathbf{z}) = k'P(r' \mid \mathbf{z}), \tag{2.4}$$

where $k, k' > 0$ are arbitrary constants that may depend on r, r', but not on $\mathbf{z}$. Notice that the LR depends on the data (r, r') only via f_r and $f_{r'}$. The speaker information in r is represented by the *whole function*

$$f_r : \mathcal{Z} \to \mathbb{R}$$

rather than by the function value $f_r(\mathbf{z})$ at some particular value of $\mathbf{z}$ (remember $\mathbf{z}$ is hidden and we are never given a fixed value for it).

The full speaker information cannot be represented by some estimate of $\mathbf{z}$. We could for example use $\hat{\mathbf{z}} = \operatorname{argmax} f_r(\mathbf{z})$ as a maximum-likelihood point-estimate for $\mathbf{z}$, but that would be throwing information away. The full information about the speaker is contained in the whole function, f_r.

The representation of the speaker information can be further understood by noticing that f_r and $f_{r'}$ represent r, r' in any posteriors for the speaker identity variable, for example:

$$P(\mathbf{z} \mid r, \pi) = \frac{\pi(\mathbf{z}) f_r(\mathbf{z})}{\left\langle f_r(\mathbf{z}') \right\rangle_{\mathbf{z}' \sim \pi}} \tag{2.5}$$

and if r and r' are known to be of the same speaker, then:

$$P(\mathbf{z} \mid r, r', H_1, \pi) = \frac{\pi(\mathbf{z}) f_r(\mathbf{z}) f_{r'}(\mathbf{z})}{\left\langle f_r(\mathbf{z}') f_{r'}(\mathbf{z}') \right\rangle_{\mathbf{z}' \sim \pi}} \tag{2.6}$$

2.3 The general case

Our speaker identity likelihood functions are applicable more generally than just to pairs of recordings. Under the independence assumptions of section 2.1, speaker identity likelihood functions can be used to answer *any* question that can be formulated in terms of partitioning a set of recordings according to speaker [22]. Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ be a set of recordings and let $A : \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m \subseteq \{1, 2, \ldots, n\}$ be a hypothesized partition of $\mathcal{R}$ into subsets belonging to $m \geq 1$ different hypothesized speakers.[2] Similarly, let $B : \mathcal{S}'_1, \ldots, \mathcal{S}'_{m'}$ be a different hypothesized partition, having m' hypothesized speakers. A and B are labels that refer to the two alternate partitioning hypotheses. Using within-speaker exchangeability and between-speaker

[2] The subsets are non-empty, non-overlapping and their union equals $\{1, 2, \ldots, n\}$. We allow $m = 1$, in which case $\mathcal{S}_1 = \mathcal{R}$.

independence, the LR comparing A to B is:

$$\begin{aligned}\frac{P(\mathcal{R} \mid A)}{P(\mathcal{R} \mid B)} &= \frac{\prod_{i=1}^{m} \left\langle \prod_{j \in \mathcal{S}_i} P(r_j \mid \mathbf{z}) \right\rangle_{\mathbf{z} \sim \pi}}{\prod_{i=1}^{m'} \left\langle \prod_{j \in \mathcal{S}'_i} P(r_j \mid \mathbf{z}) \right\rangle_{\mathbf{z} \sim \pi}} \\ &= \frac{\prod_{i=1}^{m} \left\langle \prod_{j \in \mathcal{S}_i} f_j(\mathbf{z}) \right\rangle_{\mathbf{z} \sim \pi}}{\prod_{i=1}^{m'} \left\langle \prod_{j \in \mathcal{S}'_i} f_j(\mathbf{z}) \right\rangle_{\mathbf{z} \sim \pi}}\end{aligned} \tag{2.7}$$

where, as above, the *speaker identity likelihood function* extracted from r_j is:

$$f_j(\mathbf{z}) = k_j P(r_j \mid \mathbf{z}) \tag{2.8}$$

The last equality in (2.7) follows because the arbitrary scaling constants, $\{k_j\}_{j=1}^{n}$, are the same in the numerator and denominator and cancel. This LR form is convenient because of this cancellation.

Any kind of speaker recognition problem that can be expressed in terms of partitioning recordings according to speaker, can be scored at runtime via likelihood-ratios of the form (2.7). We shall show some examples below. Moreover, any discriminative training criteria that are expressed as functions of the LR scores can then also be expressed in terms of (2.7). In summary:

> *For discriminative speaker recognition, all scoring and training calculations can be expressed in terms of the speaker identity likelihood functions.*

To better appreciate this generality, we shall present some examples in the next section.

Again, for readers that are comfortable with the idea that probability distributions convey information [23], the speaker identity likelihood functions are closely associated with posteriors for $\mathbf{z}$. Any such posterior, conditioned on a number of recordings of a speaker indexed by $\mathcal{S}$, can be computed as:

$$P(\mathbf{z} \mid \mathcal{R}, \mathcal{S}, H_1, \pi) = \frac{\pi(\mathbf{z}) \prod_{j \in \mathcal{S}} f_j(\mathbf{z})}{\left\langle \prod_{j \in \mathcal{S}} f_j(\mathbf{z}') \right\rangle_{\mathbf{z}' \sim \pi}} \tag{2.9}$$

The likelihood function, f_j, or indeed a product of any number of them, can be regarded as a *raw* form of posterior distribution for $\mathbf{z}$. We use the term raw, because the likelihood function is independent of the choice of prior, π, and is also un-normalized. Nevertheless, f_j contains all the speaker information that a given probabilistic model could extract from r_j.

2.4 Examples

The examples below should aid understanding of the meaning and applications of (2.7). We will frequently refer back to these examples in the rest of this document.

Simple verification. Let $\mathcal{R} = \{r_1, r_2\}$, $A : \mathcal{S}_1 = \{1, 2\}$ and $B : \mathcal{S}'_1 = \{1\}, \mathcal{S}'_2 = \{2\}$. Then $\frac{P(\mathcal{R}|A)}{P(\mathcal{R}|B)}$ is the canonical (single-enroll) speaker verification LR.

Multi-enroll verification. Let $\mathcal{R} = \{r_1, r_2, r_3\}$, $A : \mathcal{S}_1 = \{1, 2, 3\}$ and $B : \mathcal{S}'_1 = \{1, 2\}, \mathcal{S}'_2 = \{3\}$. Then $\frac{P(\mathcal{R}|A)}{P(\mathcal{R}|B)}$ is the speaker verification LR, with enrollment recordings r_1, r_2 and test recording r_3.

Open set classification. Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ be partitioned into m speakers, indexed by $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq \{1, \ldots, n\}$. Let $r' \neq \mathcal{R}$ be an additional (test) recording. For $1 \leq i \leq m$, let A_i be the hypothesis that r' is spoken by speaker i, as indexed by $\mathcal{S}_i$. Let A_{m+1} be the hypothesis that there are $m+1$ speakers: m of them indexed by $\mathcal{S}_1, \ldots, \mathcal{S}_m$ and r' spoken by a new speaker. Then we can score the open-set speaker classification problem as follows: For $1 \leq i \leq m+1$, the likelihood[3] for A_i is

$$L_i = \frac{P(\mathcal{R}, r' \mid A_i)}{P(\mathcal{R}, r' \mid A_{m+1})} \tag{2.10}$$

Notice that likelihood that r' was spoken by a new speaker is just $L_{m+1} = 1$. Assuming some hypothesis prior, $P(A_i)$, the hypothesis posterior is:

$$P(A_i \mid \mathcal{R}, r') = \frac{P(A_i)L_i}{\sum_{j=1}^{m+1} P(A_j)L_j} \tag{2.11}$$

Agglomerative hierarchical clustering. Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ be partitioned into m hypothesized speakers, indexed by $A : \mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq \{1, \ldots, n\}$. For $1 \leq i, j \leq m$ and $i \neq j$, let B_{ij} be the hypothesis that the speakers are correctly indexed as given by A, except that the recordings indexed by $\mathcal{S}_i$ and $\mathcal{S}_j$ are from the same speaker. That is, if we accept hypothesis B_{ij}, then

[3]The numerator of L_i by itself, $P(\mathcal{R}, r' \mid A_i)$, could also serve as likelihood for A_i. Since the denominator is independent of i, the normalized ratio, L_i, is still a valid likelihood for A_i, with the advantage that the troublesome data-dependent constants cancel.

we should reduce the number of speakers to $m-1$ and merge $\mathcal{S}_i$ and $\mathcal{S}_j$. We let

$$L_{ij} = \frac{P(\mathcal{R} \mid B_{ij})}{P(\mathcal{R} \mid A)} \tag{2.12}$$

be the (conveniently normalized) likelihood for hypothesis B_{ij}, while the likelihood for hypothesis A is just 1. These likelihood scores can be used as part of an iterative, greedy optimization[4] procedure (agglomerative hierarchical clustering) to find a partition of $\mathcal{R}$ according to speaker, with high posterior probability.

Summary. All of these examples are scored using likelihood ratios of the form (2.7) and are therefore dependent on the data only via the speaker identity likelihood functions.

We have so far demonstrated that f_j qualifies as an embedding in the sense that it contains all of the relevant information in r_j. In what follows we shall refer to $\mathbf{z}$ as the (hidden, ideal) *embedding*, while the f_j are our *meta-embeddings*.

It remains to be shown in the next chapter that these meta-embeddings live in a space with useful algebraic and geometric structure.

[4]Finding the global optimum is apparently hopelessly intractable. For a set of $n = 75$ recordings, we already have that the number of possible ways to partition this set exceeds the number of atoms (about 10^{80}) in the observable universe.

Chapter 3

The structure of meta-embedding space

Classical embeddings in machine learning live in finite-dimensional, Euclidean space, $\mathbb{R}^d$. When equipped with the standard scalar product, vector addition, dot-product and Euclidean distance, $\mathbb{R}^d$ is a vector space, inner-product space, metric space, Hilbert space, etc. Our meta-embeddings, the functions f_j, live in a slightly generalized space, which is still a vector space, inner product space, metric space and (subject to regularity conditions) a Hilbert space.

Depending on the nature of the hidden variable, $\mathbf{z} \in \mathcal{Z}$, our meta-embeddings might live in a space with very high, or even infinite dimensionality. In the theory that follows, everything can be more elegantly described in this high/infinite-dimensional space, but in practice—to be discussed in the next chapter—we will consider various finite, tractable representations of our meta-embeddings.

In this chapter we explore first the geometric (vector space) properties of meta embeddings that establish concepts such as length, distance and angle. The vector space operations include addition, scalar multiplication and inner product. Next we enrich the vector space by introducing another (elementwise) multiplicative operator, which then forms an algebra. We explain some of the properties of this algebra and finally show how these algebraic manipulations can be used to compute likelihood-ratios of the form (2.7).

The chapter concludes with a simple graphical example to aid understanding.

3.1 Geometric properties

Let $\mathbb{R}^{\mathcal{Z}}$ denote the space of all possible functions from $\mathcal{Z}$ to $\mathbb{R}$ and let our meta-embeddings live in $\mathcal{F}$, where $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. We need to confine meta-embeddings to some subset of $\mathbb{R}^{\mathcal{Z}}$ in order to exclude all kinds of pathological functions and to ensure that the operations defined on meta-embeddings are well behaved. Since our meta-embeddings are likelihoods, the function values are always non-negative, so we might consider restricting ourselves to $\mathbb{R}_+^{\mathcal{Z}}$, where $\mathbb{R}_+ = [0, \infty)$. But this does not give a vector space, because it will not be closed w.r.t. scalar multiplication. It will therefore be more convenient to work in a more general space, $\mathcal{F}$, that includes functions with negative values.

3.1.1 Functions as vectors

Consider the Euclidean vector, $\mathbf{x} \in \mathbb{R}^d$ and notice that $\mathbf{x}$ can be viewed as a *function* that maps the component index to the real line: $\{1, \ldots, d\} \to \mathbb{R}$.

Now let the hidden speaker identity variable be a d-dimensional Bernoulli vector, $\mathbf{z} \in \mathcal{Z} = \{0, 1\}^d$. That is, $\mathbf{z}$ is a vector with d components, each of which can be either 0 or 1. A function, $f : \{0, 1\}^d \to \mathbb{R}$, can be represented as a vector having 2^d real components. More generally, whenever $\mathcal{Z}$ has finite cardinality, $f : \mathcal{Z} \to \mathbb{R}$ can be represented as a vector in Euclidean space.

However, for the continuous case, when $\mathbf{z} \in \mathbb{R}^d$, the function $f(\mathbf{z})$ is like a vector with an infinite number of components. For the vector-space operations of *scalar multiplication* and *addition*, the infinite dimensionality poses no difficulty. These operations are defined as follows. Let $f, g, h \in \mathcal{F}$ and $\alpha, \beta \in \mathbb{R}$, then:

$$h = \alpha f + \beta g \qquad \Leftrightarrow \qquad h(\mathbf{z}) = \alpha f(\mathbf{z}) + \beta g(\mathbf{z}) \tag{3.1}$$

As long as we choose $\mathcal{F}$ such that it is closed under these operations and to include the constant function with value 0, then $\mathcal{F}$ is a vector space, even though the vectors are infinite-dimensional. To quality as a vector space, the properties required of the addition and scalar multiplication operators — commutativity, associativity and distributivity—follow directly from (3.1).

3.1.2 Inner product

We already have that $\mathcal{F}$ is a vector space. To further make it an *inner product space*, we need to define the inner product, which is a function, $\mathcal{F} \times \mathcal{F} \to \mathbb{R}$. For $f, g, h \in \mathcal{F}$, and $\alpha, \beta \in \mathbb{R}$, we define our inner product as:

$$\big\langle f, g \big\rangle = \big\langle f(\mathbf{z}) g(\mathbf{z}) \big\rangle_{\mathbf{z} \sim \pi} \tag{3.2}$$

which is the expectation of the product of f and g, w.r.t π, a probability distribution over $\mathcal{Z}$. If we choose π to be our hidden variable prior, then the inner product coincides with the numerator in the likelihood-ratio of (2.3). Notice that we use the same triangular bracket notation for both inner product and expectation—this ambiguous notation will be convenient later.

For (3.2) to be a valid inner product, it needs to satisfy symmetry, linearity, and positive definiteness. The symmetry, $\langle f, g \rangle = \langle g, f \rangle$ and linearity, $\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle$, follow directly from the definition. It also follows that $\langle f, f \rangle \geq 0$, which is necessary for positive definiteness. However, we also need that $\langle f, f \rangle = 0$ if and only if $f(\mathbf{z}) = 0$ everywhere. For this we need $\pi(\mathbf{z})$ to be non-zero everywhere on $\mathcal{Z}$ and we need $\mathcal{F}$ to impose certain smoothness constraints—e.g. we need to exclude from $\mathcal{F}$ functions that differ from each other only on subsets of $\mathcal{Z}$ of measure zero. In what follows, we shall assume that $\mathcal{F}$ is chosen such that it forms a valid inner-product space together with the operations defined here.[1]

Notice that (3.2) is a generalization of the usual dot product in Euclidean space. It generalizes the simple summation in the dot product by introducing the weighting by $\pi(\mathbf{z})$. For continuous $\mathcal{Z}$, summation is further generalized to integration.

Is $\mathcal{F}$ a Hilbert space? A Hilbert space is an inner product space with the additional requirement of *completeness*. If $\mathcal{F}$ is chosen to have this property, then yes, we have a Hilbert space. In what follows, we will not make use of completeness and we will therefore refer to our space with (slightly) greater generality as an inner product space.

3.1.3 Norm, distance, angle

The inner product naturally supplies the notions of length, distance and angle, making our space a metric space and a normed space. The norm (length) is defined as:

$$\|f\| = \sqrt{\langle f, f \rangle} \tag{3.3}$$

while (squared) distance is defined as:

$$\|f - g\|^2 = \langle f - g, f - g \rangle = \|f\|^2 + \|g\|^2 - 2\langle f, g \rangle \tag{3.4}$$

[1] An alternative construction of the inner product space could be to use the quotient space, populated by equivalence classes of functions, where two functions, $f(\mathbf{z})$ and $g(\mathbf{z})$ are equivalent if $P(f(\mathbf{z}) = g(\mathbf{z}) \mid \mathbf{z} \sim \pi) = 1$. See [24].

For any inner product space we have the Cauchy-Schwartz inequality:

$$\left|\langle f, g\rangle\right| \leq \|f\| \, \|g\| \tag{3.5}$$

so that the angle, θ, between f and g can be defined as:

$$\cos\theta = \frac{\langle f, g\rangle}{\|f\| \, \|g\|} \tag{3.6}$$

because then $|\cos\theta| \leq 1$, as it should.

3.1.4 Note on meta-embeddings as random variables

This note is of theoretical interest and is not essential to understanding the rest of the document.

Let $(\mathcal{Z}, \mathcal{A}, \pi)$ be a probability space, where $z \in \mathcal{Z}$ is the identity variable as defined above. $\mathcal{A}$ is a suitable sigma-algebra, consisting of subsets of $\mathcal{Z}$. The probability measure of this space, $\pi : \mathcal{A} \to [0, 1]$ is the prior on z. Under the regularity condition that our meta-embeddings are measurable functions from $\mathcal{Z}$ to $\mathbb{R}$, the meta-embeddings can be viewed as *random variables* [25]. Informally, when f_j is a meta-embedding and if we sample $z \sim \pi$, then the likelihood $f_j(z)$ is a sample from the random variable f_j.

The interested reader is encouraged to read both [24] and [26] where it is explained how the expectation of the product of two random variables, i.e. our (3.2), forms a well-defined inner product. Further reading may include appendix B of [27] and references therein.

3.2 Algebraic properties

Having described the inner-product space, we now enrich our meta-embedding space with the elementwise multiplicative operation. This operation combines in useful ways with the other operations to enable us to do the calculations we need for our ultimate goal of computing the likelihood ratios of the form (2.7).

Alert readers may notice that our addition operator does not seem to play any useful, practical role in all of this. First, the fact that addition between meta-embeddings is possible helps to establish that we have a vector space. Once this is established, all of the well-known properties of vector spaces become available to us. (For example, to derive the basic property of distance, we need subtraction, which in turn is defined by scalar multiplication and addition.) Moreover, in the next chapter, we do find a practical application for addition, in the form of mixture distributions.

3.2.1 Elementwise product

We now equip our space with the *elementwise product* (Hadamard product). Since this product is associative, we can write it using juxtaposition, defining it as:

$$h = fg \qquad \Leftrightarrow \qquad h(\mathbf{z}) = f(\mathbf{z})g(\mathbf{z}) \tag{3.7}$$

The vector space, $\mathcal{F}$, equipped additionally with the elementwise product forms an *algebra*, because the elementwise product is bilinear w.r.t. addition and scalar multiplication.[2]

We introduce the elementwise product to represent *pooling* of the relevant information extracted from multiple recordings. Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ be a set of recordings of the same speaker. If the corresponding meta-embeddings are $f_j(\mathbf{z}) = k_j P(r_j \mid \mathbf{z})$, then

$$f_{\mathcal{R}}(\mathbf{z}) = \prod_{j=1}^{n} f_j(\mathbf{z}) = \big(\prod_j k_j\big) P(\mathcal{R} \mid \mathbf{z}) \tag{3.8}$$

gives the *pooled meta-embedding*, $f_{\mathcal{R}} = \prod_j f_j$, which represents all of the speaker information in $\mathcal{R}$, in the sense that:

$$P(\mathbf{z} \mid \mathcal{R}, H_1, \pi) = \frac{\pi(\mathbf{z}) f_{\mathcal{R}}(\mathbf{z})}{\big\langle f_{\mathcal{R}}(\mathbf{z}')\big\rangle_{\mathbf{z}' \sim \pi}} \tag{3.9}$$

3.2.2 Multiplicative identity

We define the *multiplicative identity* as the constant function with value 1 everywhere: $\mathbf{1}(\mathbf{z}) = 1$. This turns out to be a (conveniently normalized) meta-embedding extracted from any *empty recording*. An empty recording is any recording that contains no speech—it could have zero duration, or otherwise contain silence, noise, or any other non-speech sounds. Since empty recordings have no speaker information, speaker identity variable posteriors conditioned on them are just equal to the prior. Letting r_ϕ be an empty recording, we have:

$$P(\mathbf{z} \mid r_\phi, \pi) = \frac{\mathbf{1}(\mathbf{z})\pi(\mathbf{z})}{\big\langle \mathbf{1}(\mathbf{z})\big\rangle_{\mathbf{z} \sim \pi}} = \pi(\mathbf{z}) \tag{3.10}$$

[2]Compare this to the complex numbers, which under addition and (real) scalar multiplication, form a 2-dimensional vector space. If complex multiplication is added, it is also an algebra. See: https://en.wikipedia.org/wiki/Algebra_over_a_field.

Notice that:

$$\|\mathbf{1}\| = \sqrt{\langle 1^2 \rangle_{\mathbf{z}\sim\pi}} = 1 \tag{3.11}$$

so that $\mathbf{1}$ represents the *diagonal axis* of length 1, with all 'components' equal to 1. The inner product, $\langle f, \mathbf{1} \rangle$, will be of special interest below. It is the length of the orthogonal projection of f onto this diagonal axis—see section 3.4 for a graphical example.

3.2.3 L1 and L2 norms

We have already introduced the norm, $\|f\| = \sqrt{\langle f, f \rangle}$. Observe that this is an L2 norm:

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\langle f^2(\mathbf{z}) \rangle_{\mathbf{z}\sim\pi}} = \|f\|_2 \tag{3.12}$$

We shall also find the L1 norm useful. Notice that when f is non-negative, we can interpret $\langle f, \mathbf{1} \rangle$ as an L1 norm:

$$\|f\|_1 = \langle |f(\mathbf{z})| \rangle_{\mathbf{z}\sim\pi} \tag{3.13}$$

It is easy to show using (3.11) and Cauchy-Schwartz (3.5) that:

$$\|f\|_1 \leq \|f\|_2 \tag{3.14}$$

If f is non-negative, then $\langle f, 1 \rangle = \|f\|_1 \leq \|f\|_2$. More generally:

$$\left|\langle f, 1 \rangle\right| \leq \|f\|_2 \tag{3.15}$$

Below we restrict attention to meta-embeddings for which $\left|\langle f, \mathbf{1} \rangle\right| < \infty$ and it is useful to know that this condition is automatically satisfied if $\|f\| = \|f\|_2 < \infty$.

3.2.4 Normalized meta-embedding

Let $f \in \mathcal{F}$ be a meta-embedding. We define the corresponding *normalized meta-embedding* as:

$$\overline{f} = \frac{1}{\langle f, \mathbf{1} \rangle} f \qquad \Leftrightarrow \qquad \overline{f}(\mathbf{z}) = \frac{f(\mathbf{z})}{\langle f(\mathbf{z}') \rangle_{\mathbf{z}'\sim\pi}} \tag{3.16}$$

If f_r is the meta-embedding for recording r and $f_r(\mathbf{z}) = kP(r \mid \mathbf{z})$, then:

$$P(\mathbf{z} \mid r, \pi) = \pi(\mathbf{z})\overline{f_r}(\mathbf{z}) \tag{3.17}$$

Similarly, for a set of recordings, $\mathcal{R}$, all of the same speaker, let $f_{\mathcal{R}}$ be the pooled meta-embedding, such that $f_{\mathcal{R}}(\mathbf{z}) = kP(\mathcal{R} \mid \mathbf{z}, H_1)$, then we can rewrite (3.9) more compactly as:

$$P(\mathbf{z} \mid \mathcal{R}, H_1, \pi) = \pi(\mathbf{z})\overline{f_{\mathcal{R}}}(\mathbf{z}) \tag{3.18}$$

We shall require that all meta-embeddings are *normalizable* in the sense that $\left|\left\langle f, 1\right\rangle\right| < \infty$. By (3.15), $\|f\| < \infty$ is a sufficient condition. Alternatively, any bounded function, $f(\mathbf{z})$ also satisfies this requirement.

3.2.5 Notations for inner product and expectation

Using the defnitions of inner product and elementwise multiplication, notice that $\left\langle f, g\right\rangle = \left\langle fg, \mathbf{1}\right\rangle$ and indeed that $\left\langle fgh, \mathbf{1}\right\rangle = \left\langle fg, h\right\rangle = \left\langle f, gh\right\rangle = \left\langle \mathbf{1}, fgh\right\rangle$, and so on. The comma in the inner product notation has no real function and it can be omitted. We shall write:

$$\left\langle f, g\right\rangle = \left\langle fg\right\rangle = \left\langle f(\mathbf{z})g(\mathbf{z})\right\rangle_{\mathbf{z}\sim\pi} \tag{3.19}$$

and more generally,

$$\left\langle \prod_j f_j \right\rangle = \left\langle \prod_j f_j(\mathbf{z}) \right\rangle_{\mathbf{z}\sim\pi} \tag{3.20}$$

In what follows, we shall interchangeably use the notations with or without comma, to respectively emphasize the dot-product or expectation interpretations.

3.3 Likelihood-ratios

Finally, we demonstrate the utility of our operations on meta-embeddings by expressing the LR formulas of chapter 2 in terms of these operations.

3.3.1 Simple LRs

Here we show how to write the LR of each of the examples of section 2.4 as an inner product between normalized meta-embeddings.

The *simple verification* likelihood ratio (2.3), rewritten in a variety of notations, is:

$$\frac{P(r, r' \mid H_1)}{P(r, r' \mid H_2)} = \frac{\left\langle f, f'\right\rangle}{\left\langle f, \mathbf{1}\right\rangle\left\langle f', \mathbf{1}\right\rangle} = \frac{\left\langle ff'\right\rangle}{\left\langle f\right\rangle\left\langle f'\right\rangle} = \left\langle \overline{f}, \overline{f'}\right\rangle \tag{3.21}$$

where f and f' are the meta-embeddings extracted from r and r'.

The *multi-enroll verification* LR, with enrollments: $r_1 \mapsto f_1$ and $r_2 \mapsto f_2$ and test: $r_3 \mapsto f_3$, is:

$$\frac{P(r_1, r_2, r_3 \mid H_1)}{P(r_1, r_2, r_3 \mid H_2)} = \frac{\langle f_1 f_2, f_3 \rangle}{\langle f_1 f_2, \mathbf{1} \rangle \langle f_3, \mathbf{1} \rangle} = \frac{\langle f_1 f_2 f_3 \rangle}{\langle f_1 f_2 \rangle \langle f_3 \rangle} = \langle \overline{f_1 f_2}, \overline{f_3} \rangle \tag{3.22}$$

For *openset classification*, with enrollment recordings, $\mathcal{R} = \{r_1, \ldots, r_n\}$ for m speakers, indexed by $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq \{1, \ldots, n\}$ and a new test recording, r', we use the meta-embeddings: $\{r_\ell \mapsto f_\ell\}_{\ell=1}^n$ and $r' \mapsto f'$ and we calculate the LR that r' is of speaker i, vs that r' is a new speaker, as:

$$\begin{aligned} L_i &= \frac{\langle f' \prod_{\ell \in \mathcal{S}_i} f_\ell \rangle \prod_{j \neq i} \langle \prod_{\ell \in \mathcal{S}_j} f_\ell \rangle}{\langle f' \rangle \prod_{j=1}^m \langle \prod_{\ell \in \mathcal{S}_j} f_\ell \rangle} \\ &= \frac{\langle f' \prod_{\ell \in \mathcal{S}_i} f_\ell \rangle}{\langle f' \rangle \langle \prod_{\ell \in \mathcal{S}_i} f_\ell \rangle} \\ &= \langle \overline{f'}, \overline{\prod_{\ell \in \mathcal{S}_i} f_\ell} \rangle \end{aligned} \tag{3.23}$$

For *agglomerative clustering* we can similarly write the LR (2.12) in terms of normalized embeddings as:

$$L_{ij} = \frac{\langle \prod_{\ell \in \mathcal{S}_i \cup \mathcal{S}_j} f_\ell \rangle}{\langle \prod_{\ell \in \mathcal{S}_i} f_\ell \rangle \langle \prod_{\ell \in \mathcal{S}_i} f_\ell \rangle} = \langle \overline{\prod_{\ell \in \mathcal{S}_i} f_\ell}, \overline{\prod_{\ell \in \mathcal{S}_j} f_\ell} \rangle \tag{3.24}$$

In summary: In each of these simple cases, the likelihood-ratio is most compactly represented as an *inner product between normalized meta-embeddings*. In what follows, we refer to suchs inner products more concisely as *normalized inner products*.

3.3.2 The general case

If we consider the most general case of LRs of the form (2.7), we find that there are cases that cannot be written as a single normalized inner product. But we show that they *can* however be written as products and ratios of such

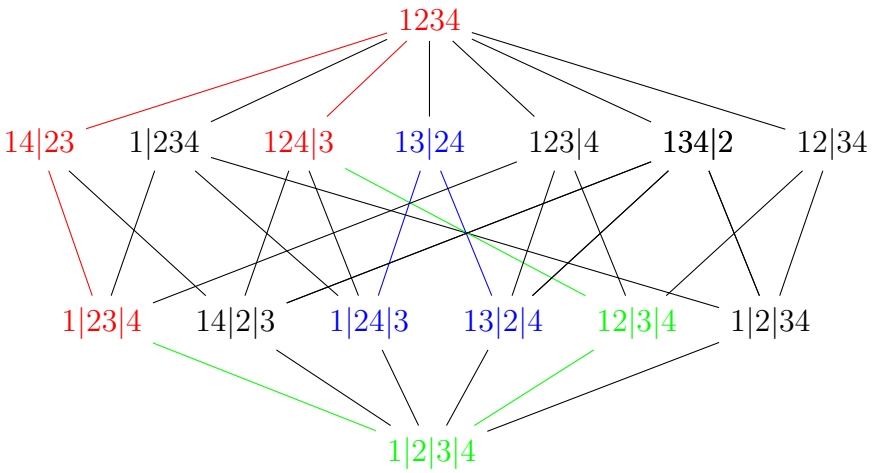


Figure 3.1: Hasse diagram of the lattice of partitions of a set of 4 recordings, ordered by 'refines'. Every downward arc is an an atomic refinement, which splits a single subset. Example 1 is highlighted in blue , example 2a in red and 2b in green.

inner products. We start with an example.

Example 1. Consider a set of four recordings, $\mathcal{R} = \{r_1, r_2, r_3, r_4\}$ with associated meta-embeddings, $\{f_1, f_2, f_3, f_4\}$ and consider the two hypotheses: $A : 1|24|3$ and $B : 13|2|4$, written in a convenient short-hand. The LR is:

$$\begin{aligned}
\frac{P(\mathcal{R} \mid A)}{P(\mathcal{R} \mid B)} &= \frac{\langle f_1 \rangle \langle f_2 f_4 \rangle \langle f_3 \rangle}{\langle f_1 f_3 \rangle \langle f_2 \rangle \langle f_4 \rangle} \\
&= \frac{\langle f_1 \rangle \langle f_2 f_4 \rangle \langle f_3 \rangle}{\langle f_1 f_3 \rangle \langle f_2 f_4 \rangle} \times \frac{\langle f_1 f_3 \rangle \langle f_2 f_4 \rangle}{\langle f_1 f_3 \rangle \langle f_2 \rangle \langle f_4 \rangle} \\
&= \frac{\langle f_1 \rangle \langle f_3 \rangle}{\langle f_1 f_3 \rangle} \times \frac{\langle f_2 f_4 \rangle}{\langle f_2 \rangle \langle f_4 \rangle} \\
&= \frac{\langle \overline{f_2}, \overline{f_4} \rangle}{\langle \overline{f_1}, \overline{f_3} \rangle}
\end{aligned} \tag{3.25}$$

We have achieved this re-arrangement by observing that we can get from A to B via the auxiliary hypothesis $C : 13|24$, where C is reached from A by joining $\{1\}\{3\}$ and B is reached from C by splitting $\{2, 4\}$. The join contributes a normalized inner product factor below the line and the split contributes another above the line.

In general, the partitions of a set form a *partial order*, where $A < C$ is

defined as: *C is finer than A*. In our example, we have both $A < C$ and $B < C$, but A and B are *not directly comparable*. If two hypotheses *are* directly comparable (if they differ by joining or splitting a single subset), then their LR can be expressed as a single normalized inner product, as we have seen in section 3.3.1.

More can be said about this partial order. It is in fact a *lattice*, where every pair of partitions has a unique supremum (least upper bound) and a unique infimum (greatest lower bound). Figure 3.1 shows an example of such a lattice, with example 1 highlighted in blue. To form the LR between any pair of hypotheses, we can choose to traverse the lattice between them via any path connected by arcs, but the shortest paths will go via the supremum or the infimum. Since the infimum and supremum always exist, there will always be a path. Every step (split or join)[3] contributes a normalized inner product factor to the final LR.

In example 1 we traversed via $C = \sup(A, B)$. If we instead traverse via $1|2|3|4 = \inf(A, B)$, we get the *same* decomposition. This is however not true in general—different paths will give different decompositions of the LR. This is shown in the next example.

Example 2a. Let's try a more complex path, highlighted in red in figure 3.1. Let $A : 1|23|4$ and $B : 124|3$. We have $\sup(A, B) = 1234$ and $\inf(A, B) = 1|2|3|4$. Whether we go up or down, we need at least three steps to traverse between A and B. There are three such paths via the supremum and three more via the infimum. Following the red highlighted path, the LR can be re-arranged as:

$$\begin{aligned}\frac{P(\mathcal{R} \mid A)}{P(\mathcal{R} \mid B)} &= \frac{\langle f_1 \rangle \langle f_2 f_3 \rangle \langle f_4 \rangle}{\langle f_1 f_2 f_4 \rangle \langle f_3 \rangle} \\ &= \frac{\langle f_1 \rangle \langle f_2 f_3 \rangle \langle f_4 \rangle}{\langle f_1 f_4 \rangle \langle f_2 f_3 \rangle} \times \frac{\langle f_1 f_4 \rangle \langle f_2 f_3 \rangle}{\langle f_1 f_2 f_3 f_4 \rangle} \times \frac{\langle f_1 f_2 f_3 f_4 \rangle}{\langle f_1 f_2 f_4 \rangle \langle f_3 \rangle} \\ &= \frac{1}{\langle \overline{f_1}, \overline{f_4} \rangle} \times \frac{1}{\langle \overline{f_1 f_4}, \overline{f_2 f_3} \rangle} \times \langle \overline{f_1 f_2 f_4}, \overline{f_3} \rangle\end{aligned} \tag{3.26}$$

Each upward step (join) contributes a normalized inner product below the line, while the downward step (split) contributes another above.

Example 2b. Let's also try a path via the infimum, highlighted in green, to traverse between the same two nodes as before. Now the LR can be

[3] In lattice theory, infimum and supremum are alternatively termed *meet* and *join*. In our usage here, we mean by *join* simply to form the union of two subsets of recordings.

re-arranged as:

$$\begin{aligned}\frac{P(\mathcal{R} \mid A)}{P(\mathcal{R} \mid B)} &= \frac{\langle f_1 \rangle \langle f_2 f_3 \rangle \langle f_4 \rangle}{\langle f_1 f_2 f_4 \rangle \langle f_3 \rangle} \\ &= \frac{\langle f_1 \rangle \langle f_2 f_3 \rangle \langle f_4 \rangle}{\langle f_1 \rangle \langle f_2 \rangle \langle f_3 \rangle \langle f_4 \rangle} \times \frac{\langle f_1 \rangle \langle f_2 \rangle \langle f_3 \rangle \langle f_4 \rangle}{\langle f_1 f_2 \rangle \langle f_3 \rangle \langle f_4 \rangle} \times \frac{\langle f_1 f_2 \rangle \langle f_3 \rangle \langle f_4 \rangle}{\langle f_1 f_2 f_4 \rangle \langle f_3 \rangle} \\ &= \langle \overline{f_2}, \overline{f_3} \rangle \times \frac{1}{\langle \overline{f_1}, \overline{f_2} \rangle} \times \frac{1}{\langle \overline{f_1 f_2}, \overline{f_4} \rangle}\end{aligned} \tag{3.27}$$

Again, each upward step (join) contributes a normalized inner product below the line, while the downward step (split) contributes another above. This rule applies in general. If we follow a path from the numerator (hypothesis A), to the denominator (hypothesis B), joins contribute below the line, while splits contribute above. (If we go from denominator to numerator, this rule is reversed.)

Equivalence of paths

As the two paths that we examined in example 2 demonstrate, there may be many different paths and therefore many different decompositions of a given LR in terms of normalized inner products. Nevertheless, as long as our calculations are exact, all such decompositions must give the same numerical result. This may serve as a useful regularization constraint if we are training neural nets to approximate inner products between meta-embeddings. (As we shall see in the next chapter, there may be various practical reasons to prefer approximate calculations.)

The equivalence of paths is not limited to shortest paths. The products and ratios of the normalized inner products of any path between two nodes must give the same numerical result, irrespective of the path. This is also true in particular of *cycles*. If we start and end at the same node, the LR we calculate thus is *unity*. The interesting thing about this is that if we are given a set of unlabelled recordings, we can use this fact to manufacture a (very) large set of constraints on the values of normalized inner products. (If we work with logarithms, this becomes a homogeneous system of linear equations.) These constraints may perhaps help to regularize neural nets in both labelled and unlabelled training regimes.

Primitive operations

Which primitive building blocks do we need to compute general LRs of the form (2.7)? Let's assume we can always extract raw (unnormalized) meta-

embeddings from any recording. Then, as (2.7) shows, we can compute any LR if we can:

- arbitrarily pool meta-embeddings (do elementwise multiplication) and
- compute arbitrary expectations of single or pooled meta-embeddings.

Alternatively, as we have shown in this section, we can compute arbitrary LRs if we can

- pool and
- compute arbitrary normalized inner products.

We can however somewhat restrict the requirement on normalized inner products. We can get away with inner products where *one of the arguments is always a single (unpooled) meta-embedding*. To do normalization, we need $\langle f, \mathbf{1} \rangle$, which respects this restriction. To further enforce this restriction, look again at the lattice in figure 3.1 and convince yourself that you can traverse between any two nodes by always adding or removing but a single element of some subset in the partition. For example, the red path does not respect this restriction, but the blue and green ones do. For such paths, all the inner product factors will have at least one unpooled argument.

When we discuss practical representations of meta-embeddings in the next chapter, we will see that for some of the representations, pooling may change the form of the representation. The form may have implications on how easy it is to do inner products. If one of the arguments is always unpooled, this may facilitate calculations.

Finally, we mention another interesting building block. It is not difficult to show that we can form arbitrary LRs from factors of the form:

$$\frac{\langle \prod_{i=1}^{m} f_i \rangle}{\prod_{i=1}^{m} \langle f_i \rangle} = \left\langle \prod_{i=1}^{m} \overline{f_i} \right\rangle \tag{3.28}$$

for $m = 1, 2, \ldots$. This factor is just the LR comparing the coarsest to the finest partition of a set of m recordings. It can be argued that this building block is not primitive, because it can be assembled via pooling and expectation (or pooling and inner products).

3.3.3 Note on cosine similarity

It is of interest to note that the normalized inner products that form our LRs are very similar to the popular cosine similarity,

$$\frac{\langle \mathbf{e}_1, \mathbf{e}_2 \rangle}{\sqrt{\|\mathbf{e}_1\|_2 \, \|\mathbf{e}_2\|_2}},$$

between embeddings $\mathbf{e}_1$ and $\mathbf{e}_2$. The difference is the use of L1 vs L2 normalization.

In the literature, there are many examples where embeddings are L2-normalized before computing dot products or distances. In speaker recognition, the first i-vector scoring recipe used cosine-similarity [3], and indeed in current PLDA scoring recipes, it is still standard practice to use L2-normalized i-vectors [28]. In face recognition, Facenet uses cosine similarity to compare embeddings [2], although subsequently [29] advises against such normalization. In [9], the embeddings for language recognition are compared with cosine similarity.

Notice that by (3.4), if embeddings have unit L2 norm, then there is no essential difference (other than the sign) between cosine similarity (inner product) and squared distance. By the Cauchy-Schwartz inequality (3.5), inner products between L2-normalized embeddings are limited to the range $[-1, 1]$, where -1 corresponds to embeddings on opposite sides of the unit hypersphere (maximum distance) and 1 to embeddings that coincide (zero distance).

In speaker recognition, cosine similarity between embeddings in Euclidean space is regarded as an uncalibrated form of *log* likelihood-ratio. (The cosine similarity is signed, just like the log likelihood-ratio.) In contrast, our inner products between L1-normalized meta-embeddings give non-negative *likelihood-ratios*. Notice also that by (3.14), our likelihood-ratio magnitudes are not limited to be at most unity.

3.4 Meta-embedding example

Let us now construct a graphical example to reinforce our geometric understanding of our meta-embeddings. In order to be able to plot the meta-embeddings, we choose perhaps the simplest possible meta-embeddings, which live in $\mathbb{R}^2$.

We do this by imagining a very weak speaker recognizer (meta-embedding extractor) that can only distinguish voices by whether they sound male or female. That is, we choose the simplest possible hidden speaker identity variable by making it discrete and binary: $\mathbf{z} \in \{\mathtt{m}, \mathtt{f}\}$. We choose the prior to have males and females equally likely:

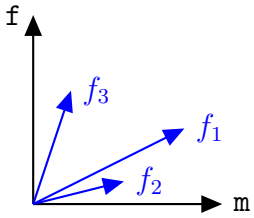
$$\pi(\mathtt{m}) = \pi(\mathtt{f}) = \frac{1}{2}$$

Seen as a function, the meta-embedding for recording r_i is $f_i(\mathbf{z}) = k_i P(r_i \mid \mathbf{z})$. Since $\mathbf{z}$ has only two possible values, we can represent the meta-embedding

as a *two-dimensional vector*: $f_i = \big[k_i P(r_i \mid \mathtt{m}), k_i P(r_i \mid \mathtt{f})\big]$. Because of the prior weighting, $\pi(\mathtt{m}) = \pi(\mathtt{f}) = \frac{1}{2}$, the inner product and dot product differ by a factor 2: if $f_i = [f_{i1}, f_{i2}]$, then:

$$\big\langle f_i, f_j \big\rangle = \frac{1}{2}(f_{i1}f_{j1} + f_{i2}f_{j2}) \tag{3.29}$$

For three hypothetical recordings, let the meta-embeddings be $f_1 = [2, 1]$, $f_2 = [1.2, 0.3]$, $f_3 = [0.5, 1.5]$ and we can plot these meta-embeddings as:



The first thing to notice is the difference between embedding and meta-embedding: The ideal embedding here is binary, $\mathbf{z} \in \{\mathtt{m}, \mathtt{f}\}$, while the meta-embedding is continuous: $f_i \in \mathbb{R}^2$. Here, and in general, the meta-embedding lives in a more complex space than $\mathbf{z}$.

Since likelihoods are positive, the meta-embeddings are confined to the positive quadrant. We should never be working outside of this quadrant, but to see the space as a vector space, we need to be aware of the existence of the other quadrants. Indeed, when we design our neural nets to extract meta-embeddings, we will have to make sure they end up being in the positive quadrant.

Because of the arbitrary scaling constants, k_i, the lengths of our meta-embedding vectors do not carry information—but from the directions we see that f_1 and f_2 are probably male, while f_3 is probably female. Our end-goal is not to infer the genders of the speakers, but to infer whether the speakers of different recordings are the same or not. A little thought should convince the reader that:

- The smallest possible likelihood-ratio, $\frac{P(r_i, r_j|H_1)}{P(r_i, r_j|H_2)} = 0$, would be obtained if we were certain that one speaker is male and the other female. This would happen when we have:

  $$P(r_i \mid \mathtt{m}) \gg P(r_i \mid \mathtt{f}) \qquad \text{and} \qquad P(r_j \mid \mathtt{m}) \ll P(r_j \mid \mathtt{f})$$

  so that f_i is on the horizontal (male) axis, while f_j is on the vertical (female) axis. Then indeed, the inner product (and dot product) between these two orthogonal meta-embeddings would be zero and so would the LR.

- The largest possible LR with the weak, binary hidden variable would be just 2, which would be obtained when we are certain that both recordings are of the same gender. We need to consider details of the normalization to see how this works out.

Figure 3.2 shows the same three meta-embeddings, together with their normalized versions, with normalization as defined by (3.16). Remember the prior-weighting: when $f_i = [f_{i1}, f_{i2}]$, then the normalization constant is

$$\langle f_i, \mathbf{1} \rangle = \frac{1}{2}(f_{i1} + f_{i2})$$

Inner products between the normalized meta-embeddings give the likelihood-ratios,

$$\frac{P(r_i, r_j \mid H_1)}{P(r_i, r_j \mid H_2)} = \langle \overline{f_i}, \overline{f_j} \rangle$$

Since the discrimination given by the binary hidden variable is weak, the likelihood-ratios are not too far from the neutral value of 1. The LR, $\langle \overline{f_1}, \overline{f_2} \rangle = 1.2$ is greater than 1, slightly favouring the hypothesis that the speakers of r_1 and r_2 are the same. The other two LRs are less than one, favouring the H_2 hypothesis in each case. The LR, $\langle \overline{f_2}, \overline{f_3} \rangle = 0.64$ is *stronger* (further from 1) than $\langle \overline{f_1}, \overline{f_3} \rangle = 0.8$, because we are more certain of the maleness of f_2 than we are of f_1. The largest possible LR of 2 would be obtained when both normalized meta-embeddings coincide on the horizontal or vertical axis, at $[2, 0]$, or $[0, 2]$.[4]

Figure 3.3 shows what happens when we know that r_1 and r_2 are of the same speaker and we pool their meta-embeddings, giving us more certainty that this is a male speaker and more certainty that this speaker is different from the (probably female) speaker of r_3.

[4]Remember the prior weighting of $\frac{1}{2}$.

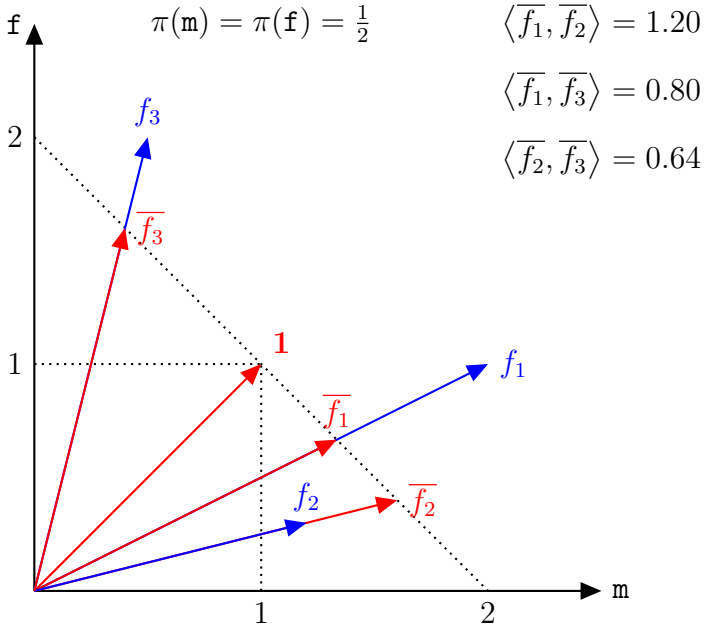


Figure 3.2: Normalized meta-embeddings (red) and their inner products. Raw meta-embeddings (blue) are normalized by scaling them so that their projections on **1** are unity.

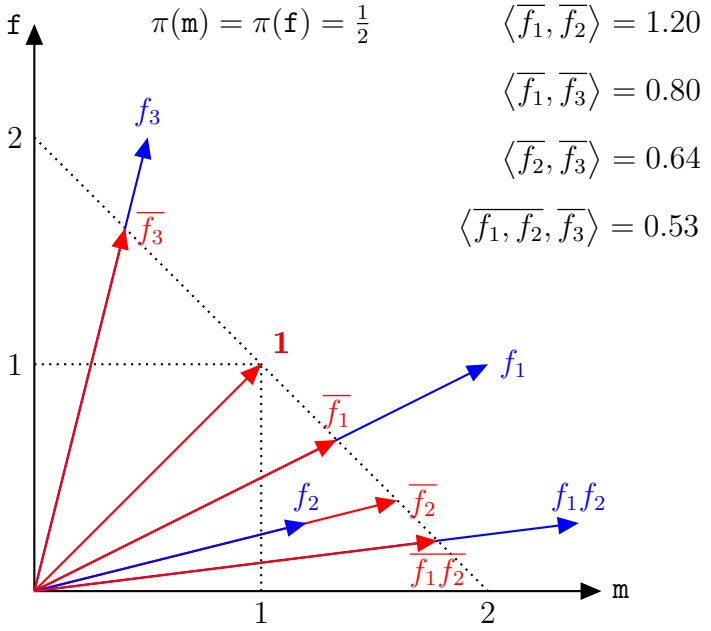


Figure 3.3: Pooling: If we know that f_1 and f_2 are from the same speaker, we can pool them, using the elementwise product $f_1 f_2$. This increases the certainty that this is a male speaker. Correspondingly, the strength (difference from 1) of $\langle \overline{f_1 f_2}, \overline{f_3} \rangle$ is more than either of $\langle \overline{f_1}, \overline{f_3} \rangle$ and $\langle \overline{f_2}, \overline{f_3} \rangle$. Pooling has increased the certainty that the speaker represented by $f_1 f_2$ is not the same speaker as the probably female f_3.

Chapter 4

Practical meta-embeddings

We have thus far developed a theoretical idea of the nature of meta-embeddings. We are now ready to explore a few proposals of how to practically represent meta-embeddings. The end-goal is to train a neural net that takes voice recordings as input and outputs representations for the corresponding meta-embeddings. We shall discuss training criteria for the neural net later. Here we are interested in the form that practical meta-embedding representations might take and we shall explore several possibilities.

Before proceeding to these possibilities, let us consider in general, some desirable properties of our representations. Ideally, we need all of the following properties for our meta-embeddings:

Non-negativity: $f(\mathbf{z}) \geq 0$ everywhere.

Normalizability: $\langle f, \mathbf{1} \rangle < \infty$.

Pooling should be tractable, and the pooled result, $f_i f_j$, should have the same representation as f_i and f_j.

Expectation: $\langle f(\mathbf{z}) \rangle_{\mathbf{z} \sim \pi}$ should be tractable for any meta-embedding f, whether it is raw or pooled.

Backpropagation of derivatives through pooling and expectation is necessary for training.

Recall that if we can pool and do expectations, then we can also do any inner products, since $\langle f, g \rangle = \langle fg \rangle$.

The first two proposals below, based on exponential family distributions, meet all of these requirements, although the expectations are somewhat computationally expensive. In some of the other proposals, various approximations and compromises have to be made.

4.1 Multivariate Gaussian

This recipe supposes a continuous speaker identity variable, $\mathbf{z} \in \mathbb{R}^d$. A trainable neural net processes each recording, r_j, and outputs the corresponding meta-embedding, f_j, in the form of a $(d + D)$-dimensional representation:

$$r_j \mapsto \mathbf{e}_j = (\mathbf{a}_j, \mathbf{b}_j)$$

where $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{b}_j = (b_{1j}, \ldots, b_{Dj}) \in \mathbb{R}^D_+$, where $D \geq 1$ and the $b_{ij} \geq 0$. While the representation, $\mathbf{e}_j$, is finite-dimensional, it parametrizes the *infinite-dimensional* meta-embedding, f_j, defined as:

$$f_j(\mathbf{z}) = \exp\big[\mathbf{a}_j'\mathbf{z} - \frac{1}{2}\mathbf{z}'\mathbf{B}_j\mathbf{z}\big] \tag{4.1}$$

where $\mathbf{B}_j$ is a d-by-d, positive semi-definite precision matrix, composed as a conical combination:

$$\mathbf{B}_j = \sum_{i=1}^{D} b_{ij}\mathbf{E}_i \tag{4.2}$$

where $\{\mathbf{E}_i\}_{i=1}^D$ are fixed, d-by-d, positive semi-definite definite matrices—they are fixed in the sense of being independent of the input data (recordings), but the elements of these matrices are still trainable, together with the parameters of the neural net that extracts the $\mathbf{e}_j$. These matrices can be full, low rank, diagonal, etc.

Since $\mathbf{z} \in \mathbb{R}^d$ is hidden, we are free to choose the dimensionality, d. Experience with PLDA suggests $100 \leq d \leq 200$ is a good choice. This can also be compared to the Facenet embeddings, which are 120-dimensional. The size, D, of the precision parametrization can be used to trade off computational complexity vs capacity. To keep the complexity significantly less than that which would be needed for fully specified precision matrices, we probably want to constrain $D \ll \frac{d(d+1)}{2}$.

Scalar multiplication could be easily done in this framework but in practice, we probably won't need it. Addition of the meta-embeddings would give mixtures of Gaussians, with more complex representations, but for this representation we don't need addition.

4.1.1 Elementwise product

The important elementwise product is done by simply adding the representation vectors. For $\mathcal{R} = \{r_1, \ldots, r_n\}$, we can extract the individual representations, $\{\mathbf{e}_j\}_{j=1}^n$, which respectively represent the $\{f_j\}_{j=1}^n$. The representation

for $f_{\mathcal{R}} = \prod_{j=1}^{n} f_j$ is then $\mathbf{e}_{\mathcal{R}} = \sum_{j=1}^{n} \mathbf{e}_j$. The representation $\mathbf{e}_j$ is essentially logarithmic, which gives us the benefit of automatic positivity, as well as easy elementwise multiplication. The disadvantage is somewhat complex expectation computation.

Note on group structure

For future reference, it might be useful to note that since Gaussians are closed under (associative) multiplication, our space of Gaussian meta-embeddings is a *monoid* (a semigroup, with identity). Since we exclude precision matrices with negative eigenvalues, our monoid lacks inverse elements, which would have made it a full group. The identity element is a Gaussian with zero mean and zero precision, which is just our previously defined identity, $\mathbf{1}(\mathbf{z}) = 1$.

4.1.2 Prior

To do expectations we need to define the prior. The simplest choice is the standard Gaussian:

$$\pi(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}) = \frac{e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}}{\sqrt{(2\pi)^d}} \tag{4.3}$$

4.1.3 Expectation

Since our logarithmic representation is closed under elementwise multiplication, our expectations can always be performed on representations of the form (4.1). Dropping the subscript j to avoid clutter, we derive an expression for $E(\mathbf{a}, \mathbf{B}) = \big\langle f \big\rangle$, where $f(\mathbf{z}) = \exp[\mathbf{a}'\mathbf{z} - \frac{1}{2}\mathbf{z}'\mathbf{B}\mathbf{z}]$:

$$\begin{aligned} E(\mathbf{a}, \mathbf{B}) &= \big\langle f \big\rangle \\ &= \int_{\mathbb{R}^d} f(\mathbf{z})\pi(\mathbf{z})\, d\mathbf{z} \\ &= \int_{\mathbb{R}^d} f(\mathbf{z})\mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})\, d\mathbf{z} \\ &= \int_{\mathbb{R}^d} \frac{\exp\big[\mathbf{a}'\mathbf{z} - \frac{1}{2}\mathbf{z}'(\mathbf{I} + \mathbf{B})\mathbf{z}\big]}{\sqrt{(2\pi)^d}}\, d\mathbf{z} \end{aligned} \tag{4.4}$$

If we define $\boldsymbol{\mu} = (\mathbf{I} + \mathbf{B})^{-1}\mathbf{a}$, we can rewrite this as:

$$\begin{aligned}
E(\mathbf{a}, \mathbf{B}) &= \big\langle f \big\rangle \\
&= \int_{\mathbb{R}^d} \frac{\exp\big[\boldsymbol{\mu}'(\mathbf{I} + \mathbf{B})\mathbf{z} - \frac{1}{2}\mathbf{z}'(\mathbf{I} + \mathbf{B})\mathbf{z}\big]}{\sqrt{(2\pi)^d}}\, d\mathbf{z} \\
&= \frac{\exp\big[\frac{1}{2}\boldsymbol{\mu}'(\mathbf{I} + \mathbf{B})\boldsymbol{\mu}\big]}{|\mathbf{I} + \mathbf{B}|^{\frac{1}{2}}} \int_{\mathbb{R}^d} \frac{|\mathbf{I} + \mathbf{B}|^{\frac{1}{2}}}{\sqrt{(2\pi)^d}} \exp\big[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})'(\mathbf{I} + \mathbf{B})(\mathbf{z} - \boldsymbol{\mu})\big]\, d\mathbf{z} \\
&= \frac{\exp\big[\frac{1}{2}\boldsymbol{\mu}'(\mathbf{I} + \mathbf{B})\boldsymbol{\mu}\big]}{|\mathbf{I} + \mathbf{B}|^{\frac{1}{2}}} \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, (\mathbf{I} + \mathbf{B})^{-1})\, d\mathbf{z} \\
&= \frac{\exp\big[\frac{1}{2}\boldsymbol{\mu}'(\mathbf{I} + \mathbf{B})\boldsymbol{\mu}\big]}{|\mathbf{I} + \mathbf{B}|^{\frac{1}{2}}} \\
&= \frac{\exp\big[\frac{1}{2}\mathbf{a}'\boldsymbol{\mu}\big]}{|\mathbf{I} + \mathbf{B}|^{\frac{1}{2}}}
\end{aligned} \tag{4.5}$$

Both $\boldsymbol{\mu}$ and the determinant $|\mathbf{I} + \mathbf{B}|$ can be found by Cholesky decomposition of the positive definite matrix $\mathbf{I} + \mathbf{B}$.

4.1.4 Stochastic expectation

Pooling is fast (addition) but expectation is slower (Cholesky decomposition). Unfortunately, our recipe above requires a new Cholesky decomposition for every LR calculation. In applications where many verification trials are processed on limited hardware, this will have a significant impact on speed. Even if we have enough CPU power available at runtime for some applications, training remains a problem because we are going to be training our meta-embedding extractors with discriminative criteria that typically require evaluating very many trials. Let us therefore consider some plans for speeding up scoring of trials by making stochastic approximations.

We envisage that such stochastic approximation could work out to be especially cheap at training time, in the same way that stochastic evaluation of the ELBO in *variational autoencoder* (VAE) training typically requires but a single stochastic sample per training example [30]. Given sufficiently many examples, the stochastic approximation errors tend to cancel, provided the errors are independent.

We will not be too concerned about expectations of a single factor, of the form $\big\langle f \big\rangle$. These expectations are required to normalize trial sides (typically meta-embeddings extracted from single recordings) and there are far fewer

trial sides than there are trials. The calculations for which we need speed, are expectations of the form $\langle \overline{f}_i \, \overline{f}_j \rangle$. We consider three solutions below.

Prior sampling

An obvious choice, perhaps naive, would be to form the expectation by sampling from the prior, π. The expectation could be approximated via m such samples as:

$$\left\langle \overline{f}_i(\mathbf{z})\overline{f}_j(\mathbf{z}) \right\rangle_{\mathbf{z}\sim\pi} \approx \frac{1}{m} \sum_{\tilde{\mathbf{z}}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \overline{f}_i(\tilde{\mathbf{z}})\overline{f}_j(\tilde{\mathbf{z}}) \tag{4.6}$$

We fear this might be inaccurate, because in high-dimensional space, samples from π will be very unlikely to hit the peak[1] of $\overline{f}_i \, \overline{f}_j$ and therefore an affordable number of samples might not be able to sufficiently explore the volume under the peak. Experiments might be needed to check whether this is indeed a problem. Let us nevertheless explore in more detail the calculation required for each sample:

$$\begin{aligned} \overline{f}_i(\tilde{\mathbf{z}})\overline{f}_j(\tilde{\mathbf{z}}) &= c_i c_j \exp\left[\mathbf{a}_i'\tilde{\mathbf{z}} + \mathbf{a}_j'\tilde{\mathbf{z}} - \frac{1}{2}\tilde{\mathbf{z}}'\mathbf{B}_i\tilde{\mathbf{z}} - \frac{1}{2}\tilde{\mathbf{z}}'\mathbf{B}_j\tilde{\mathbf{z}}\right] \\ &= c_i c_j \exp\left[\mathbf{a}_i'\tilde{\mathbf{z}} + \mathbf{a}_j'\tilde{\mathbf{z}} - \frac{1}{2}\mathbf{b}_i'\tilde{\boldsymbol{\gamma}} - \frac{1}{2}\mathbf{b}_j'\tilde{\boldsymbol{\gamma}}\right] \end{aligned} \tag{4.7}$$

where using (4.2), we have defined $\tilde{\boldsymbol{\gamma}} = [\tilde{\gamma}_1, \ldots, \tilde{\gamma}_D] \in \mathbb{R}^D$, with $\tilde{\gamma}_\ell = \tilde{\mathbf{z}}'\mathbf{E}_\ell\tilde{\mathbf{z}}$ and where c_i, c_j reflect the normalizations. The meta-embedding representations, $\mathbf{a}_i, \mathbf{b}_i$ and $\mathbf{a}_j, \mathbf{b}_j$, and the trainable constants, $\{\mathbf{E}_\ell\}_{\ell=1}^D$, are as defined in the beginning of section 4.1.

Notice that $\tilde{\boldsymbol{\gamma}}$ is independent of the meta-embeddings and therefore independent of the data and can be precomputed. If the $\mathbf{E}_\ell$ are constrained to be either diagonal or rank-one, then the calculations $\tilde{\mathbf{z}}'\mathbf{E}_\ell\mathbf{z}$ will be cheap. Keep in mind that we need m different samples, $\tilde{\mathbf{z}}$, and therefore also m different versions of $\tilde{\boldsymbol{\gamma}}$. In stochastic minibatch training, if we update the $\{\mathbf{E}_\ell\}$ after every minibatch, we will have to update all of the $\tilde{\boldsymbol{\gamma}}$ also, in which case it might make sense to also resample m new values for $\tilde{\mathbf{z}}$. This type of strategy, which is well-known also in VAE [30], is termed *doubly stochastic* by Michalis Titsias [31].

An advantage of (4.7) is that it generalizes to expectations of more (or fewer) than two factors—the argument in the exponent will have independent terms for each of the factors in the expectation.

[1] The product is still Gaussian and therefore has single peak.

Experiments would have to be conducted to verify if an affordably small m could give reasonable accuracy. Keep in mind that in such experiments, we can compare accuracy against the exact calculation (4.5).

Posterior sampling

If meta-embedding f_i represents recording r_i, then

$$\pi(\mathbf{z})\overline{f_i}(\mathbf{z}) = P(\mathbf{z} \mid r_i, \pi)$$

is a properly normalized Gaussian, from which we can sample.[2] We can now rewrite the expectation as:

$$\left\langle \overline{f_i}(\mathbf{z})\overline{f_j}(\mathbf{z}) \right\rangle_{\mathbf{z}\sim\pi} = \left\langle \overline{f_j}(\tilde{\mathbf{z}}) \right\rangle_{\tilde{\mathbf{z}}\sim\pi\overline{f_i}} \approx \frac{1}{m} \sum_{\tilde{\mathbf{z}}\sim\pi\overline{f_i}} c_j \exp\left[\mathbf{a}_j'\tilde{\mathbf{z}} - \frac{1}{2}\mathbf{b}_j'\tilde{\boldsymbol{\gamma}}\right] \tag{4.8}$$

This might be more accurate than prior sampling, because when f_i and f_j represent the same speaker, their peaks should overlap. If they are of different speakers, their peaks might be far apart, but then we do want a value close to zero for the expectation. With this variant, we may be able to use smaller values for m.

This method is asymmetric. Which version is best—should we sample from f_i, or from f_j? If f_i is much sharper than f_j, so that f_j is almost constant compared to f_j, then one sample would suffice, while if we sampled the other way round, we would need many samples to properly explore the volume of f_i. Pooling generally sharpens peaks, so it would be more accurate to sample from the side that consists of the most pooled factors. Unfortunately, fast pooling and sampling are conflicting requirements—after adding natural parameters to compute the pooling, we need a Cholesky factorization before we can sample.

Whether we pool or not, a disadvantage of posterior sampling vs prior sampling is that we have reintroduced the need for Cholesky factorization. Fortunately, we need this only for one side of the trial and we are still avoiding per-trial Cholesky decompositions. Also note that for the sampled side of the trial we do not need to separately compute the normalization constant, because the samples already come from the normalized distribution.

Another option would be to change the representation. We could let:

$$\mathbf{B}_i = \Big(\sum_{\ell=1}^{D} b_{\ell i}\mathbf{T}_\ell\Big)\Big(\sum_{\ell=1}^{D} b_{\ell i}\mathbf{T}_\ell\Big)' \tag{4.9}$$

[2]The same is not always true of f_i or $\overline{f_i}$, which might have a singular precision, in which case sampling could not be done.

where the $\mathbf{T}_\ell$ are triangular and the $b_{\ell i}$ are no longer constrained to be non-negative. This gives a free Cholesky decomposition, and fast, approximate evaluation of LRs of the form $\langle \overline{f_i}, \overline{f_j} \rangle$, but it complicates pooling for more complex LR calculations.

4.2 Zero-mean Gaussians

Next, let us place a restriction on our multivariate Gaussian embeddings in the quest for fast and accurate stochastic expectation solutions. By restricting meta-embeddings to *zero-mean Gaussians with strictly positive definite precisions*, the Fourier transforms of these Gaussians always exist and are also multivariate Gaussians—from which we can sample in the frequency domain. By Parseval's theorem, we can compute the expectation integrals instead in the frequency domain. By the convolution theorem, products of Gaussians in the original $\mathcal{Z}$ domain become convolutions in the frequency domain. These convolutions can be stochastically implemented by *adding* samples drawn from individual embeddings. We now have made both pooling and expectation *simultaneously easy*—a property which has thus far eluded us. The disadvantage is that we have lost the capacity to have means and it is not immediately obvious whether this would be a serious drawback for the application at hand. A more detailed explanation follows.

4.2.1 Representation

Let the Fourier transform of meta-embedding f_i be denoted by $\tilde{f}_i$. If f_i is a zero-mean Gaussian with strictly positive definite precision, then $\tilde{f}_j$ is also a zero-mean Gaussian, having covariance equal to the precision of f_i. Since there is a bijection between f_i and $\tilde{f}_i$, we may as well directly extract representations for the $\tilde{f}_i$. Moreover, these representations can be conveniently chosen to facilitate sampling. That is, every $\tilde{f}_i$ is represented by a d-by-d square (possibly triangular) matrix. Sampling is done by forming matrix vector products, where the d-dimensional vectors are sampled from the standard Gaussian. This is just the *reparametrization trick*, familiar from the VAE literature, which plays well with gradient backpropagation through the stochastic part of the calculation [30].

More parsimonious representations are possible. We can represent each meta-embedding via *samples*, rather than via a matrix of parameters. The net that extracts the embeddings is simply required to produce Gaussian samples, using any convenient method. One way to implement this is to

generate samples from a factor analysis model—where samples drawn from a full rank diagonal Gaussian are added to D-dimensional samples multiplied by a D-by-d factor loading matrix, where typically $D \ll d$. The parameter count of the factor-analysis representation is $d(D+1)$, rather than d^2 for full, or $\frac{d(d+1)}{2}$ for triangular covariance square roots.

4.2.2 Frequency domain pooling and expectation

We are interested in expectations of the form $\left\langle \prod_{i=1}^{m} f_i \right\rangle_\pi$, which we rewrite via Parseval's and convolution theorems as:

$$\int_{\mathbb{R}^n} \pi(\mathbf{z}) \prod_{i=1}^{m} f_i(\mathbf{z})\, d\mathbf{z} = \int_{\mathbb{R}^n} \pi(\mathbf{x}) \tilde{f}_1(\mathbf{x}) * \tilde{f}_2(\mathbf{x}) * \cdots * \tilde{f}_m(\mathbf{x})\, d\mathbf{x} \tag{4.10}$$

where $*$ denotes convolution and where we do not need complex conjugation, because everything is real-valued. Since π is standard Gaussian, its Fourier transform is also standard Gaussian. The RHS is the expected value of π, w.r.t. a Gaussian, formed by the convolution of m zero-mean Gaussians, $\{\tilde{f}_i\}$. The sum of independent samples of all the $\tilde{f}_i$ will give a sample from the convolution. These sums can then be inserted into π and averaged to approximate the required expectation.

4.2.3 Could it work?

4.3 Exponential family Gaussian mixture

The multivariate Gaussian of the previous section is of course an exponential family distribution. We can derive very similar recipes—also with logarithmic representations—from other exponential families. We develop one such example here.

Although a Gaussian mixture is not an exponential family, the *joint* distribution for the continuous and discrete (state) variable *is* exponential family. We do this with a D-component mixture of d-dimensional Gaussians, by choosing our hidden speaker identity variable as: $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = (x_1, \ldots, x_D)$ is a one-hot vector of size D, while $\mathbf{y} \in \mathbb{R}^d$. Our meta-embedding for recording j is:

$$f_j(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{D} \Big(w_{ij} \exp\big[\mathbf{a}'_{ij}\mathbf{y} - \frac{1}{2}\mathbf{y}'\mathbf{B}_{ij}\mathbf{y}\big] \Big)^{x_i} \tag{4.11}$$

where $x_i \in \{0, 1\}$, $w_{ij} > 0$, $\mathbf{a}_{ij} \in \mathbb{R}^d$ and the $\mathbf{B}_{ij}$ are d-by-d positive semi-definite. To see that this is exponential family, take the logarithm and re-arrange:

$$\log f_j(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D} x_i \log w_{ij} + (x_i \mathbf{y}')\mathbf{a}_{ij} - \frac{1}{2} \operatorname{tr}\big[(x_i \mathbf{y}\mathbf{y}')\mathbf{B}_{ij}\big] \tag{4.12}$$

where we see the sufficient statistics are: $\{x_i, x_i\mathbf{y}, x_i\mathbf{y}\mathbf{y}'\}_{i=1}^{D}$ and the natural parameters are $\{\log w_{ij}, \mathbf{a}_{ij}, \mathbf{B}_{ij}\}_{i=1}^{D}$. To form more compact representations, we can let the natural parameters be linear functions of smaller vectors (of which some components have to be constrained to be non-negative). As above, elementwise multiplication (pooling) is accomplished by simple vector addition of these representations.

4.3.1 Prior

We choose a parameterless prior, of the same form as the meta-embeddings:

$$\pi(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{D} \Big(\frac{\mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{I})}{D} \Big)^{x_i} \tag{4.13}$$

and

$$\begin{aligned} \log \pi(\mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{D} x_i \big(-\frac{1}{2}\mathbf{y}'\mathbf{y} - \frac{d}{2}\log(2\pi) - \log(D) \big) \\ &= -\frac{d}{2}\log(2\pi) - \log(D) + \sum_{i=1}^{D} -\frac{1}{2} \operatorname{tr}\big[(x_i \mathbf{y}\mathbf{y}')\mathbf{I}\big] \end{aligned} \tag{4.14}$$

As in the multivariate Gaussian case, the practical function of the prior is to add $\mathbf{I}$ to the possibly semi-definite matrices, $\mathbf{B}_{ij}$, of the meta-embeddings, to make them properly positive definite.

4.3.2 Expectation

Dropping the subscript j, let $f(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{D} \Big(w_i \exp\big[\mathbf{a}_i'\mathbf{y} - \frac{1}{2}\mathbf{y}'\mathbf{B}_i\mathbf{y}\big] \Big)^{x_i}$, then:[3]

$$\begin{aligned} E\big(\{w_i, \mathbf{a}_i, \mathbf{B}_i\}_{i=1}^{D}\big) &= \big\langle f \big\rangle \\ &= \sum_{i=1}^{D} \frac{w_i}{D} \int_{\mathbb{R}^d} f(\mathbf{x} = i, \mathbf{y}) \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{I})\, d\mathbf{y} \\ &= \sum_{i=1}^{D} \frac{w_i}{D} \int_{\mathbb{R}^d} \exp\big[\mathbf{a}_i'\mathbf{y} - \frac{1}{2}\mathbf{y}'\mathbf{B}_i\mathbf{y}\big] \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{I})\, d\mathbf{y} \\ &= \sum_{i=1}^{D} \frac{w_i}{D} \frac{\exp\big[\frac{1}{2}\mathbf{a}_i'\boldsymbol{\mu}_i\big]}{|\mathbf{I} + \mathbf{B}_i|^{\frac{1}{2}}} \end{aligned} \tag{4.15}$$

where $\boldsymbol{\mu}_i = (\mathbf{I} + \mathbf{B}_i)^{-1}\mathbf{a}_i$.

This computation is much the same as for the multivariate Gaussian case, except that we have to do D such calculations every time. Obviously, if we want to use a large value for D, then the individual calculations need to be cheap. Since Cholesky decomposition of full matrices $\mathbf{I} + \mathbf{B}_i$ requires $\mathcal{O}(d^3)$ computation, we need to simplify these calculations. There are various way to do that. We can let the $\mathbf{B}_{ij}$ differ from each other by scaling, by low-rank modifications, or by forming them via Kronecker products, etc.

4.4 Mixtures

Let us compare the previous solution with one where the state is not considered part of the speaker identity variable. The meta-embedding for recording r_j is:

$$f_j(\mathbf{z}) = \sum_{i=1}^{D} w_{ij} \tag{4.16}$$

[3] We abuse notation by writing $\mathbf{x} = i$ to indicate that component i is the hot element (value 1) in the otherwise zero one-hot vector, $\mathbf{x}$.

4.5 Free form, inspired by exponential family distribution

4.6 Discrete Factorial

4.7 Mixture with fixed components

4.8 Mixture with shifted components

4.9 Kernel approximation

Let us think in terms of inner products, rather than expectations. If we want fast LR computation, we need fast inner product computation.

4.10 Mean embedding

Chapter 5

Discriminative training

5.1 Pairs

5.2 Triplet loss

First pubished in [2].

Good paper: [29]: explains original triplet loss, variants for hinge vs softplus loss and gives nice recipe for mining moderately hard triplets,

5.3 Multiclass classification

[8, 9]

5.4 Pseudolikelihood

Bibliography

[1] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *NIPS*, 2000.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832

[3] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech*, Brighton, UK, September 2009.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[5] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Interspeech*, 2011.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017. [Online]. Available: https://arxiv.org/abs/1706.08612

[7] D. Snyder, P. Ghahremani, and D. Povey, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *IEEE Workshop on Spoken Language Technology*, 2016.

[8] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech*, Stockholm, 2017.

[9] G. Gelly and J. L. Gauvain, "Spoken language identification using lstm-based angular proximity," in *Interspeech*, Stockholm, 2017.

[10] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Tech. Rep. CRIM-06/08-13, 2005.

[11] ——, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, keynote presentation.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *9th European Conference on Computer Vision*, Graz, Austria, 2006.

[13] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision*, 2007.

[14] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE Trans. PAMI*, vol. 34, no. 1, January 2012.

[15] L. Vilnis and A. McCallum, "Word representations via Gaussian embedding," in *ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6623

[16] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in PLDA," *IEEE Trans. ASLP*, vol. 22, no. 4, 2014.

[17] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *IEEE ICASSP*, 2013.

[18] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *Interspeech*, 2013.

[19] P. Kenny, T. Stafylakis, J. Alam, V. Gupta, and M. Kockmann, "Uncertainty modeling without subspace methods for text-dependent speaker recognition," in *Speaker Odyssey: The Speaker and Language Recognition Workshop*, Bilbao, 2016.

[20] Y. S. Chow and H. Teicher, *Probability theory: Independence, interchangeability, martingales*, 3rd ed., ser. Springer Texts in Statistics. New York: Springer, 1997.

[21] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[22] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[23] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[24] K. Siegrist, "Random: Probability, mathematical statistics, stochastic processes," 1997, see section 3.11 Vector Spaces of Random Variables. [Online]. Available: http://www.math.uah.edu/stat/expect/Spaces.html

[25] P. Billingsley, *Probability and Measure*, 3rd ed. John Wiley & Sons, 1995.

[26] P. Ouwehand, "Spaces of random variables," AIMS, lecture, November 2010. [Online]. Available: http://users.aims.ac.za/~pouw/Lectures/Lecture_Spaces_Random_Variables.pdf

[27] D. Bigoni, "Uncertainty quantification with applications to engineering problems," Ph.D. dissertation, Thechnical University of Denmark, 2015, see Appendix B: Probability theory and functional spaces.

[28] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, Florence, Italy, 2011.

[29] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017. [Online]. Available: http://arxiv.org/abs/1703.07737

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv*, 2013. [Online]. Available: http://arxiv.org/abs/1312.6114

[31] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational Bayes for non-conjugate inference," in *ICML*, 2014. [Online]. Available: www.jmlr.org/proceedings/papers/v32/titsias14.pdf