A Geometric Look at Double Descent Risk: Volumes, Singularities, and Distinguishabilities

Prasad Cheema  
The University of Sydney  
prasad.cheema@sydney.edu.au

Mahito Sugiyama  
The National Institute of Informatics  
mahito@nii.ac.jp

Abstract

The appearance of the double-descent risk phenomenon has received growing interest in the machine learning and statistics community, as it challenges well-understood notions behind the U-shaped train-test curves. Motivated through Rissanen’s minimum description length (MDL), Balasubramanian’s Occam’s Razor, and Amari’s information geometry, we investigate how the logarithm of the model volume: \( \log V \), works to extend intuition behind the AIC and BIC model selection criteria. We find that for the particular model classes of isotropic linear regression, statistical lattices, and the stochastic perceptron unit, the \( \log V \) term may be decomposed into a sum of distinct components. These components work to extend the idea of model complexity inherent in AIC and BIC, and are driven by new, albeit intuitive notions of (i) Model richness, and (ii) Model distinguishability. Our theoretical analysis assists in the understanding of how the double descent phenomenon may manifest, as well as why generalization error does not necessarily continue to grow with increasing model dimensionality.

1 Introduction

Model selection is a problem which underpins the field of machine learning. Key to its formulation is the notion of learning an appropriate predictor, \( h^* : \mathbb{R}^D \to \mathbb{R} \) from an underlying model class, \( \mathcal{H} \), based on \( N \) input training examples \( \{(x_i, y_i)\}_{i=1}^N \), with each \( (x_i, y_i) \in \mathbb{R}^D \times \mathbb{R} \). Typically, the predictor, \( h^* \), is chosen so as to minimize some risk functional; that is, \( h^* = \arg \min_{h \in \mathcal{H}} R(h) \) with \( R(h) = \mathbb{E}_{p(x,y)}[L(h(x), y)] \), where \( L : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) is the risk functional, and \( p(x, y) \) denotes the probability density function (pdf) over the data. Fundamentally, the aim of such an approach is to ensure that \( h^* \) provides good generalization capability, so that after training it minimizes the out-of-sample test error [15]. This is historically estimated via the Akaike information criterion (AIC) [5], the Bayesian information criterion (BIC) [32], or through cross validation [15]. AIC and BIC are derived based on asymptotic assumptions in the sample size \( N \), and work similarly. Moreover, both criteria suggest that out-of-sample error increases as \( \mathcal{O}(D^1) \), suggesting that an over-parameterized model should generalize poorly, which is an idea consistent with traditional empirical evidence, via the U-shaped train-test curves [15].

Recently, however, particular classes of highly parameterized models such as deep neural networks, and random forests have been shown to generalize extremely well, working in contrary to the implied

1Technically AIC has a \( 2D \) model complexity term, and BIC has a \( D \log N \) model complexity term. We will refer to the implied effects of both model complexities simply as the “\( \mathcal{O}(D) \) terms”.

Preprint. Under review.
\(\mathcal{O}(D)\) model complexity effects. In fact, strong empirical evidence has been presented by Belkin et al. [13], where it was shown that a double descent risk phenomenon may be observed for a variety of models which transition into highly parameterized regimes. This phenomenon is shown in Figure 1 with many additional experiments made clear in [26]. In an effort to explain such trends Belkin et al. [14] have tried to infer some similarities between ReLU networks and traditional kernel models, and Geiger et al. [19] have tried to connect the double descent cusp-like behaviour with diverging norms, through a neural tangent kernel framework. In addition, double descent risk has been explored in a variety of simpler (and shallow) model classes [21] [9] [17], with various risk asymptotics established [12] [23]. Lastly and rather interestingly, it has been found that there exist certain parallels between double descent risk, and the notion of the jamming transition which occurs in physical materials which undergo a phase transition [34] [20].

In this paper, we attempt to understand the double descent phenomenon from a geometrically driven perspective. This will be achieved via the notion of a model volume, and shown to have interpretations which stem from information theory (coding theory, and signal analysis), as well as Occam’s razor. We find that for a variety of simple statistical models, the model volume can decrease with increasing \(D\), which ultimately serves to overpower the magnitude of the lower order model complexity term: \(\mathcal{O}(D)\). This idea is also clarified in Figure 1.

### 1.1 Model Selection and Occam’s Razor

In the late 90s and early 2000s, extensions to the base AIC and BIC formulations were developed by Rissanen [31] and Balasubramanian [11], which include additional model-specific terms. From the perspective of coding theory, Rissanen developed a notion of stochastic model complexity, which builds upon Shannon’s information criteria used for lossless encoding [33]. Upon this notion, Rissanen formalized an extension of binary Shannon entropy to continuous function classes, via the discretization of the model manifold over approximately equivalent model classes. This approach establishes an intuition behind “model distinguishability”, which is also echoed by Balasubramanian. In particular, under Rissanen’s construction information is encoded in nats (as opposed to bits) and it is formally recognized as the Minimum Description Length (MDL). This is shown in Equation (1).

\[
-\log(p(x)) = -\log(\hat{L}) + \frac{D}{2} \log \left( \frac{N}{2\pi e} \right) + \log \int_{\Theta} \sqrt{\det (I(\theta))} d\theta + o(1),
\]

where \(x = \{x_i\}_{i=1}^N\) denotes a random vector of \(N\) data samples, \(\hat{L} \in \mathbb{R}\), is the likelihood function evaluated at its optimal parameter setting, with \(\Theta\) being the space of possible parameter settings, and \(I(\theta) \in \mathbb{R}^{D \times D}\) denotes the Fisher information matrix (FIM), which is traditionally used as a lower bound on the variance of unbiased estimators, and in Jeffrey’s prior [15]. In a parallel fashion, Balasubramanian approached the problem of model selection, albeit from a more Bayesian perspective, which curiously yielded the same equation as in Rissanen’s MDL (with the exception of one additional term, and ignoring constants) [10]. To achieve this he takes an alternate route based on specifying a Jeffrey’s prior over the underlying parameter space, which is noted to act as an appropriate measure for the density of distinguishable distributions [11]. Ultimately, in both Rissanen’s and Balasubramanian’s model selection criteria, there is a term which acts like the \(\mathcal{O}(D)\) model complexity used in AIC and BIC, and an additional term: \(\log \int_{\Theta} \sqrt{\det (I(\theta))} d\theta\), which they collectively referred to as a model distinguishability-like term. Moreover, since these methods are built upon the log-marginal: \(-\log(p(x))\), they also appeal to a Bayesian Occam’s razor-like principle. Ultimately, this paper aims to explore this additional term: \(\log \int_{\Theta} \sqrt{\det (I(\theta))} d\theta\), through a geometric lens. Moreover, it will be made clear that this term also describes the underlying log-model volume, which we denote by \(\log V\) [7]. Ultimately, it will be shown that, \(\log V\) can in fact decrease with increasing \(D\) for several simple model classes, thus serving to counter-act the behaviours of the \(\mathcal{O}(D)\) model complexity term. By implication, this suggests that certain model classes have the power to generalize well when transitioning into over-parameterized regimes.

### 1.2 Information Geometry

Information geometry concerns the application of differential geometry to statistical models. In particular, it considers a statistical manifold, \(\mathcal{M} = \{ p(x; \theta) \} \), over a \(\theta\) co-ordinate system. A Riemannian metric, \(\mathcal{G}\), can be placed on \(\mathcal{M}\), where \(\mathcal{G} : T_p(\mathcal{M}) \times T_p(\mathcal{M}) \rightarrow \mathbb{R}_{\geq 0}\) for each point
p ∈ M, with T_p(M) defined as the local tangent space at point p on the manifold. Principally, G is a generalization of the inner product on Euclidean spaces to Riemannian manifolds. In addition, Amari defines a dually coupled affine co-ordinate system on statistical manifolds. Dually coupled co-ordinates arise naturally from the dually flat property which is intrinsic to many information manifolds. These co-ordinates are known as the θ (e-flat) and η (m-flat) co-ordinates for the exponential family in particular, and are related through the Legendre transformation η = ∇ψ(θ) and θ = ∇φ(η) via two convex functions ψ, φ : R^D → R [8]. The θ and η co-ordinates for exponential models correspond to the natural and expectation parameters, respectively. Furthermore, the FIM defines a natural Riemannian metric tensor: \( G_{ij} = \mathbb{E}[\partial_i \log(p(x; \theta)) \partial_j \log(p(x; \theta))] = \mathcal{I}_{ij} \) [29, 8]. Thus the motivation for using information geometry is clear as Rissanen’s MDL, and Balasubramanian’s Occam Razor, depend on the FIM, which is, geometrically speaking, the metric tensor. Consequently, the term \( \log \int_0 \sqrt{\text{det}(\mathcal{I}(\theta))} d\theta \) has a clear definition in differential geometry as being the log-volume of the underlying information manifold; that is, the square root of the determinant of the Fisher information matrix is the manifold volume [8, 25]. Unfortunately, the FIM is singular \( \text{rank}(\mathcal{I}(\theta)) < D \) for many modern statistical models, implying that classical assumptions such as the asymptotic normality of the maximum likelihood estimator, does not hold, and that the BIC is not necessarily equal to the Bayes marginal likelihood [43]. The implication of this will be explored for the studied models.

## 2 Double Descent Risk and the Volume of Statistical Models

In the following subsections we will explore how the notion of model volume suggests that in some models, increasing D can result in a decreasing, or even saturating volume, which ultimately acts to counter-balance the poor generalization properties traditionally implied by the \( O(D) \) model complexity. This will be demonstrated for three popular models in machine learning: (1) Isotropic linear regression, (2) Statistical lattice models, and (3) Stochastic perceptron units.

### 2.1 Isotropic Linear Regression

We define isotropic linear regression as: \( y = X\beta + \epsilon \), for a given dataset \( \{x_i, y_i\}_{i=1}^N \), with each \( (x_i, y_i) \in \mathbb{R}^D \times \mathbb{R} \), where \( \epsilon \sim N(0, \sigma^2) \) and each \( x_{ij} \sim N(0, 1) \). Furthermore, we place a power constraint on \( \beta \in \mathbb{R}^D \) such that \( \mathbb{E}[\beta^T \beta] < P \) (for reasons which will become clear soon). It will be shown that for the regime \( D \leq N \) the model volume is well defined (FIM is non-singular), but in the regime of interest \( D > N \) a singular geometry arises leading to zero determinants, thus rendering the volume expression of Equation (1) not meaningful. This singularity will be an issue central to every model investigated, and in particular for linear regression we draw inspiration from innovations in signal processing literature in order to help circumvent it. We begin by revealing the general form of \( \log V \) for linear regression, which is elucidated in Theorem 1 where \( B^D(\sqrt{P}) \) refers to the volume of a \( D \)-ball of radius \( \sqrt{P} \).

**Theorem 1 (Regression Log Volume).** \( \log V = \frac{1}{2} \log \det(\mathbb{E}[X^T X]) + \log \left( B^D(\sqrt{P})/\sigma^D \right) \)

**Corollary 1.** For \( D \leq N \), \( \log V = D/2 \log N + \log \left( B^D(\sqrt{P})/\sigma^D \right). \)

The proofs of Theorem 1 and Corollary 1 are provided in Appendix A.1 and A.2 respectively. Theorem 1 makes clear that the \( \log V \) expression separates into two rather distinct components, (i) \( \frac{1}{2} \log \det(\mathbb{E}[X^T X]) \) which we claim possesses a notion of model richness, and (ii) \( \log \left( B^D(\sqrt{P})/\sigma^D \right) \), which is the volume of a \( D \)-ball scaled against noise, which defines a degree of model distinguishability. Considering richness, it is clear that it involves the expectation over the data. And whilst not explicit in Theorem 1, it also contains information regarding the model architecture, as this term arises as a by product of the FIM calculation, which necessitates the second derivative of the underlying model parameter space. Thus, this term serves to connect the underlying data distribution, against the (usually) implicitly made geometric assumptions of the model’s architecture. Considering now distinguishability, in the calculation of \( \log V \) this term manifests as the volume implied by the constraints underlying the parameter space. In the case of Theorem 1, \( \mathbb{E}[\beta^T \beta] \leq P \), implying \( \beta \) behaves according to a ball geometry, wherein it is known that \( \lim_{D \to \infty} B^D(R) = 0 \). This implies

\[ \text{The volume of a } D \text{-ball, or radius } R \text{ is: } B^D(R) = \frac{\pi^{D/2}}{\Gamma(D/2+1)} R^D, \text{ where } \Gamma(D) := (D - 1)! \]
that the addition of new features at increasingly high $D$ can work to counteract the $O(D)$ model complexity contribution, traditionally implied by AIC and BIC. Moreover, distinguishability is noted to behave as a function of the model noise. This carries with it a nice intuition as an increasingly noisy model will find it difficult to distinguish between similar datapoints. Thus, even though the traditional understanding of out-of-sample test error, as implied by AIC and BIC is driven by: (i) Training error, and (ii) Model complexity, we can see that model: (iii) Richness, and (iv) Distinguishability, also seem to play an important role. Ultimately, all the distinct model components (i) - (iv) seem to necessitate investigation, as each will respond in a unique way against the effects of, $N$, $D$, and the underlying model architecture.

From Corollary [1], it is shown that the $\frac{1}{2} \log \circ \det (E [X^\top X])$ richness term simplifies down to $D/2 \log(N)$ if $D \leq N$, which is able to add directly onto the already existing $O(D)$ term in AIC and BIC. However, extending this analysis to $D > N$, is difficult as rank $(X^\top X) = N$, even though $X^\top X \in \mathbb{R}^{D \times D}$, meaning $\dim \ker(X^\top X) = D - N > 0$. Thus we have an increasingly large null-space as $D$ increases, serving to exacerbate the singular geometry. By implication, $\det(X^\top X) = 0$, meaning that a naive volume calculation lacks meaning. To circumvent this issue, we draw inspiration from theories in communication theory. In particular, this line of research is often concerned with the same underlying system: $Y = X\beta + \varepsilon$, however, the interpretations of the two terms, $X$ and $\beta$, are appreciably different. In particular, $\beta$ is interpreted as an input signal, and $X$ is the signal mixing matrix. In other words, the known quantity in this literature is $\beta$, and the unknown quantity $X$ acts like a confounding factor, which is contrary with the traditional statistical goal of regression, whereby $X$ is known data, and $\beta$ are the unknown coefficients. Moreover, the signal $\beta$, is typically power constrained: $E[\beta^\top \beta] < P$, which in statistics often arises due to regularization, or through consideration of a prior distribution [40] [15]. For communication channels, a quantity of central interest is the channel capacity: $C$. It acts as an upper-bound on the rate of reliable information transfer for a communication channel, and it is mathematically equivalent to the supremum of the mutual information: $\sup_{p(\beta)} I(\beta; y)$, wherein we consider $\beta$ and $y$ as random variables [16].

For the problem of isotropic linear regression with real feature variables, $C$ follows Theorem 2 (see Appendix [A.4]).

**Theorem 2 (Channel Capacity).** The channel capacity for the system $Y = X\beta + \varepsilon$ with $X_{ij} \sim N(0, 1), \varepsilon \sim N(0, \sigma^2 I)$, and $E[\beta^\top \beta] \leq P$, where $\text{SNR} = P/\sigma^2$ refers to the signal-to-noise ratio, is given as,

$$ C = \frac{1}{2} E \left[ \log \circ \det \left( I_N + \frac{\text{SNR}}{D} X X^\top \right) \right]. $$

(2)

The importance of considering $C$, is that it bears an equivalency (asymptotic in $N$) to the $D/2 \log(N/2\pi e) + \log V$ term in Equation (1). Specifically, its relationship can be summarized in Equation (3), with further clarification provided in Theorems 6 and 7 in Appendix [A.3]. Thus, $C$ is also reflective of out-of-sample test error.

$$ \lim_{N \to \infty} \left[ C - \frac{D}{2} \log \frac{N}{2\pi e} + \log \int_{\Theta} \sqrt{\det(I(\theta))}d\theta \right] = 0. $$

(3)

However, not only does $C$ relate to out-of-sample test error through Equation (3), it: (i) Allows us to perform calculations for $D > N$, circumventing the singularity issue, and (ii) Implicitly encodes the required forms of richness and distinguishability. In fact, as a consequence of point (i), it is possible to derive upper and lower bounds (Theorem 9 in Appendix [A.5]), and further it can be made clear that the channel capacity saturates as $D \to \infty$, by studying such bounds (Corollary 2 in Appendix [A.6]). The presence of saturation (that is, an asymptotic behaviour) in such linear models has been recently observed in the work of Hastie et al. [23], using a bias-variance trade off. Moreover, Xu & Raginsky [44] have shown that generalization error relates intimately to the notion of input-output mutual information—the supremum of which is known to be $C$. It is interesting that such properties arise naturally in different fields, which in turn have different research aims.

We further study the interpretation of $C$. In Remark 1 (see Appendix [A.7]), $C$ is shown to be upper bounded by an expression containing: $\frac{1}{2} \log \circ \det (E [X^\top X])$, the regression richness term. However, here it is scaled with a factor of $\text{SNR}/D$, and added against the identity matrix: $I_D$. It is the presence of this identity matrix which acts as a regularizer against the null space of $X^\top X$. This allows one to investigate in the domain of $D > N$, by working with (a slightly modified) $\log V$ which avoids the singularity issues.

**Remark 1 (Regularized Richness Bound).** $C \leq \frac{1}{2} \log \circ \det \left( I_D + \frac{\text{SNR}}{D} E[X^\top X] \right)$. 

4
In addition to richness, we make clear that the notion of distinguishability arises implicitly in $C$, when considering a sphere packing argument. A common example of sphere packing as it arises in additive white Gaussian noise (AWGN) is provided for the reader in Appendix A.8. [40] Here however, we focus on showing how it manifests in $C$, considering the model $Y = X\beta + \varepsilon$, where each $\beta_i \sim \mathcal{N}(0, P/D)$, and where we consider $X$ initially fixed (that is, non-random). We base this calculation on two parts: (i) Considering what the model is trying to achieve without noise: $Y = X\beta$, and (ii) The impact of noise. Considering (i), we have established that $\beta$ is zero-mean Gaussian, which implies that $Y$ will have an ellipsoidal shape. That is, it will be scaled and rotated due to $X$. Moreover, considering that $X$ is ultimately a linear map, it is known that: $\text{Vol}(X\beta) = \sqrt{\det(X^TX)}\text{Vol}(\beta)$. And since, $X : \mathbb{R}^D \to \mathbb{R}^N$, where each $\beta_i \perp \beta_j$, we are ultimately dealing with a space wherein $\text{Vol}(\beta) = \mathbb{B}^N(\sqrt{NP/D})$. This results in $\text{Vol}(X\beta) = V_2 = \sqrt{\det(X^TX)}\cdot\mathbb{B}^N(\sqrt{NP/D})$, in which most vectors transformed under $X\beta$ will lie inside with high probability. Considering now point (ii), since $\text{dim}(\varepsilon) = N$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, the volume induced by the $N$-ball in regards to the presence of noise is, $V_1 = \mathbb{B}^N(\sqrt{N\sigma^2})$. Ultimately a log-volume ratio can be calculated as: $\log(V_2/V_1) = (1/2) \log \det(\text{SNR} \cdot X^TX)$, which derives $C$ for the case of $\text{SNR} \gg 1$. As a last point of connection, notice that the distinguishability term in Theorem 1 may be re-expressed as: $\log(\text{SNR}^{D/2}\mathbb{B}^D(1))$, which is simply the unit $D$-ball scaled against the $D$-th root of the signal-to-noise ratio. Therefore, not only does $C$ encode in it, a notion of richness, it also strongly encodes a notion of volume distinguishability, the intuition of which is made clear in Figure 2. From a communication theory perspective, each transmitted signal vector $X$ should be able to (i) Be decoded by the receiver (each transmitted codeword should map to distinct codes), and (ii) Maximize the total amount of uniquely decodeable codes that exists. In this case, the transmitted signals can be interpreted as $\beta' = X\beta$, and thus through training one hopes to find such a $\beta$, which would allow one to map the new, and unseen data, $X$, efficiently throughout the entire ellipsoid, thereby maximizing the generalization capability of the $\beta$ in question. Typically, if $\|\beta\|^2_2$ is large, then this can become increasingly difficult, as the outer ellipsoid will grow in volume relative to the noise spheres, providing yet another intuition for why we regularize models.

Empirical results for the out-of-sample test error in the case of isotropic linear regression model, and the impact of the magnitude of $\|\beta\|^2_2$ is made clear in Figure 3. To generate these results we used sklearn.linear_model.Ridge, as (i) The ridge regression hyper-parameter, $\alpha$, is used to emulate the effect of a power constraint, and (ii) We intend to keep the code as simple as possible to maximize reproducibility. For these experiments, we generated data according to $X_{ij} \sim \mathcal{N}(0, 1)$, and we assumed that there exists a true underlying generating process, of some dimension, which we choose wlog as 150. We keep increasing $D$ (the number of feature variables of the proposed model fit), up to and beyond 150, ranging from $D = 1$, up to $D = 2500$. In this way, we transition from tall $X$ matrices to long $X$ matrices, and are able to demonstrate a very clear transition from the classical, to the modern train-test risk regimes. In addition, 10-fold cross validation is performed in order to produce the train-test curves, and we opt for three values of the ridge hyper-parameter: $\alpha = (10^{-2}, 10^0, 10^2)$, chosen wlog, in order to demonstrate how the double descent behaviour can be slowly switched on-and-off, and considered $N = (300, 600, 900)$ datapoints. Moreover, the random seed is fixed between runs, and each $\beta_i$ is sampled from $\mathcal{N}(0, P = 0.25)$, with additive noise of $\varepsilon \sim \mathcal{N}(0, 1)$, so that we work with a fixed SNR between each case. Lastly, the experiments were run on a 64 bit Windows OS, with Intel(R) Core(TM) i7-7500U CPU @ 2.70 GHz, since these experiments are designed to be straight-forward, and reproducible by everyone.

Firstly, in Figure 3 we can see that the classical (U-shaped), and modern (double descent) regimes are visible. As noted by many researchers, the peak of the observed double descent phenomenon (the interpolation threshold), occurs at $N = D$ [13, 13, 23]. When considering the train-test curves based on, $\alpha = 10^{-2}$, it is evident that out-of-sample test error indeed peaks at $N = D$. However, for continually increasing $D$, all the test risk curves reach a point of saturation, regardless of the choice
of \( \alpha \). This was implied to occur due to the saturating effect of \( C \), when \( D \to \infty \). Moreover, it is clear that for each run, the train-test curves asymptote to the same error magnitude, which results due to the fixed SNR between each run. Such a behaviour is predicted to occur from Hastie et al. [24], wherein they claim that the asymptotic behaviour for risk curves (in linear problems), is highly dependent upon SNR. And indeed here, SNR arises naturally in \( \log V \), and when considering \( C \). In addition to this, the peaking behaviour of double descent risk coincides with the peaking locations of \( \| \beta \|_2^2 \), and it scales strongly dependent upon the choice of \( \alpha \). Such a behaviour was anticipated to occur via the considerations of model distinguishability, and sphere packing. In other words, the outer ellipsoid: 
\[
V_2 = \sqrt{\det(XX^\top)} \cdot B^N(\sqrt{NP/D}),
\]
has a volume dependent on \( P \), which is in turn reflective of the maximum magnitude allowed by \( \| \beta \|_2^2 \). Evidently if \( P \) explodes, then, \( V_2 \) will explode in volume, relative to the noise spheres: \( V_1 \), and the relative sizes of these volumes is indicative of out-of-sample test error. In other words, when only weak regularisation is applied (\( \alpha = 10^{-2} \)), it is not possible to properly constrain the magnitude of \( \| \beta \|_2^2 \), especially at the interpolation threshold (\( N = D \)), and thus \( V_2 \) like-wise explodes in magnitude. However as we progressively strengthen the effect of regularisation: \( \alpha = (10^0, 10^2) \), we observe a gradual reduction, and even a complete elimination of the double descent risk pathology. Recently, Nakkiran et al. [27] have also investigated the impact of ridge regularization on the double descent peaking phenomenon, wherein they observe that an optimally tuned regularizer can work to eliminate the presence of this peaking behaviour. However, the intuition behind this has not yet been elucidated from a geometric angle.

In addition, the double descent peak is observed to shift to the right with increasing \( N \). Information theoretically, increasing \( N \) proliferates the total number of possible encodings which may be able to explain the observed data. This is an interpretation which is consistent with Rissanen’s original derivation of MDL, in which he states: “the number of distinguishable models grows with the length of the data, which seems reasonable. In view of this we define the model complexity (as seen through the data)” [30]. Interestingly, this can imply that when a model’s total distinguishability is insufficient (weak regularization, and or insufficient noise), it is possible for the model to generalize well on one quantity of data, \( N_1 \), and then upon re-training on some new data such that \( N_2 > N_1 \), to then generalize poorly, due to the ability of the double-descent cusp to shift towards the right. Similar ideas have been uttered recently by Nakkiran et al., in that: “for a fixed architecture and training procedure, more data (can) actually hurt” [26].

Finally, we make clear that the concept of distinguishability (as it stems from \( \log V \)), has an intuition which is shared by many schools of thought on double descent risk, and generalization error. For example: (i) It bears strong intuitive similarities to the idea of a jamming transition, which refers to the tight packing of particles in physics, when a material transitions from fluid to solid. This idea is used by Geiger et al. [20, 19] to clarify several ideas on double descent risk. Moreover, (ii) Many statistical generalization theorems are based on the notion of “sphere coverings” over constrained function spaces, in which the sphere covering number, and sphere packing number, closely relate [41].

Figure 3: How double descent risk manifests for \( N = (300, 600, 900) \) and \( \alpha = (10^{-2}, 10^0, 10^2) \). Risk curves (upper row) and \( \| \beta \|_2^2 \) plots (bottom row).
2.2 Statistical Lattice Models

Thus far it has been established that in the case of isotropic linear regression, the double descent phenomenon and notions of generalization, possess strong geometric intuition. Here we show that this extends to a family of statistical lattice models which include Boltzmann machines (or Ising models) [1], log-linear models [6], and the matrix balancing problem [36]. We will encode the hierarchical structure of the domain of the probability distribution via a lattice, which will be shown to naturally lead into the η co-ordinates (m-flat) from information geometry (see 1.2), and analyze the learning of distributions over a lattice structured domain. In order to achieve this, we formalize lattice domains using partial orders.

Formally, a partially ordered set (poset) is a tuple, \((\mathcal{P}, \leq_{\mathcal{P}})\), where \(\mathcal{P}\) is a set of elements, and \(\leq_{\mathcal{P}}\) denotes an ordering structure, such that (1) \(\forall p \in \mathcal{P}\), \(p \leq_{\mathcal{P}} p\) (reflexivity), (2) If \(p \leq_{\mathcal{P}} q\) and \(q \leq_{\mathcal{P}} r\), then \(p \leq_{\mathcal{P}} r\) (antisymmetry), and (3) If \(p \leq_{\mathcal{P}} q\) and \(q \leq_{\mathcal{P}} r\), then \(p \leq_{\mathcal{P}} r\). Note that not every element may be directly comparable to every other element in the set (which would be known as a total ordering). In addition, a poset \((\mathcal{P}, \leq_{\mathcal{P}})\) is called a lattice if every pair of elements \(p, q \in \mathcal{P}\) has the least upper bound \(p \lor q\) and the greatest lower bound \(p \land q\) [18]. We assume that \(\mathcal{P}\) is finite. In working with posets it is common to consider the zeta function, \(\zeta : \mathcal{P} \times \mathcal{P} \rightarrow \{0, 1\}\) such that \(\zeta(p, q) = 1_{p \leq_{\mathcal{P}} q}\) \([23]\). The lattice structure always gives us the \(\theta\) and \(\eta\)-coordinates of information geometry: \(\log \mathcal{F}(p) = \sum_{q \in \mathcal{P}} \zeta(q, p) \theta_q = \sum_{\eta \leq_{\mathcal{P}}} \theta_\eta\) and \(\eta_\theta = \sum_{q \in \mathcal{P}} \zeta(p, q) \mathcal{F}(q) = \sum_{\eta \geq_{\mathcal{P}}} \mathcal{F}(q)\) \([36]\). For example, for Boltzmann machines with \(d\) binary variables, the lattice space \(\mathcal{P} = \{0, 1\}^n\), where \(p = (p_1, \ldots, p_n) \leq_{\mathcal{P}} q = (q_1, \ldots, q_n)\) if \(p_i \leq q_i\) for all \(i \in \{1, \ldots, n\}\). The size \(D = |\mathcal{P}| = 2^n\) in this case. The metric tensor for the information manifold of the proposed lattice structure is shown in Theorem 3 which was previously derived by Sugiyama et al. \([36]\). We assume that \(\mathcal{P} = \{1, \ldots, |\mathcal{P}|\}\) such that \(1\) corresponds to the least element without loss of generality.

**Theorem 3 (Lattice Metric Tensor \([36]\)).** \(G_{ij} = \sum_{p \in \mathcal{P}} \zeta(i, p) \zeta(j, p) \mathcal{F}(p) - \eta_i \eta_j = \eta_{i \lor j} - \eta_i \eta_j\).

In Theorem [3] we can replace \(\sum_{p \in \mathcal{P}} \zeta(i, p) \zeta(j, p) \mathcal{F}(p)\) with \(\eta_{i \lor j}\) as we assume that \(\mathcal{P}\) is a lattice and \(\eta_{i \lor j}\) always exists, which says that we only consider those poset structures in which the \(\eta\) co-ordinate values are shared (nested) between \(\eta_i\) and \(\eta_j\) for the off-diagonal terms in the metric tensor. In Theorem 3 it is clear that this metric tensor is expressed in terms of the \(\eta\) co-ordinates. The equivalent metric tensor in terms of the \(\theta\) co-ordinates is available, but it is much more difficult to work with (requires the Möbius function instead of the zeta function). Moreover, in this definition of the metric tensor, the first row and column are always zeros, resulting in again, a singular geometry. Luckily this happens since \(-\theta_1\) corresponds to the partition function, it is generally removed in practice. \([37]\) so that we effectively work with co-ordinates \(\eta' = (\eta_2, \ldots, \eta_D)\), resulting in \(G' > 0\). Based on this set-up, it is possible to derive the upper and lower bounds for \(\log \mathcal{V}\). For lattice models, the \(\eta\) co-ordinates lie compactly in the unit \(D\)-hypercube, that is \(\Omega = [0, 1]^D\), which makes evaluations much simpler \([35]\). In fact, it is possible to perform the re-parametrization: \(\delta = f(\eta)\), which allows \(\delta\) to be sampled from a Dirichlet distribution. This makes the \(\eta\) co-ordinate more intuitive to work with, and provides us with a tractable way to evaluate the volume integral via sampling. We provide details of this re-parametrization in Appendix B.1. The \(\log \mathcal{V}\) bounds which result are shown in Theorem 4 with the proof clarified in Appendix B.2.

**Theorem 4 (Lattice Log Volume Bounds).** \(\log \mathcal{V}\) is bound as in Equation (4), where \(\mathcal{G} = \mathcal{M}^\top \mathcal{M}\), \(\delta = f(\eta)\) is a re-parametrization, and \(\Gamma(D) = (D - 1)!\) is the standard Gamma function.

\[
\mathbb{E} \left[ \sum_{i=1}^{D} \log \left( \mathcal{M}_{ii}(\delta) \right) \right] + \log \left( \frac{1}{\Gamma(D)} \right) \leq \log \mathcal{V} \leq \log \left( \mathbb{E} \left[ \prod_{i=1}^{D} G_{ii}(\delta) \right] \right) + \log \left( \frac{1}{\Gamma(D)} \right) \tag{4}
\]

For poset models, we once again find that \(\log \mathcal{V}\) can be decomposed into two distinct terms (i) A model richness term which this time is driven by the diagonal elements found in the metric tensor, and (ii) The model distinguishability term, this time relating to the volume of the probability simplex. Intuitively: (i) The metric tensor is used to describe notions of angles, and lengths (divergences for dually flat manifolds), and so the product of its diagonal elements does encode a notion of complexity

\(^3\)Since elements from a poset are denoted by \(p\), we denote a probability measure as \(\mathbb{P}\) when referencing poset systems.
dependent on the choice of the model architecture. And once again, notice that this term manifests in the presence of the expectation operator, which summarizes how the model architecture relates to the data distribution. As for point (ii), we have a probability simplex geometry, of which the volume is:

$$\log V = \frac{1}{\sqrt{\det(I)}}.$$ 

Ultimately, the resulting behaviour of $\log V$ as $D \to \infty$ is clarified in Remark 2 (see Appendix B.3).

**Remark 2 (Limiting Lattice Volume).** $\lim_{D \to \infty} V = 0$

### 2.3 The Stochastic Perceptron Unit

Finally, we extend the current thesis to the classic stochastic perceptron unit. In particular, we consider a simple perceptron unit of the form: $y = f(w \cdot x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and $f(\cdot)$ is the sigmoid non-linearity. Moreover, we follow Amari’s parameterization from [4], which considers the input data distributed as, $x \sim \mathcal{N}(0, I)$, and defines, $w = \|w\|$, and $\xi \sim \mathcal{N}(0, 1)$, so that we can deal with the term: $w \cdot x = w\xi \sim \mathcal{N}(0, w^2)$. Based on this set-up, we obtain Theorem 5

**Theorem 5 (Log-Volume of Perceptron).**

$$\log V = \log \int_{\Omega_w} \sqrt{\det(I)} \ E_{\xi} [f'(w\xi)] \ E_{\xi} [f'(w\xi)]^{D-1} \ dw + \log \frac{\|w\|^D}{\sigma^D}$$

(5)

Theorem 5 makes clear that $\log V$ for the stochastic perceptron unit similarly decomposes into distinct distinguishability and richness components. The distinguishability term closely parallels that already derived for isotropic linear regression (Theorem 1) as it is driven by the volume of a $D$-ball, and scaled against noise. However, the richness term here strongly encodes the model architecture - as opposed to linear regression, as it contains the derivatives of the non-linearity function, in relation to the norm of the input weights, with the distribution over $x$. From observing this $\log V$ decomposition, it would appear beneficial for the network architecture to be modeled with non-linearities whose derivatives are well-constrained, so that as $D \to \infty$, we can better guarantee that $V \to 0$, implying good generalizability. As it turns out, the popular sigmoid, ReLU, and tanh activation functions do satisfy this property, as common to all we have: $|f'(\cdot)| \leq 1$. However, the perceptron unit has a strongly singular geometry, which does pose a marked difficulty when extending its geometry to deeper, and more complex network architectures. On this note, a possible solution may lie in the work of Sun & Nielsen [38]. In particular, they have tried to circumvent the pathologies of singular semi-Riemannian geometries via a clever use of lightlike manifold structures. Doing so, they observe that under certain conditions for deep neural network (DNN) architectures, a Balasubramanian Occam’s razor-like term may be developed as: $- \log p(x) \approx - \log(L) + \frac{D}{2} \log N - O(DN^2)$ (Equation (21) in [38]), wherein the higher order term is developed based on a combined understanding of the empirical FIM, and $\log V$. At large $D$, and for fixed $N$, this term dominates, which they theorize may imply the strong compression, and generalization properties for DNNs; a conclusion which runs parallel to the thesis of this paper.

### 3 Conclusion

Motivated through the works of Rissanen, and Balasubramanian, investigation of the $\log V$ term has shown that for a variety of statistical models, a geometrical understanding of out-of-sample test error and double descent risk involves considering: (i) Model training error, (ii) Model complexity, (iii) Model richness, and (iv) Model distinguishability. Such ideas carry an appealing intuition, and are firmly rooted in information geometry, coding theory, as well as in the principle of Occam’s razor. Future work in this area could entail an investigation into the practicality of such terms for the purpose of empirical model selection. Regardless however, the FIM does appear to be a key object in deepening one’s understanding of many statistical models, especially from a geometric perspective. However, its singular nature does necessitate a considered, and careful analysis, especially for models based upon deep, latent architectures.
References

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[2] Alan Agresti. *Categorical Data Analysis*. Wiley, 3 edition, 2012.

[3] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[4] Shun-ichi Amari. Information geometry of neural networks—an overview—. In *Mathematics of Neural Networks*, pages 15–23. Springer, 1997.

[5] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[6] Shun-ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[7] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016.

[8] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[9] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International Conference on Learning Representations*, 2020.

[10] Vijay Balasubramanian. A geometric formulation of occam’s razor for inference of parametric distributions. *arXiv:adap-org/9601001*, 1996.

[11] Vijay Balasubramanian. MDL, Bayesian inference, and the geometry of the space of probability distributions. *Advances in Minimum Description Length: Theory and Applications*, pages 81–98, 2005.

[12] Peter L. Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv:1906.11300*, 2019.

[13] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[14] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv:1802.01396*, 2018.

[15] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[16] Bertrand S Clarke and Andrew R Barron. Jeffreys’ prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

[17] Stéphane d’Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. *arXiv:2003.01054*, 2020.

[18] Brian A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2 edition, 2002.

[19] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d’Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[20] Mario Geiger, Stefano Spigler, Stéphane d’Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.
[21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. arXiv:1904.12191, 2019.

[22] Gerhard Gierz, Karl H. Hofmann, Klaus Keimel, Jimmie D. Lawson, Michael Mislove, and Dana S. Scott. Continuous Lattices and Domains. Cambridge University Press, 2003.

[23] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. arXiv:1903.08560, 2019.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[25] Harold Jeffreys. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 186(1007):453–461, 1946.

[26] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In International Conference on Learning Representations, 2020.

[27] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. arXiv preprint arXiv:2003.01897, 2020.

[28] Mark S Pinsker. Information and Information Stability of Random Variables and Processes. Holden-Day, 1964.

[29] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In Breakthroughs in Statistics, pages 235–247. Springer, 1992.

[30] Jorma J Rissanen. Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 42(1):40–47, 1996.

[31] Jorma J Rissanen. Stochastic complexity in learning. Journal of Computer and System Sciences, 55(1):89–95, 1997.

[32] Gideon Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 1978.

[33] Claude E Shannon. A mathematical theory of communication. Bell System Technical Journal, 27(3):379–423, 1948.

[34] Stefano Spigler, Mario Geiger, Stéphane d’Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. arXiv:1810.09665, 2018.

[35] Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Information decomposition on structured space. In 2016 IEEE International Symposium on Information Theory, pages 575–579, 2016.

[36] Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Tensor balancing on statistical manifold. In Proceedings of the 34th International Conference on Machine Learning, volume 70, pages 3270–3279, 2017.

[37] Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Legendre decomposition for tensors. In Advances in Neural Information Processing Systems, pages 8811–8821, 2018.

[38] Ke Sun and Frank Nielsen. Lightlike neuromanifolds, occam’s razor and deep learning. arXiv:1905.11027, 2019.

[39] Emre Telatar. Capacity of multi-antenna gaussian channels. European Transactions on Telecommunications, 10(6):585–595, 1999.

[40] David Tse and Pramod Viswanath. Fundamentals of Wireless Communication. Cambridge university press, 2005.
[41] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2013.

[42] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

[43] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25. Cambridge University Press, 2009.

[44] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.
Appendix

A Isotropic Linear Regression

A.1 Proof of Theorem 1

In order to prove Theorem 1, we first make note of Remark 3, which develops the FIM (metric tensor) for linear regression.

**Remark 3 (Linear Regression FIM).** The FIM (Riemannian metric tensor) for the proposed linear regression model is \( G(\beta) = E[X^\top X] / \sigma^2 \).

**Proof.** The proposed regression can be represented probabilistically as:

\[
p(y_i | x_i, \beta; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2}(y_i - x_i^\top\beta)^2 \right)
\]

\[
\Rightarrow \log p(y_i | x_i, \beta; \sigma^2) = -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y_i - x_i^\top\beta)^2
\]

\[
\Rightarrow \nabla_\beta^2 \log p(y_i | x_i, \beta; \sigma^2) = -\frac{x_i x_i^\top}{\sigma^2},
\]

where we have assumed variance \( \sigma^2 \) is known. Generalizing this to the data matrix \( X \), negating the above expression, and taking the expectation, we obtain the Fisher information matrix (Riemannian metric tensor) as required. \( \square \)

The proof of Theorem 1 thus completes as follows:

**Proof.** Since \( V = \int_B \sqrt{\det G(\beta)} \, d\beta \), where, \( G(\beta) = E[X^\top X] / \sigma^2 \).

\[
V = \int_B \sqrt{\det \left( \frac{E[X^\top X]}{\sigma^2} \right)} \, d\beta
\]

\[
= \sqrt{\det \left( \frac{E[X^\top X]}{\sigma^2} \right)} \int_B d\beta
\]

\[
= \sqrt{\det \left( \frac{E[X^\top X]}{\sigma^2} \right)} \int_{\beta_1^2 + \cdots + \beta_D^2} \cdot \cdot \cdot \int_B \, d\beta
\]

\[
= \sqrt{\det \left( \frac{E[X^\top X]}{\sigma^2} \right)} \cdot \mathbb{B}^D(\sqrt{P}),
\]

Thus, arriving at \( \log V = \frac{1}{2} \log \circ \det \left( \frac{E[X^\top X]}{\sigma^2} \right) + \log \frac{\mathbb{B}^D(\sqrt{P})}{\pi^{D/2}} \), where \( \mathbb{B}^D(\sqrt{P}) \) is the volume of a \( D \)-ball of radius \( \sqrt{P} \). \( \square \)

We make clear that the condition used to evaluate the integral in Theorem 1 is based on integrating over the domain: \( \beta^\top \beta \leq P \), where \( \beta \in \mathbb{R}^D \). However, the proposed power constraint in this paper is defined as \( E[\beta^\top \beta] \leq P \), so that ultimately \( \beta \) is considered to be a random vector. This extension is necessary later when defining channel capacity, wherein it will be assumed that each \( \beta_t \sim \mathcal{N}(0, P/D) \), implying that \( \beta^\top \beta \sim \frac{P}{D} \chi^2(D) \Rightarrow E[\beta^\top \beta] = P \). The choice of distribution on each \( \beta_t \) is driven by a maximum entropy argument, based on the first and second statistical moments. Specifically, it is known that the maximal entropy distribution with a specified mean (which we take as zero), and a specified covariance, \( E[\beta \beta^\top] \), is Gaussian, and further this covariance naturally relates to the required power constraint through: \( E[\beta^\top \beta] = \text{tr}(E[\beta \beta^\top]) \). Thus, based on a maximum entropy argument we opt for \( \beta_t \sim \mathcal{N}(0, P/D) \). Additionally, it is known that the norm of a \( D \)-dimensional random vector of sub-Gaussian components, distributed as \( \beta_t \sim \mathcal{N}(0, P/D) \) will take values close to \( \sqrt{P} \) (the radius of the \( D \)-sphere in question) with high probability. This can be seen in the concentration inequality: \( P \left( ||\beta||_2 - \sqrt{P} \geq t \right) \leq 2 \exp \left( -ct^2/K^4 \right) \), for all \( t > 0 \), where \( c \) is a constant, and \( K = \max_i ||\beta||_2 \), where \( || \cdot ||_2 \) is the sub-Gaussian norm (see Equation (3.3) in [42]). Thus the domain of the \( D \)-ball integral will contain most \( \beta \) vectors with a high probability.

12
A.2 Proof of Corollary 1

Proof. Consider \( X = [x_1, \ldots, x_D] \), such that the \( i \)-th \( X \) component: \( x_i \sim \mathcal{N}(0, I) \). Now, considering \( (X^\top X) \in \mathbb{R}^{D \times D} \), we arrive at a random matrix, with \( \chi^2 \)-distributed elements along its diagonals. In other words, the \( i \)-th diagonal element: \( (X^\top X)_{ii} \sim \chi^2(N) \), implying that \( \mathbb{E}(X^\top X)_{ii} = N \). Moreover, we obtain zeros in the off-diagonals due to the iid sampling conditions, meaning: \( \mathbb{E}(X^\top X)_{ij} = 0 \) since \( \mathbb{E}[x_i^\top x_j] = \sum_{k=1}^D \mathbb{E}[x_{ik}] \mathbb{E}[x_{jk}] = 0 \), so that due to independence we get: \( x_i \perp x_j \Rightarrow \mathbb{E}[x_i x_j] = \mathbb{E}[x_i] \mathbb{E}[x_j] \), and the summation term follows from the linearity of expectation. Thus, \( \sqrt{\det (\mathbb{E}[XX^\top])} = \sqrt{\det (N \cdot I_{D \times D})} = N^{D/2} \). Taking the logarithm we obtain \( D/2 \log N \) for the richness term, as required.

A.3 Channel Capacity Redundancy Theorems

Here, we make clear that channel capacity (which manifests as the supremum of a mutual information), is related to the MDL generalization term via the min-max KL risk. These ideas are well-known, and stated by Clarke & Barron in §2 and §5.1 of [16], and by Rissanen in §1 and §4 of [30].

Theorem 6 (Redundancy-Capacity Theorem [16]). The maximum channel capacity equals to the min-max KL divergence,

\[
\sup_{p(\beta)} KL(p(y|\beta)p(\beta)||p(y)) = \inf_{p \in \mathcal{P}} (KL(p(y|\beta)||p(y))) ,
\]

Theorem 7 (Redundancy-Generalization Theorem [16]). In the infinite limit of \( N \), the min-max KL risk approaches the MDL generalization estimator:

\[
\lim_{N \to \infty} \left[ \inf_{p \in \mathcal{P}} \sup_{\beta \in \mathcal{B}} (KL(p(y|\beta)||p(y))) - \frac{D}{2} \log \left( \frac{N}{2\pi} \right) - \log \left( \int_{\mathcal{B}} \sqrt{\det(I(\beta))}d\beta \right) \right] = 0
\]

Theorem 6 relates to Theorem 7 through the appearance of the \( \inf_{p \in \mathcal{P}} \sup_{\beta \in \mathcal{B}} (KL(p(y|\beta)||p(y))) \) term (the min-max KL risk). Specifically, these theorems make clear of an asymptotic relationship between the channel capacity, and the out-of-sample test error.

A.4 Proof of Theorem 2

In this subsection we provide a proof for \( C \) as it appears in Theorem 2. We note that there exists a proof for a similar looking system, involving, \( X_{ij}, \xi_i \sim \mathcal{CN}(0, 1) \) being circularly complex in Telatar’s seminal paper [39]. However, since properties over complex random matrix spaces do not necessarily lend themselves to real random matrices, we must work to provide a quick alternate proof for the case of \( X_{ij} \sim \mathcal{N}(0, 1) \), and with more general noise term, \( \xi \sim \mathcal{N}(0, \sigma^2) \) - which are the required assumptions for this particular paper, and for the problem of isotropic linear regression in general. To prove Theorem 2, we call upon Theorem 9.2.1 in Pinsker [28], which relates the mutual information between real Gaussian random vectors in a convenient form via the logarithm of the determinant. This is expressed as Theorem 8 below.

Theorem 8 (Pinsker’s Mutual Information [28]). Let \( \xi = (\xi_1, \ldots, \xi_n) \), \( \eta = (\eta_1, \ldots, \eta_m) = (\xi_{n+1}, \ldots, \xi_{n+m}) \), and \( (\xi, \eta) = (\xi_1, \ldots, \xi_{n+m}) \) be mean zero Gaussian random vectors taking values in some corresponding measurable Cartesian product spaces, \( X = X_1 \times \ldots \times X_n \), \( Y = Y_1 \times \ldots \times Y_m = X_{n+1} \times \ldots \times X_{n+m} \), and where \( Z = X \times Y \). Then,

\[
I(\xi; \eta) = \frac{1}{2} \log \left( \frac{\det(A_\xi) \cdot \det(A_\eta)}{\det(A_{\xi,\eta})} \right),
\]

where \( A_\xi = \mathbb{E}[\xi \xi^\top] \), and \( A_\eta = \mathbb{E}[\eta \eta^\top] \).

We now move onto proving Theorem 2, and begin by relating the notation used by Pinsker, to that used within this paper.
We once again emphasise that we initially consider a fixed (that is, non random) value of $X = X$ to perform calculations. Through maximum entropy arguments, we consider $E[\beta \beta^T] = \frac{D}{2} I_D$, which is based on the power constraint of $E[\beta^T \beta] \leq P$, coupled with the fact that each $\beta_i \sim N(0, P/D \cdot I)$. For details on the choice of this prior in regards to maximum entropy, see the end of Appendix [A.1] and in addition see § 4.1 of Telatar as to why this is a natural choice for: bounds in total.

In this section we derive upper and lower bounds for the channel capacity, $C$. Moreover, the expression thus far has been calculated for a fixed $\beta \beta^T$. We ultimately desire a form for $I$ accurately we have an expression for $A$, and lower bounds are required separately, for the cases of $D \leq N$, and $D > N$. This results in four bounds in total.

**Theorem 9 (Channel Capacity Bounds).** The channel capacity $C : \mathbb{R} \to \mathbb{R}$ is the digamma function, is bounded as follows:

$$\frac{D}{2} \log \left( \frac{2SNR}{D} \right) + \frac{1}{2} \sum_{i=1}^{D} \Psi \left( \frac{N-i+1}{2} \right) \leq C \leq \left\{ \begin{array}{ll} \frac{D}{2} \log \left( \frac{N}{D} SNR + 1 \right) & \text{for } D \leq N, \\ \frac{N}{2} \log (SNR + 1) & \text{for } D > N. \end{array} \right.$$
Proof. Lower Bound \((D \leq N)\):

\[ A, B \succ 0 \Rightarrow \det(A + B)^{1/n} \geq \det(A)^{1/n} + \det B^{1/n} \]  
(Minkowski’s inequality). We take \(n = 1\), fix an \(X = X\), and proceed as follows:

\[
\det \left( I_D + \frac{\text{SNR}}{D} X^\top X \right) \geq 1 + \det \left( \frac{\text{SNR}}{D} X^\top X \right) 
\]

\[
= 1 + \left( \frac{\text{SNR}}{D} \right)^D \det (X^\top X).
\]

Consider that for any \(a, b\) such that \(a > b\) we have \(\log_m(a) > \log_m(b)\), for \(m > 1\). Thus,

\[
\log \circ \det \left( I_D + \frac{\text{SNR}}{D} X^\top X \right) \geq \log \left( 1 + \det \left( \frac{\text{SNR}}{D} X^\top X \right) \right) 
\]

\[
> \log \circ \det \left( \frac{\text{SNR}}{D} X^\top X \right) 
\]

\[
= D \log \left( \frac{\text{SNR}}{D} \right) + \log \circ \det (X^\top X). 
\]

\[
\Rightarrow \frac{1}{2} \mathbb{E} \left[ \log \circ \det \left( I_D + \frac{\text{SNR}}{D} X^\top X \right) \right] \geq \frac{D}{2} \log \left( \frac{\text{SNR}}{D} \right) + \frac{1}{2} \mathbb{E}[\log \circ \det(X^\top X)].
\]

Now, since each \(X_{ij} \sim N(0, 1)\), it follows that \(X^\top X \sim \mathcal{W}(N, \Lambda)\); that is, \(X^\top X\) is Wishart distributed, where \(N\) is the degrees of freedom, and \(\Lambda\) is the Wishart scale matrix, which in this case is the identity matrix. For the Wishart distribution, there is a known expansion for the expected log-determinant as follows:

\[
\mathbb{E}[\log \circ \det(X^\top X)] = \sum_{i=1}^{D} \Psi \left( \frac{N - i + 1}{2} \right) + D \log 2 
\]

with \(\Psi(\cdot)\) being the standard digamma function \([15]\). Thus we arrive at:

\[
\mathcal{C} \geq \frac{D}{2} \log \left( \frac{2\text{SNR}}{D} \right) + \frac{1}{2} \sum_{i=1}^{D} \Psi \left( \frac{N - i + 1}{2} \right),
\]

where the lower bound for \(D > N\) follows similarly.

Upper Bound \((D \leq N)\):

From Jensen’s inequality,

\[
\mathcal{C} \leq \frac{1}{2} \log \circ \det \left( I_D + \frac{\text{SNR}}{D} \mathbb{E}[X^\top X] \right) 
\]

\[
= \frac{1}{2} \log \circ \det \left( \text{diag}_D \left( 1 + \frac{\text{SNR}}{D} \mathbb{E}[\chi^2(N)] \right) \right) 
\]

\[
= \frac{1}{2} \log \prod_{i=1}^{D} \left( 1 + \frac{\text{SNR} \cdot N}{D} \right) 
\]

\[
= \frac{D}{2} \log \left( \frac{N}{D} \text{SNR} + 1 \right). 
\]

The proof for the upper bound in \(D > N\) follows similarly, where we would consider a \(\chi^2(D)\) term instead.

We can see how the bound behaves in Figure 5. In particular, it seems to suggest that the lower bound is much tighter than the upper bound for high SNR. However, it exhibits a small “dip” at the transition point of \(D \to N\). This is primarily due to the \(\frac{D}{2} \log (2\text{SNR}/D)\) term in the lower bound, which for smaller SNR becomes negative if \(\text{SNR} < D/2\). Figure 5 naturally leads one to question if something may be said about the limiting nature of the upper and lower bounds as \(D \to \infty\). We thus establish Corollary 2.
Corollary 2 - Convergence of Channel Capacity Limit

In order to establish convergence we consider the following mild conditions: (1) A high SNR ($\text{SNR} \gg 1$), and (2) $\Psi(x) \approx \log(x) - \frac{1}{2x}$, which is a common approximation to the digamma function, for large $x$.

**Corollary 2 (Channel Capacity Convergence).** For $D > N$, and under mild conditions we have,

$$
\lim_{D \to \infty} \mathcal{C} = \lim_{D \to \infty} \mathcal{C} = \frac{N}{2} \log(\text{SNR}),
$$

where $\mathcal{C}$ and $\mathcal{C}$ refer to the upper and lower capacity bounds which are established in Theorem 9.

**Proof.**

$$
\mathcal{C} \approx \frac{N}{2} \log \left( \frac{2\text{SNR}}{D} \right) + \frac{1}{2} \sum_{i=1}^{N} \left[ \log \left( \frac{D - i + 1}{2} \right) - \frac{1}{2(D - i + 1)} \right]
$$

$$
= \frac{1}{2} \sum_{i=1}^{N} \left[ \log \left( \frac{2\text{SNR}(D - i + 1)}{2D} \right) - \frac{1}{2(D - i + 1)} \right]
$$

$$
= \frac{1}{2} \sum_{i=1}^{N} \left[ \log \left( \frac{2\text{SNR}(1 - i/D + 1/D)}{2} \right) - \frac{1}{2(D - i + 1)} \right]
$$

$$
\Rightarrow \lim_{D \to \infty} \mathcal{C} = \frac{1}{2} \sum_{i=1}^{N} \log(\text{SNR})
$$

$$
= \frac{N}{2} \log(\text{SNR})
$$

$$
= \lim_{D \to \infty} \mathcal{C}, \text{ for } \text{SNR} \gg 1
$$

\qed
A.7 Proof of Remark 1

Proof. This observation is a direct consequence of Sylvester’s determinant theorem, also known as the Weinstein–Aronszajn identity, where we observe
\[
\log \circ \det \left( I_D + \frac{\text{SNR}}{D} X X^\top \right) = \log \circ \det \left( I_N + \frac{\text{SNR}}{D} E \left[ X^\top X \right] \right)
\]
Consequently we can write:
\[
C = \frac{1}{2} E \left[ \log \circ \det \left( I_D + \frac{\text{SNR}}{D} X X^\top \right) \right]
\]
\[
= \frac{1}{2} E \left[ \log \circ \det \left( I_N + \frac{\text{SNR}}{D} E \left[ X^\top X \right] \right) \right]
\]
\[
\leq \frac{1}{2} \log \circ \det \left( I_N + \frac{\text{SNR}}{D} E \left[ X^\top X \right] \right)
\]

A.8 Sphere Packing in AWGN

For the additive white Gaussian noise (AWGN) case, we consider the system:
\[
y = x + \varepsilon, \text{ where } x, y \in \mathbb{R}^N \text{ and each } \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2).
\]
Under this system, each instantiation of \( x = x \) is approximately surrounded by a noise sphere of radius \( \sqrt{E[\varepsilon^\top \varepsilon]} = \sqrt{N\sigma^2} \). Moreover, we can define an approximate radius using \( y \) as:
\[
\sqrt{E[y^\top y]} = \sqrt{E[x^\top x] + E[\varepsilon^\top \varepsilon]} = \sqrt{NP + N\sigma^2},
\]
which describes the set of all possible codewords decodable when transmitting a signal \( x \). This is made geometrically clear in Figure 6. In this paper we extend this simple example to the ellipsoid case, which has an additional transformation factor of: \( \sqrt{\det XX^\top} \).

Figure 6: Channel capacity is equivalent to sphere packing on a hypersphere for the case of AWGN.

B Statistical Lattice Models

B.1 Reparameterizing the Dual Geometry of Lattices

Lattices are useful structures, as they allow one to efficiently encode information hierarchically. A geometry over lattice-type structures based on modelling higher order feature interactions via log probabilities has been derived in the work of Sugiyama et al. [35]. The well-known log-linear model for binary variables in question is formulated as,
\[
\log p(x) = \sum_i \theta_{(i)} x_i + \sum_{i<j} \theta_{(i,j)} x_i x_j + \sum_{i<j<k} \theta_{(i,j,k)} x_i x_j x_k + \cdots + \theta_{(1\ldots,n)} x_1 \ldots x_n - \psi,
\]
where \( x \in \{0, 1\}^n \), each \( \theta \in \mathbb{R} \) denotes the connection strength of a particular higher order interaction, each \( x_i \in \{0, 1\} \) denotes a binary valued variable which activates a particular connection strength, \( \psi \in \mathbb{R} \) denotes the normalization constant for the probability model [6, 2]. Under this structure, \( \mathbb{P}(x) \) is a member of the exponential family of distributions. If we define a particular instance of the partial ordering as \( x = (x_1 \ldots x_n) \leq y = (y_1 \ldots y_n) \), where \( x_i \leq y_i \) for all \( i \in \{1 \ldots n\} \), and denote by \( \Sigma(x) \) as the set of indices of “1” in \( x \), then when can instantiate a lattice, and can condense the representation of the above log-linear model as:
\[
\log p(x) = \sum_s \delta(s, x) \theta_{\Sigma(s)} = \sum_{s \subseteq x} \theta_{\Sigma(s)}, \text{ where } \psi = -\theta_0.
\]
Figure 7: An example of a lattice structure for the binary domain $\{0, 1\}^3$. Each arrow denotes the partial order between elements in the lattice.

Hence the lattice is a natural representation of this hierarchical structure over the sample space of $\{0, 1\}^n$. Sugiyama et al. [35] studied geometric structure of statistical lattice models and showed that distributions over not only $\{0, 1\}^n$ but any lattices belong to the exponential family. Note that posets are originally used in [35], which is a more general structure than lattices. Although we treat only lattices in this paper, most interesting statistical models (such as Boltzmann machines) are lattices. We thus proceed in this direction as lattice structures entail a simple co-ordinate representation of the metric tensor as we have described in Theorem 3. An example of a lattice structure for $\{0, 1\}^3$ is illustrated in Figure 7.

As Amari notes [8], the exponential family of distributions induces a statistical manifold which possesses an interesting dualistic structure. That is, two co-ordinate systems can be dually connected and allow one to generalize notions such as the Pythagoras theorem in Euclidean manifolds, to more general statistical manifolds. In particular, for the exponential family the first of these co-ordinate systems is given by $\theta = (\theta_1, ..., \theta_D)$ (as defined in the specified log-linear model of this subsection), and the second is given by $\eta = (\eta_1, ..., \eta_D)$. Note that in the case of a binary log-linear model, we have $D = 2^n$ and

$$
\eta_{(i)} = E[x_i] = \Pr(x_i = 1) \\
\eta_{(i,j)} = E[x_i x_j] = \Pr(x_i = 1, x_j = 1) \\
\eta_{(1,...,n)} = E[x_1...x_n] = \Pr(x_1 = 1, ..., x_n = 1),
$$

and $(\theta, \eta)$ are explicitly dually connected via the Legendre transformation [8, 35]. Note that since the $\eta$ co-ordinate system is defined probabilistically, and is thus constrained to be in $[0, 1]^D$, which is convenient, and simplifies many calculations. Moreover, note that the $\eta$ co-ordinate system is built hierarchically, in the sense that it adheres to a partial ordering similar to the following example structure:

$$
\eta_1 \geq \eta_2, \eta_3 \geq \eta_4, \eta_3 \geq \eta_5, \ldots
$$

This ordering is model specific and always uniquely determined from the lattice structure, and it is thus difficult to perform integrations over such arbitrary orderings. However, owing to the probabilistic nature of the $\eta$ co-ordinate system, it is possible to impose the following recursive re-parameterisation:

$$
\eta_1 = \delta_1 \\
\eta_2 = \eta_1 + \delta_2 \\
\eta_3 = \eta_1 + \delta_3 \\
\vdots \\
\eta_D = \eta_{D-1} + \delta_D
$$

where each $\delta_i \in [0, 1]$ for $1 \leq i \leq D$, and $\sum_{i=1}^{D} \delta_i = 1$. Geometrically speaking, $\delta = \{\delta_i\}_{i=1}^D$ represents points in a $D$-simplex. We can then proceed to formally encode the lattice ordering constraints via an additional zeta matrix, resulting in $\eta = Z\theta$, where each $Z_{ij} \in \{0, 1\}$ is the value
of zeta function \( \zeta(q_i, q_j) = 1_{q_i \leq q_j} \) for the corresponding elements \( q_i \) and \( q_j \) in the lattice. In other words, points from the \( D \) simplex, \( \delta \), can be transformed into \( \eta \) co-ordinates for the poset manifold through a linear mapping. In order to re-express the volume integral via the \( \delta \) co-ordinates, it is necessary to calculate the determinant of the Jacobian transformation matrix between the co-ordinate systems. However this is trivially one, since \( Z \) is by construction upper triangular, resulting in

\[
\det \left( \frac{\partial \eta}{\partial \delta} \right) = 1.
\]

Therefore it is possible to calculate the log-volume integral as,

\[
\log \left( \int_{\Delta_D} \sqrt{\det (G(\delta))} \cdot \det \left( \frac{\partial \eta}{\partial \delta} \right)^2 d\delta \right) = \log \left( \int_{\Delta_D} \sqrt{\det (G(\delta))} d\delta \right),
\]

where the coordinate transformation was performed using the square of the determinant, as the metric tensor is rank \((0,2)\) - that is, it is a doubly covariant object, and we define the \( D \)-simplex as \( \triangle_D \). In this form it is natural to re-express the volume integral via the expectation operator, where the expectation is taken with respect to a Dirichlet distribution over \( \delta \), as the Dirichlet distribution represents a pdf over the probability simplex. In other words we consider,

\[
\log \left( \int_{\Delta_D} \sqrt{\det (G(\delta))} d\delta \right) = \log \left( \int_{\Delta_D} \sqrt{\det (G(\delta))} \cdot \frac{w(\delta)}{w(\delta)} d\delta \right)
\]

\[
= \log \left( \mathbb{E} \left[ \frac{\sqrt{\det (G(\delta))}}{w(\delta)} \right] \right),
\]

where \( w(\delta) = \prod_{i=1}^{D} \frac{\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{D} \alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1} := \text{Dir}(\delta; \alpha) \), with \( \alpha = (\alpha_1, \ldots, \alpha_D) \), and \( \Gamma : \mathbb{R} \rightarrow \mathbb{R} \) being the standard Gamma function. Here, the choice of \( \alpha \) controls the manner in which sampling is performed over \( \Delta_D \). We opt for a uniform exploration over the \( D \)-simplex, which equates to requiring that \( \alpha_d = 1 \) for all \( d \in D \). Doing so means that we get, \( w(\delta) = \prod_{i=1}^{D} \frac{1}{\Gamma(\sum_{i=1}^{D} 1)} = 1^{(D)}, \) where \( \Gamma(D) = (D - 1)! \). Ultimately, Equation (12) becomes,

\[
\log \left( \mathbb{E} \left[ \frac{\sqrt{\det (G(\delta))}}{w(\delta)} \right] \right) = \log \left( \mathbb{E} \left[ \sqrt{\det (G(\delta))} \right] \right) - \log \Gamma(D),
\]

implying that bounding the volume can be equivalently achieved by bounding the behaviour of \( \log \left( \mathbb{E} \left[ \sqrt{\det (G(\delta))} \right] \right) \), and then appending the \( \log \Gamma(D) \) term. The upper and lower bounds on this volume integral, are shown in Appendix B.2.
B.2 Proof of Theorem 4

In this subsection we proceed to find lower and upper bounds for the \( \log V \) in the case of the prescribed lattice geometry, by exploiting the re-parameterization of the \( \eta \) co-ordinate system.

**Proof. Volume Upper Bound:**

From Hadamard’s inequality: \( |\det(A)| \leq \prod_{i=1}^{D} A_{ii}, \) for \( A \in \mathbb{R}^{D \times D} \). Thus:

\[
|\det(G(\delta))| = \det(G(\delta)) \quad (G(\delta) \succ 0) \\
\leq \prod_{i=1}^{D} G_{ii}(\delta),
\]

\[
\Rightarrow \quad \sqrt{\det(G(\delta))} \leq \prod_{i=1}^{D} G_{ii}(\delta)
\]

\[
\iff \quad \mathbb{E}\left[\sqrt{\det(G(\delta))}\right] \leq \mathbb{E}\left[\prod_{i=1}^{D} G_{ii}(\delta)\right]
\]

\[
\iff \quad \log \left(\mathbb{E}\left[\sqrt{\det(G(\delta))}\right]\right) \leq \log \left(\prod_{i=1}^{D} G_{ii}(\delta)\right)
\]

\[
\iff \quad \log \left(\mathbb{E}\left[\sqrt{\det(G(\delta))}\right]\right) - \log \Gamma(D) \leq \log \left(\prod_{i=1}^{D} G_{ii}(\delta)\right) - \log \Gamma(D)
\]

\[
\iff \quad \log V \leq \log \left(\frac{\mathbb{E}\left[\prod_{i=1}^{D} G_{ii}(\delta)\right]}{\Gamma(D)}\right),
\]

**Volume Lower Bound:**

\[
\log \left(\mathbb{E}\left[\sqrt{\det(G(\delta))}\right]\right) \geq \mathbb{E}\left[\log \left(\sqrt{\det(G(\delta))}\right)\right] \quad (14)
\]

\[
= \frac{1}{2} \mathbb{E}\left[\log \circ \det(G(\delta))\right], \quad (15)
\]

where Inequality (14) is Jensen’s inequality. Moreover, \( G \succ 0 \Rightarrow \exists M \) s.t. \( G = MM^\top \), where \( M \) is a triangular matrix (Cholesky decomposition). Thus,

\[
\det(G) = \det(MM^\top) \\
= \det(M) \cdot \det(M^\top) \\
= \det(M) \cdot \det(M) \\
= \det(M)^2 \\
= \left(\prod_{i=1}^{D} M_{ii}\right)^2, \quad (16)
\]
\[ \Rightarrow \quad \frac{1}{2} \mathbb{E} [\log \circ \det (\mathcal{G})] = \frac{1}{2} \mathbb{E} \left[ \log \left( \prod_{i=1}^{D} \mathcal{M}_{ii}(\delta) \right)^2 \right] = \mathbb{E} \left[ \sum_{i=1}^{D} \log (\mathcal{M}_{ii}(\delta)) \right], \]

\[ \Rightarrow \quad \log \left( \mathbb{E} \left[ \sqrt{\det (\mathcal{G}^2)} \right] \right) \geq \mathbb{E} \left[ \sum_{i=1}^{D} \log (\mathcal{M}_{ii}(\delta)) \right] \]

\[ \Leftrightarrow \quad \log \left( \mathbb{E} \left[ \sqrt{\det (\mathcal{G}^2)} \right] \right) - \log \Gamma(D) \geq \mathbb{E} \left[ \sum_{i=1}^{D} \log (\mathcal{M}_{ii}(\delta)) \right] - \log \Gamma(D) \]

\[ \Leftrightarrow \quad \log V \geq \mathbb{E} \left[ \sum_{i=1}^{D} \log (\mathcal{M}_{ii}(\delta)) \right] + \log \left( \frac{1}{\Gamma(D)} \right) \]

This decomposition provides insight into how \( \log V \) operates, because of its split into the distinct additive components of richness, and distinguishability. Such ideas were shown similarly for Theorem 1 in the case of isotropic linear regression, and in Theorem 5 for the stochastic perceptron unit.

**B.3 Proof of Remark 2**

Here we show that as \( D \to \infty \), \( V \to 0 \) which implies that \( \log V \to -\infty \). This is an important indicator in the increase of generalization performance, as sufficiently large \( D \) can therefore overpower the \( O(D) \) model complexity term, present in traditional AIC and BIC.

**Proof.** As this represents a volume integral a trivial lower bound is zero. It follows that

\[ 0 \leq V \leq \mathbb{E} \left[ \frac{\sqrt{\prod_{i=1}^{D} \mathcal{G}_{ii}(\delta)}}{\Gamma(D)} \right] \]

\[ \Rightarrow \quad \lim_{D \to \infty} 0 \leq \lim_{D \to \infty} V \leq \lim_{D \to \infty} \mathbb{E} \left[ \frac{\sqrt{\prod_{i=1}^{D} \mathcal{G}_{ii}(\delta)}}{(D-1)!} \right]. \]

Since each \( \mathcal{G}_{ii}(\delta) \in [0, 1] \), the factorial in the denominator strongly dominates, so that:

\[ \lim_{D \to \infty} \mathbb{E} \left[ \frac{\sqrt{\prod_{i=1}^{D} \mathcal{G}_{ii}(\delta)}}{(D-1)!} \right] = 0. \]

Thus from via an application of squeeze theorem we see that,

\[ 0 \leq \lim_{D \to \infty} V \leq 0, \quad \Rightarrow \quad \lim_{D \to \infty} V = 0. \]

\[ \square \]
C Stochastic Perceptron Unit

C.1 Stochastic Perceptron Metric Tensor

We consider the following stochastic perceptron unit,

\[ p(x, y; w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (y - f(w \cdot x))^2 \right) p(x), \]

and note that its metric tensor has been derived previously by Amari for the case of \( \sigma^2 = 1 \). This is made clear in Theorem 10. Moreover, we follow Amari’s notation from [4], and consider the input data distributed as, \( x \sim \mathcal{N}(0, I) \), and write: \( w = \|w\| \), and \( \xi \sim \mathcal{N}(0, 1) \), so that we can go on to define: \( w \cdot x = w \xi \sim \mathcal{N}(0, w^2) \).

**Theorem 10 (Perceptron Metric Tensor [4]).** The metric tensor for the stochastic perceptron model is given as:

\[ \mathcal{G}(w) = c_1(w) I + (c_2(w) - c_1(w)) e_w e_w^T, \]

such that

\[ c_1(w) = \frac{1}{\sqrt{2\pi}} \int f(w\xi) (1 - f(w\xi)) \exp \left( -\frac{1}{2\sigma^2} \xi^2 \right) d\xi, \]

\[ c_2(w) = \frac{1}{\sqrt{2\pi}} \int f(w\xi) (1 - f(w\xi)) \xi^2 \exp \left( -\frac{1}{2\sigma^2} \xi^2 \right) d\xi \]

where \( w = \|w\|_2 \), \( e_w = \frac{w}{\|w\|} \), and \( f : \mathbb{R} \to \mathbb{R} \) is the sigmoid activation function.

Its proof can be inferred from another one of Amari’s works [5], where instead it is performed with respect to the tanh activation function. Based on this, we note that Theorem 10 can be written into an expectation form, with the inclusion of a generic noise term \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \). This is shown in Proposition 1.

**Proposition 1 (Metric Tensor Expectation).** The metric tensor can be condensed into expectation form as:

\[ \mathcal{G}(w) = \frac{1}{\sigma^2} \left( \mathbb{E}_{\xi} \left[ f'(w\xi) \right] + e_w e_w^T \mathbb{E}_{\xi} \left[ f'(w\xi)(\xi^2 - 1) \right] \right). \]

**Proof.** Take the expectation with respect to the Gaussian measure on \( \xi \).

The derivative, and noise terms arise from realizing that, if:

\[ \log p(y, x; w, \sigma^2) = -\frac{1}{2\sigma^2} (y - f(w \cdot x))^2 + \log p(x) + C. \]

Then \( \frac{\partial p}{\partial w_i} = -\frac{1}{\sigma^2}(y - f(w \cdot x)) \cdot f'(w \cdot x) \cdot x_i. \)

\( \Rightarrow \mathcal{G}_{ij}(w) = \mathbb{E} \left[ \frac{\partial}{\partial w_i} \log p \cdot \frac{\partial}{\partial w_j} \log p \right] \]

\[ = \frac{1}{\sigma^2} \mathbb{E} \left[ f'(w \cdot x)^2 x_i x_j \right], \]

where we can notice: \( y - f(w \cdot x) = \varepsilon \), which implies \( \mathbb{E}[\varepsilon^2] = \sigma^2 \).

Changing the \( x_i, x_j \) term into the required \( \xi^2 \) which appears, requires calculating the inner products: \( \langle w, w \rangle_G, \) and \( \langle v, v \rangle_G, \) where \( v \perp w \). This is described in greater depth from Amari’s work [5], and is in fact used to help prove its appearance in Theorem 10 in \( c_2(w) \) (or lack thereof in \( c_1(w) \)).

This is a form which allows the integral to be evaluated simply via sampling methods. Based on this, it is possible to derive the formula for the volume of the information manifold induced by the stochastic perceptron unit, where we write it in a form which makes clear of the contributions of the richness term, and the distinguishability term.
C.2 Proof of Theorem 5

Amari originally states the required form of $V$ shown via Equation (19) without proof, in [4]. However, in here we proceed to provide a short proof, and extend it slightly so that it is expressed in a form which makes clear of the existence of richness, and distinguishability terms.

Proof. Applying the matrix-determinant Lemma over the metric tensor in Theorem 10, we have

$$\det(G(w)) = \det(c_1(w)I) \left(1 - (c_2(w) - c_1(w)) \cdot e_w^T (c_1(w)I)^{-1} e_w \right)$$

$$= c_1(w)^D \cdot \left(1 - \frac{c_2(w) - c_1(w)}{c_1(w)} \right)$$

$$= c_1(w)^{D-1} c_2(w)$$

(18)

$$\Rightarrow V = \int_{\Omega} \sqrt{c_2(w)c_1(w)^{D-1}} dw.$$  

(19)

Since integration is performed over a $D$-ball of radius $w$, Amari writes: $dw = B^D(1)w^{D-1}dw$, which allows us to write,

$$\log V = \log B^D(1) + \log \int_{\Omega} \sqrt{c_2(w)c_1(w)^{D-1}} dw.$$  

Noticing that $c_1(w) = \frac{1}{\sigma^2} \mathbb{E}_\xi [f'(w\xi)]$, and $c_2(w) = \frac{1}{\sigma^2} \mathbb{E}_\xi [\xi^2 f'(w\xi)]$:

$$\log V = \log B^D(1) + \log \left(\frac{1}{\sigma^2}\right)^{D/2} \int_{\Omega} \sqrt{\mathbb{E}_\xi [\xi^2 f'(w\xi)] \mathbb{E}_\xi [f'(w\xi)]^{D-1}} dw$$

$$= \log \frac{B^D(1)}{\sigma^D} + \log \int_{\Omega} \sqrt{\mathbb{E}_\xi [\xi^2 f'(w\xi)] \mathbb{E}_\xi [f'(w\xi)]^{D-1}} dw$$

\[\square\]