NOTES ON KULLBACK–LEIBLER DIVERGENCE AND FISHER INFORMATION

IORDAN GANEV

Version 0.1

Contents

1. Introduction 1
2. Divergence 2
3. Statistical learning 5
4. Examples 8
References 10
Appendix A. Symmetries 12

1. Introduction

1.1. **Parameterized models.** In statistical learning, one often assumes a "true" probability density function $q(x)$ generating observed data. This density function is usually not directly tractable, but approximated with members of a parameterized family of densities $p(x \mid w)$, where the parameters w belong to a subset W of Euclidean space. For example, normal distributions are parameterized by their mean and variance, in which case we have $w = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$. In essence, one seeks $w \in W$ minimizing the "distance" between $p(x \mid w)$ and $q(x)$.

In this way, the question arises of how to define "distances" between distributions; this is answered by the notion of Kullback–Leibler (KL) divergence. In fact, the KL divergence defines a new geometry on W, where the "distance" between w_1 and w_2 is given by the KL divergence of $p(x \mid w_1)$ from $p(x \mid w_2)$. In general, this is not a metric (in particular, it usually differs from the Euclidean metric on W), but provides the probabilistically correct geometry on which to implement learning algorithms. Moreover, Fisher information can be thought of as the curvature of this probabilistic geometry on W.

1.2. **Outline.** In these expository notes, we begin by defining entropy, KL divergence, and Fisher information (Section 2). Our running example is the normal distribution. In Section 3, we relate these concepts back to statistical learning, with an emphasis on models with unit variance Gaussian noise. We examine a number of examples in detail in Section 4, including linear regression, two-layer neural networks, and models with parameter space symmetries. Finally, in Appendix A, we include somewhat advanced

material showing how symmetries can lead to singular models. We assume intermediate, mathematical familiarity with probability and statistics. References and motivation include [Car23, Bis13, Wat09, KR23].

1.3. **Notation.** We comment on some notation used in these notes. For an open set $W \subseteq \mathbb{R}^d$, the second-order Taylor expansion at $w_0 \in W$ of a real analytic function $f : W \to \mathbb{R}$ is:

$$f(w) = f(w_0) + (w - w_0)^T \nabla_{w_0} f + (w - w_0)^T \nabla^2_{w_0} f (w - w_0)^T + \dots$$

for w in an open neighborhood of w_0 in W, where $\nabla_{w_0} f \in \mathbb{R}^d$ is the gradient at w_0 and $\nabla^2_{w_0} f \in \mathbb{R}^{d \times d}$ is the Hessian matrix at w_0. We assume familiarity with tensor products, and freely use the isomorphism $\mathbb{R}^{m \times n} \simeq \mathbb{R}^m \otimes \mathbb{R}^n$ between the space of m by n matrices and the tensor product of $\mathbb{R}^m$ and $\mathbb{R}^n$. Additionally, we note that the natural map $\mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^m \otimes \mathbb{R}^n$ can be expressed as taking the pair of vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ the m by n matrix xy^T. Finally, we write 'log' for the natural logarithm.

2. Divergence

2.1. **Entropy.** Recall that a *probability density function*, or just *density*, on $\mathbb{R}^k$ is a function $p : \mathbb{R}^k \to \mathbb{R}$ satisfying:

(1) $p(x) \geq 0$ for all $x \in \mathbb{R}^k$.
(2) p is almost everywhere continuous.
(3) The Riemann integral[1] $\int_{\mathbb{R}^k} p(x)dx$ is equal to one.

The *cross-entropy* of a density p relative to a density q is defined as:

$$H[q, p] := -\int_{\mathbb{R}^k} q(x) \log p(x) dx$$

as long as the support of q is contained in the support of p (up to a set of measure zero); otherwise, the cross entropy is $+\infty$. If $p(x) = q(x) = 0$ for some x, then we set the product $q(x) \log p(x)$ to be equal to zero. The cross entropy is equal to the expected value of $-\log p(X)$ for a random variable X with density q. One can show that, for any p and q, we have:

$$H[q, p] \geq H[q, q] \tag{2.1}$$

with equality if and only if $p = q$ almost everywhere. The *entropy* of a density q is defined as its cross entropy relative to itself:

$$H[q] := H[q, q] = -\int_{\mathbb{R}^k} q(x) \log q(x) dx$$

[1]For $k = 1$, the bounded Riemann integral $\int_a^b p(x)dx$ exists if and only if $p(x)$ is almost everywhere continuous on $[a, b]$, i.e., if and only if the set of points where p is discontinuous has Lebesgue measure zero [Shr04, Theorem 1.3.8]. Taking limits (if they exist) gives the Riemann integral $\int_{\mathbb{R}} p(x)dx$. This generalizes to $\mathbb{R}^k$.

Finally, the *relative entropy* of a density p relative to a density q is defined as:

$$D_{\mathrm{KL}}(q||p) := -\int_{\mathbb{R}^k} q(x)\log\frac{p(x)}{q(x)}dx = H[p,q] - H[q]$$

This quantity is also known as the *Kullback–Leibler divergence*, or just *KL divergence*. Again, the definition requires that the support of q is almost everywhere contained in the support of p; otherwise, the KL divergence is $+\infty$. By virtue of 2.1, the KL divergence is non-negative, and is equal to zero if and only if $p(x) = q(x)$ almost everywhere. The KL divergence is not a metric, and is not even symmetric. Still, it is the correct information-theoretic way to measure the "distance" that p is from q. One interpretation of the KL divergence is as the expected excess surprise from sampling from p as an approximation for the actual density q.

Example 2.1. Let $r > 0$. The uniform density is defined as $p(x) = 1/r$ for $x \in [0,r]$ and zero otherwise. Its entropy is $H[p] = \log(r)$, so increases with r, and is negative for $0 < r < 1$. Suppose q and p are uniform densities on $[0,r]$ and on $[0,s]$, respectively. Suppose further that $r \leq s$ so that the support of q is contained in the support of p. Then their relative entropy is $D_{\mathrm{KL}}(p||q) = \log(s) - \log(r)$.

Example 2.2. The normal density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_{>0}$ is defined as $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$. Its entropy is calculated[2] to be:

$$H[p] = \log\left(\sigma\sqrt{2\pi e}\right) = \log(\sigma) + \log\left(\sqrt{2\pi}\right) + \frac{1}{2},$$

which increases with σ (and doesn't depend on μ). Suppose q and p are normal densities with means μ_0, μ and variances σ_0^2 and σ^2, respectively. The cross entropy[3] of q and p is

$$H[q,p] = \log(\sigma) + \log\left(\sqrt{2\pi}\right) - \frac{\sigma_0 + (\mu-\mu_0)^2}{2\sigma^2}$$

and hence the relative entropy of p relative to q is:

$$D_{\mathrm{KL}}(q||p) = \log\frac{\sigma}{\sigma_0} + \frac{\sigma_0^2 + (\mu-\mu_0)^2}{2\sigma^2} - \frac{1}{2}.$$

If $\sigma_0 = \sigma$, then this reduces to a multiple of the squared Euclidean distance between the means: $\frac{1}{2\sigma^2}|\mu - \mu_0|^2$.

2.2. **Families.** Let $\mathsf{PD}(\mathbb{R}^k)$ denote the set of probability density functions on $\mathbb{R}^k$. Let $W \subseteq \mathbb{R}^d$ be a non-empty open subset of Euclidean space $\mathbb{R}^d$. An *(analytic) W-family of densities* is the assignment of a density on $\mathbb{R}$ for every $w \in W$:

$$W \to \mathsf{PD}(\mathbb{R}^k), \qquad w \mapsto p(-|w)$$

We think of W as the set of parameters for a family of distributions. We make the following assumptions (some of which may be weakened without much effect on the results below):

[2]One uses the substitution $u = \frac{x^2}{2\sigma^2}$ and a value of the Gamma function: $\Gamma(3/2) = \sqrt{\pi}/2$.

[3]The substitution is $u = \frac{(x-\mu_0)^2}{2\sigma_0^2}$, and one uses $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(3/2) = \sqrt{\pi}/2$.

- The density $p(-|w)$ is analytic for any $w \in W$.
- For each $x \in \mathbb{R}$, the map $W \to \mathbb{R}$ defined as $w \mapsto p(x|w)$ is analytic.
- The entropy $H[p(-|w)]$ converges for every $w \in W$.
- The relative entropy $D_{\mathrm{KL}}\left(p(-|w_0)||p(-|w_1)\right)$ converges for every $w_0, w_1 \in W$
- The support of $p(- \mid w)$ is independent of w.

A W-family is also knowns as a *statistical model* parameterized by W. By slight abuse of notation, we set:

$$D_{\mathrm{KL}}(w_0||w_1) := D_{\mathrm{KL}}\left(p(-|w_0)||p(-|w_1)\right)$$

In this way, we pull back the relative entropy to a real-valued binary function on $W \subseteq \mathbb{R}^d$.

Example 2.3. Let $W = \mathbb{R} \times \mathbb{R}_{>0}$. For $w = (\mu, \sigma^2) \in W$, set $p(x|w)$ to be the normal density with mean μ and variance σ^2. From Example 2.2, we have $D_{\mathrm{KL}}(w_0||w_!) = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_1 - \mu_0)^2}{2\sigma_1^2} - \frac{1}{2}$.

Definition 2.4. Fix $w_0 \in W$. The *KL divergence from* w_0 is the function

$$D_{w_0} := D_{\mathrm{KL}}(w_0 \mid\mid -) : W \to \mathbb{R}$$

so that $D_{w_0}(w) = D_{\mathrm{KL}}(w_0 \mid\mid w)$. When w_0 is clear from context, we write simply D for D_{w_0}. Meanwhile, the *Fisher information* at $w_0 \in W$ is defined as the Hessian of the function D_{w_0} at $w = w_0$. This gives a map:

$$I : W \to \mathbb{R}^{d \times d}, \qquad w_0 \mapsto \nabla^2_{w_0}(D_{w_0})$$

The proof of the following lemma is a straightforward verification.

Lemma 2.5. *Fix $w_0 \in W$. The gradient of the divergence from w_0 vanishes at $w = w_0$:*

$$\nabla_{w_0}(D_{w_0}) = 0$$

Example 2.6. Continuing the example of the normal distribution family, the gradient of $D = D(w_0||-)$ at w is:

$$\nabla_w D = \left(\frac{\mu - \mu_0}{\sigma^2} \, , \, \frac{\sigma^2 - \sigma_0^2 + (\mu - \mu_0)^2}{\sigma^3} \right),$$

which evaluates to zero if and only if $w = w_0$. The Hessian is:

$$\nabla^2_w D = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{2(\mu - \mu_0)}{\sigma^3} \\ \frac{2(\mu - \mu_0)}{\sigma^3} & \frac{-1}{\sigma^2} + \frac{3\left(\sigma_0^2 + (\mu - \mu_0)^2\right)}{\sigma^4} \end{bmatrix}$$

At $w = w_0$, we obtain the Fisher information, which is positive definite:

$$I(w_0) = \nabla^2_{w_0} D = \frac{1}{\sigma_0^2} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

Returning to the general case, the fact that $D_{w_0}(w_0) = 0$ means that D vanishes up to first order at w_0, and the curvature of the KL divergence from w_0 is given by the Fisher information matrix at w_0. Since D has a global minimum at w_0, the Fisher information $I(w_0)$ is always positive semi-definite. We introduce some terminology. A W-family is:

- *identifiable* if the map $W \to \mathsf{PD}(\mathbb{R}^k)$ is injective.
- *positive definite* if the Fisher information matrix $I(w)$ is positive definite for every $w \in W$.
- *regular* if it is both identifiable and positive definite
- *singular* if it is not regular.
- *strictly singular* if it is not identifiable and not positive definite.

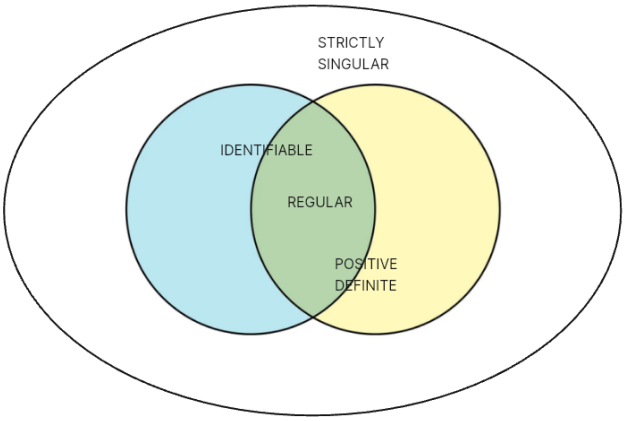

FIGURE 1. The outer circle represents all possible W-families. The left inner circle is the set of identifiable families, while the right inner circle is the set of positive definite families. They overlap in the set of regular families. All models that are not regular are singular, while those that are both non-identifiable and indefinite are strictly singular.

See Figure 1 for a diagrammatic description of these definitions. Our terminology differs from that in [Wato9, Section 1.2]; we find the latter presentation potentially inconsistent.

We end this discussion by noting that the Fisher information is the same as the (negative) expected value of the Hessian $\nabla_{w_0} \log p(X|w)$ for a random variable X valued in $\mathbb{R}^k$ with density $p(x|w_0)$. In symbols:

$$I(w_0) = -\mathbb{E}_{X \sim p(-|w_0)} \left[\nabla^2_{w=w_0} \log p(x|w) \right]$$

3. Statistical learning

3.1. **Sampling.** We focus on Bayesian supervised learning with N-dimensional input space and M-dimensional output space. A *sample* is a random variable (X, Y) valued in $\mathbb{R}^N \times \mathbb{R}^M$ with density function

$$q(x, y) : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$$

We can write $q(x, y) = q(y|x)q(x)$ as the product of the distribution of the input $q(x)$ and the conditional distribution $q(y|x)$ of the output given an input. Assumptions:

- The input distribution $q(x)$ is known. We will refer to this as the *sampling density*.
- The joint distribution $q(x, y)$, while unknown, has finite variance.

3.2. **Models.** To model the unknown distribution $q(y|x)$, we choose a family of conditional distributions parametrized by an open subset $W \subseteq \mathbb{R}^d$. Formally, we have a map:

$$p : W \times \mathbb{R}^N \to \mathsf{PD}(\mathbb{R}^M), \qquad (w, x) \mapsto p(-|x, w)$$

such that $w \mapsto p(y|x, w)$ is an analytic map on W for any $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$. Since the input distribution $q(x)$ is known, the modeled sampling distribution for a given $w \in W$ is:

$$p_w(x, y) := p(y|x, w)q(x)$$

which we may also denote $p(x, y|w)$. Comparing to the true distribution $q(x, y)$ via relative entropy, we get a function on W:

$$\begin{aligned} K : W &\to \mathbb{R} \\ w &\mapsto D_{\mathrm{KL}}(q||p_w) \end{aligned}$$

This is our "loss metric"; statistical learning techniques seek $w \in W$ that minimizes this function. There is a $w_0 \in W$ such that $q = p_{w_0}$ almost everywhere if and only if the fiber $K^{-1}(0)$ is nonempty. The assumption of analyticity implies that this fiber is a real analytic set.

3.3. **Unit variance Gaussian noise.** One source of the modeled conditional distributions is from parameterized families of functions $\mathbb{R}^N \to \mathbb{R}^M$. Specifically, an analytic function $f : W \times \mathbb{R}^N \to \mathbb{R}^M$ can be thought of as a family of functions $\mathbb{R}^N \to \mathbb{R}^M$, and each such family gives rise to a conditional distribution by adding unit variance Gaussian noise:

$$p(y|x, w) = \frac{1}{(\sqrt{2\pi})^M} \exp\left(-\frac{1}{2}\|y - f(w, x)\|^2\right) \tag{3.1}$$

Suppose the true conditional distribution is also unit variance Gaussian, i.e.,

$$q(y|x) = \frac{1}{(\sqrt{2\pi})^M} \exp\left(\frac{1}{2}\|y - \mu(x)\|^2\right)$$

for some function $\mu : \mathbb{R}^{n_0} \to \mathbb{R}$. Since we are working with unit variance Gaussians, the entropy of $p_w(x, y) = p(y|x, w)q(x)$ relative to the the true density $a(x, y) = q(y|x)q(x)$ is obtained as the expected value of the squared distance between the outputs at the two parameter values:

$$K(w) = \frac{1}{2} \int_{\mathbb{R}^{n_0}} q(x)\|f(w, x) - \mu(x)\|^2 dx$$

(See Example 2.2 above.) As mentioned, we seek points in the fiber $K^{-1}(0)$; these points will all be critcal points of K. We note the inclusion, which may be proper:

$$\{w \in W \mid \mu = f(w, -) \text{ almost everywhere}\} \subseteq K^{-1}(0).$$

Depending on the nature of μ and q, the fiber $K^{-1}(0)$ may be empty.

3.4. **Unit variance Gaussian noise and same model class.** Continuing the set-up from above, suppose the true conditional distribution is also unit variance Gaussian, with mean coming from the W-family, so that:

$$q(y|x) = \frac{1}{(\sqrt{2\pi})^M} \exp\left(\frac{1}{2}\|y - f(w_0, x)\|^2\right)$$

for some $w_0 \in W$. Hence, $K = D_{KL}(w_0||-)$ and $w_0 \in K^{-1}(0)$. Set:

$$\mathsf{Jac}_w = df(-, x)_w \in \mathbb{R}^{M\times d}$$

to be the Jacobian of $f(-, x) : W \to \mathbb{R}^M$. We have:

$$\nabla_w K = \int_{\mathbb{R}^N} q(x)\mathsf{Jac}_w^T \left(f(w, x) - f(w_0, x)\right) dx \in \mathbb{R}^d,$$

which vanishes at $w = w_0$. To be clear, the transpose Jac_w^T a d by M matrix which is multiplied by the M-vector $f(w, x) - f(w_0, x)$. Similarly, we have:

$$\nabla^2_{w_0} K = \int_{\mathbb{R}^N} q(x)\mathsf{Jac}_{w_0}^T \mathsf{Jac}_{w_0} dx = \mathbb{E}[\mathsf{Jac}_{w_0}] \in \mathbb{R}^{d\times d}$$

3.5. **Free energy.** Let $\{(X_1, Y_1), (X_2, Y_2), \dots\}$ be a sequence of random variables valued in $\mathbb{R}^N \times \mathbb{R}^M$ that is independent and identically distributed according to q. Given a sample $D_n = \{(x_i, y_i)\}_{i=1}^n$ of these random variables, we have the empirical cross entropy for any $w \in W$:

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^n \log p(y_i|x_i, w)$$

The assumption of finite variance and the central limit imply that $L_n(w)$ limits to the cross entropy $H[q||p_w]$ as $n \to \infty$.

Now suppose we have a prior density $\phi : W \to \mathbb{R}$ on the parameter space W such that $\phi(w) > 0$ for all $w \in W$. Given data D_n, Bayes' rule yields the posterior density:

$$p(w|D_n) = \frac{p(D_n|w)\phi(w)}{p(D_n)}$$

The denominator of the posterior is known as the *partition function* or *evidence*; using the law of total probability and the fact that $p(D_n|u) = \prod_{i=1}^n p(y_i|x_i, u) = \exp\left(-nL_n(u)\right)$, the partition function is given by:

$$Z_n = \int_W p(D_n|u)\phi(u)du = \int_W \exp\left(-nL_n(u)\right)\phi(u)du$$

Hence, the posterior takes the form:

$$p(w|D_n) = \frac{\exp\left(-nL_n(w)\right)\phi(w)}{Z_n}$$

Taking logs of both sides yields:

$$\log p(w|D_n) = \log\phi(w) + F_n - nL_n(w)$$

where $F_n = -\log Z_n$ is the *free energy*. A later version of these notes may feature a more detailed discussion of free energy; for now, we move on to examples.

4. Examples

In each of the following examples, we consider a specific model class $f : W \times \mathbb{R}^N \to \mathbb{R}^M$. We add unit variance Gaussian noise, fix a sample density $q(x)$, and obtain a family of distributions $p_w(x, y) = p(y|x, w)q(x)$ as in equation 3.1. We assume the true density belongs to our model class so that $q = p_{w_0}$ for some $w_0 \in W$, and form the relative entropy function $K = D_{\mathrm{KL}}(w_0||-)$. For each example, we compute the gradient $\nabla_w K$ at an arbitrary point in W, and the Hessian $\nabla^2_{w_0} K$ at the point w_0.

4.1. **Linear regression.** We model inputs in $\mathbb{R}^N$ and outputs in $\mathbb{R}^M$ via an affine linear transformation $x \mapsto Ax + b$ together with unit variance Gaussian noise. Specifically, the parameter space is

$$W = \mathbb{R}^{M\times N} \oplus \mathbb{R}^M = \mathbb{R}^M \otimes \mathbb{R}^{N+1}$$

The map f is given by:

$$f : W \times \mathbb{R}^N \to \mathbb{R}^M, \qquad (A, b, x) \mapsto Ax + b$$

Computing derivatives, we obtain the Jacobian:

$$\mathsf{Jac}_w := df(-, x)_w = \mathrm{id}_M \otimes \begin{bmatrix} x \\ 1 \end{bmatrix} \in \mathbb{R}^M \otimes \mathbb{R}^M \otimes \mathbb{R}^{N+1}$$

It follows that:

$$\nabla_w D = \int_{\mathbb{R}^N} q(x)\,(Ax + b - A_0x - b_0) \otimes \begin{bmatrix} x \\ 1 \end{bmatrix} dx \in \mathbb{R}^M \otimes \mathbb{R}^{N+1}$$

$$\nabla^2_{w_0} D = \mathrm{id}_M \otimes \int_{\mathbb{R}^N} q(x) \begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x^T & 1 \end{bmatrix} dx = \mathrm{id}_M \otimes \begin{bmatrix} \mathbb{E}[XX^T] & \mathbb{E}[X] \\ \mathbb{E}[X^T] & 1 \end{bmatrix}$$

where we use express the Hessian as an element of

$$\begin{aligned} W \otimes W &= (\mathbb{R}^M \otimes \mathbb{R}^{N+1}) \otimes (\mathbb{R}^M \otimes \mathbb{R}^{N+1}) \\ &\simeq \mathbb{R}^{M\times M} \otimes \left(\mathbb{R}^{N\times N} \oplus \mathbb{R}^N \oplus \mathbb{R}^N \oplus \mathbb{R}\right) \end{aligned}$$

The Hessian, and hence the Fisher information, does not depend on w_0. We conclude that the Fisher information is positive definite if and only if variance-covariance matrix of the input sampling density p is positive definite.

4.2. **Composition.** Suppose we have smooth maps $f : U \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$ and $g : V \times \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ where U and V are open subsets of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. The parameter space in this case is $W = U \times V \subseteq \mathbb{R}^{d_1+d_2}$ and the model is the composition:

$$W \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_2}, \qquad (u, v, x) \mapsto g(v, f(u, x))$$

To compute the gradient and Hessian of the function K, we set:

$$\begin{aligned} A &= dg(v, f(-, x))_u = dg(v, -)_{f(u,x)} \circ df(-, x)_u && \in \mathbb{R}^{n_2\times d_1} \\ B &= dg(-, f(u, x))_v && \in \mathbb{R}^{n_2\times d_2} \end{aligned}$$

where we use the chain rule in the first line. Then:

$$\nabla_{(u,v)} K = \int_{\mathbb{R}^{n_0}} p(x) \begin{bmatrix} A^T \\ B^T \end{bmatrix} (g(v, f(u,x)) - g(v_0, f(u_0,x)))dx$$
$$\nabla^2_{(u_0,v_0)} K = \int_{\mathbb{R}^{n_0}} p(x) \begin{bmatrix} A_0^T A_0 & A_0^T B_0 \\ B_0^T A_0 & B_0^T B_0 \end{bmatrix} dx = \begin{bmatrix} \mathbb{E}[A_0^T A_0] & \mathbb{E}[A_0^T B_0] \\ \mathbb{E}[B_0^T A_0] & \mathbb{E}[B_0^T B_0] \end{bmatrix}$$

where A_0 and B_0 are obtained from A and B by substituting (u_0, v_0) for (u, v).

4.3. **Two-layer neural network.** Consider a two-layer neural network with layer sizes (n_0, n_1, n_2) and activation[4] $\sigma : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$. Then the parameter space consists of two matrices and two bias vectors:

$$W = (\mathbb{R}^{n_1 \times n_0} \times \mathbb{R}^{n_1}) \times (\mathbb{R}^{n_2 \times n_1} \times \mathbb{R}^{n_2})$$

We have the feedforward function[5]:

$$F : W \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_2}, \qquad (u, b, v, c, x) \mapsto v\sigma(ux + b) + c$$

As this is a composition, we can apply the analysis of the previous section with $f((u,b),x) = \sigma(ux+b)$ and $g((v,c),z) = vz + c$. This yields:

$$A = (V \circ d\sigma_z) \otimes \begin{bmatrix} x \\ 1 \end{bmatrix} \qquad\qquad B = \mathrm{id}_{n_2} \otimes \begin{bmatrix} \sigma(z) \\ 1 \end{bmatrix}$$

where $z = ux + b \in \mathbb{R}^{n_1}$, and $d\sigma_z \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_1}$ is the Jacobian of σ at z. Hence:

$$A_0^T A_0 = \begin{bmatrix} x \\ 1 \end{bmatrix} \otimes \left(d\sigma_{z_0}^T \circ v_0^T v_0 \circ d\sigma_{z_0} \right) \otimes \begin{bmatrix} x \\ 1 \end{bmatrix} \qquad \in \mathbb{R}^{n_0+1} \otimes \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_0+1}$$
$$A_0^T B_0 = \begin{bmatrix} x \\ 1 \end{bmatrix} \otimes \left(d\sigma_{z_0}^T \circ v_0^T \right) \otimes \begin{bmatrix} \sigma(z_0) \\ 1 \end{bmatrix} \qquad \in \mathbb{R}^{n_0+1} \otimes \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_1+1}$$
$$B_0^T B_0 = \begin{bmatrix} \sigma(z_0) \\ 1 \end{bmatrix} \otimes \mathrm{id}_{n_2} \otimes \begin{bmatrix} \sigma(z_0) \\ 1 \end{bmatrix} \qquad \in \mathbb{R}^{n_1+1} \otimes \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_1+1}$$

where $w_0 = (u_0, b_0, v_0, c_0)$ and $z_0 = u_0 x + b_0$.

4.4. **Parameter space symmetries.** Let G be a Lie group acting on W, and let $f : W \times \mathbb{R}^N \to \mathbb{R}^M$ be such that $f(gw, x) = f(w, x)$ for any $w \in W$, $x \in \mathbb{R}^N$, and $g \in G$. Fixing $w_0 \in W$ and a sampling density, we form the corresponding function $K : W \to \mathbb{R}$ as above. Suppose the orbit of w_0 under G is of positive dimension. Then the Hessian $\nabla^2_{w_0} K$ of K at w_0 not positive definite; indeed, the infinitesimal action of G provides vectors in the kernel of the Hessian. See Appendix A for more details.

[4]While the activation is often pointwise, meaning the same function $\mathbb{R} \to \mathbb{R}$ is applied in each coordinate, we consider the more general case.

[5]Somewhat unconventionally, we use u and v to denote matrices, since U and V are reserved for open subsets of Euclidean space.

4.4.1. *Simple example.* Consider the action of $G = \mathbb{R}_{>0}$ on $W = \mathbb{R}^2$ and via $\lambda \cdot (u, v) = (\lambda u, \lambda^{-1} v)$. The Lie algebra is $\mathrm{Lie}(\mathbb{R}_{<0}) = \mathbb{R}$, and the infinitesimal action of $\xi = 1 \in \mathbb{R}$ at $w = (u, v)$ is the (tangent) vector $\xi_w = (u, -v)$. The function:

$$f : W \times \mathbb{R} \to \mathbb{R}, \qquad ((u, v), x) \mapsto vux$$

is G-invariant in the W factor. The corresponding function K is given by:

$$K(w) = \frac{1}{2} \int_{\mathbb{R}^N} q(x)(vux - v_0 u_0 x)^2 dx$$

The Hessian at $w_0 = (u_0, v_0)$ is easily seen to be:

$$\nabla^2_{w_0} K = \mathbb{E}[X^2] \begin{bmatrix} v_0^2 & u_0 v_0 \\ u_0 v_0 & u_0^2 \end{bmatrix}$$

If $u_0 = v_0 = 0$, the Hessian is zero. Otherwise, the vector $(u_0, -v_0)$ belongs to the kernel of the Hessian, and this is precisely a vector provided by the infinitesimal action. We conclude that the Hessian is not invertible, regardless of the value of w_0.

4.4.2. *Neural network examples.* More sophisticated examples in the context of neural networks come from choosing activations with symmetries. Specifically, in the two-layer case with layer sizes (n_0, n_1, n_2) and activation $\sigma : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$, the parameter space consists of two matrices and two bias vectors:

$$W = (\mathbb{R}^{n_1 \times n_0} \times \mathbb{R}^{n_1}) \times (\mathbb{R}^{n_2 \times n_1} \times \mathbb{R}^{n_2})$$

There is an action of $G = \mathrm{GL}_{n_1}(\mathbb{R})$ on the parameter space via:

$$g \cdot (u, b, v, c) = (gu, gb, vg^{-1}, c)$$

Consider the following cases:

- For the trivial activation $\sigma = \mathrm{id}_{n_1}$, the feedforward function is invariant for the action of the entire group $G = \mathrm{GL}_{n_1}(\mathbb{R})$. The orbit of any point other than zero is of positive dimension.
- For the pointwise ReLU activation $\sigma(z) = (\mathrm{ReLU}(z_1), \ldots, \mathrm{ReLU}(z_{n_1}))$, the feedforward function is invariant for the subgroup of G consisting diagonal matrices with positive entries along the diagonal. The orbit of any nonzero point is of positive dimension. Note that the feedforward is not smooth when using ReLU activations; however, the theory still holds for piecewise smooth functions.
- For radial activations (see [GLW23]), the feedforward function is invariant for the subgroup of orthogonal matrices in G. The orbit of any nonzero point is of positive dimension.

For further analysis of parameter space symmetries, see [ZGW+23].

References

[Bis13] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer (India) Private Limited, 2013.

[Car23] Liam Carroll, *Distilling Singular Learning Theory — LessWrong*, 2023.

[GLW23] Iordan Ganev, Twan van Laarhoven, and Robin Walters, *Universal approximation and model compression for radial neural networks*, arXiv, 2023. arXiv:2107.02550.

[KR23] Mohammad Emtiyaz Khan and Håvard Rue, *The Bayesian learning rule*, J. Mach. Learn. Res. **24** (2023), no. 1, 281:13328–281:13373.

[Shr04] Steven E. Shreve, *Stochastic Calculus for Finance II: Continuous-Time Models*, Springer Science & Business Media, 2004 (en).

[Wat09] Sumio Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2009.

[ZGW+23] Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy, *Symmetries, flat minima, and the conserved quantities of gradient flow*, ICLR (2023).

Appendix A. Symmetries

A.1. **Jacobians, gradients, and Hessians.** Let $X \subset \mathbb{R}^n$ be an open subset of Euclidean space $\mathbb{R}^n$, and let $f : X \to \mathbb{R}^m$ be a differentiable function. Let $f_1, \dots, f_m : X \to \mathbb{R}$ be the components of f, so that $f(x) = (f_1(x), \dots, f_m(x))$. The Jacobian of f, also know as differential of f, at $x \in X$ is the following matrix of partial derivatives evaluated at x, using the standard coordinates $(x_1, \dots, x_n)$ on $\mathbb{R}^n$:

$$df_x := \begin{bmatrix} \left.\frac{\partial f_1}{\partial_{x_1}}\right|_x & \left.\frac{\partial f_1}{\partial_{x_2}}\right|_x & \cdots & \left.\frac{\partial f_1}{\partial_{x_n}}\right|_x \\ \left.\frac{\partial f_2}{\partial_{x_1}}\right|_x & \left.\frac{\partial f_2}{\partial_{x_2}}\right|_x & \cdots & \left.\frac{\partial F_2}{\partial_{x_n}}\right|_x \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial f_m}{\partial_{x_1}}\right|_x & \left.\frac{\partial f_m}{\partial_{x_2}}\right|_x & \cdots & \left.\frac{\partial f_m}{\partial_{x_n}}\right|_x \end{bmatrix}$$

The differential df_x is a matrix in $\mathbb{R}^{m\times n}$, and hence defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, or, more precisely, from the tangent space T_xX of X at x to the tangent space $T_{f(x)}\mathbb{R}^m$ of $\mathbb{R}^m$ at $f(x)$. Observe that if f itself is linear, then, as matrices, $df_x = f$ for all points $x \in X$. If $g : \mathbb{R}^m \to \mathbb{R}^p$ is another differentiable map, then the chain rule implies that, for all $x \in \mathbb{R}^n$, we have:

$$d(g \circ f)_x = dg_{f(x)} \circ df_x.$$

In the special case the $m = 1$, the Jacobian is a $1 \times n$ row vector, and the *gradient* $\nabla_x f$ of f at $x \in X$ is defined as the transpose of the Jacobian df_x:

$$\nabla_v f := (df_x)^T = \begin{bmatrix} \left.\frac{\partial f}{\partial x_1}\right|_x & \cdots & \left.\frac{\partial f}{\partial x_n}\right|_x \end{bmatrix}^T = \left(\left.\frac{\partial f}{\partial x_i}\right|_x \right)_{i=1}^n \in \mathbb{R}^n$$

Moreover, the Hessian of f at x is defined as the matrix of second partial derivatives:

$$\nabla_x^2 f := \begin{bmatrix} \left.\frac{\partial^2 f}{\partial x_1 \partial x_1}\right|_x & \cdots & \left.\frac{\partial^2 f}{\partial x_1 \partial x_n}\right|_x \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 f}{\partial x_n \partial x_1}\right|_x & \cdots & \left.\frac{\partial^2 f}{\partial x_n \partial x_n}\right|_x \end{bmatrix} \in \mathbb{R}^{n\times n}$$

The Hessian is symmetric, and can be realized at the Jacobian matrix of the map $\nabla f : X \to \mathbb{R}^n$ taking x to the gradient $\nabla_x f$; that is: $\nabla_x^2 f = d\left(\nabla f\right)_x$.

A.2. **Group actions.** Let G be a Lie group acting smoothly on an open set $X \subseteq \mathbb{R}^n$. In other words, we have a smooth map:

$$a : G \times X \to X$$

that satisfies (1) the unit axiom, namely $a(1_G, x) = x$ for all $x \in X$ where $1_G \in G$ is the identity of G, and (2) the associativity axiom, namely $a(g, a(h, x)) = a(gh, x)$ for all $g, h \in G$ and $x \in X$. We write $\rho(g) = a(g, -) : X \to X$ for the automorphism of X

corresponding to $g \in G$. Let $\mathfrak{g}$ be the Lie algebra of G, realized as the tangent space $T_{1_G}G$ to G at the identity. The differential of the map $a(-,x) : G \to X$ at 1_G gives a map

$$d(a(-,x))_1 : \mathfrak{g} \to T_x X$$

The *infinitesimal action* of a Lie algebra element $\xi \in \mathfrak{g}$ at $x \in X$ is defined as the image of ξ under this map, which is a tangent vector to x:

$$\overline{\xi}_x := d(a(-,x))_1(\xi) \in T_x X \simeq \mathbb{R}^n$$

The infinitesimal action can be described in terms of the exponential map $\exp : \mathfrak{g} \to G$ as:

$$\overline{\xi}_x = \frac{d}{d_t}\bigg|_{t=0} a\left(\exp(t\xi), x\right),$$

noting that $\mathfrak{g}$ is a vector space, hence it makes sense to scale a Lie algebra element $\xi \in \mathfrak{g}$ by a scalar $t \in \mathbb{R}$.

A.3. **Equivariant maps.** Let $a : G \times X \to X$ be a group action as above, setting $\rho(g) = a(g,-) : X \to X$. We abbreviate $a(g,x)$ as just gx. A smooth function $f : X \to \mathbb{R}$ is G-invariant function if $f(gx) = f(x)$ for all $g \in G$ and $x \in X$.

Lemma A.1. *Suppose $f : X \to \mathbb{R}$ is G-invariant, and let $x \in X$.*

(1) *For any $g \in G$, the gradient of f at gx is obtained from that at x via via the transpose of the Jacobian matrix of $\rho(g)$ at gx:*

$$\nabla_{gx} f = d\rho(g^{-1})^T_{gx}(\nabla_x f)$$

(2) *For any $\xi \in \mathfrak{g}$, the infinitesimal action $\overline{\xi}_x$ is orthogonal to the gradient:*

$$\langle \nabla_x f, \overline{\xi}_x \rangle = 0$$

Sketch of proof. For the first claim, we first compute differentials:

$$df_{gx} = d(f \circ \rho(g^{-1}))_{gx} = df_x \circ d(\rho(g^{-1}))_{gx}$$

where we use the invariance of f and chain rule. Taking transposes gives the result. For the second claim, we note that $f \circ a(-,x)$ is a constant function $G \to \mathbb{R}$ taking every element to $f(x)$. Hence $d(f \circ a(-,x))_{1_G} = 0$. The result follows from chain rule and taking transposes. □

Corollary A.2. *Suppose $f : X \to \mathbb{R}$ is G-invariant, and let $x \in X$ be a critical point. Then:*

(1) *For any $g \in G$, the point gx is a critical point of f, of the same type (local minimum, local maximum, or saddle point) as x.*
(2) *For any $\xi \in \mathfrak{g}$, the infinitesimal action $\overline{\xi}_x$ belongs to the kernel of the Hessian of f at x:*

$$\nabla^2_x f\left(\overline{\xi}_x\right) = 0$$

Sketch of proof. The first claim is immediate from the previous lemma and the fact that $\rho(g)$ is a diffeomorphism commuting with f. The first claim implies that:

$$\nabla f \circ a(-,x)(g) = \nabla_{gx} f = 0$$

for all $g \in G$. Hence the differential $d(f \circ a(-, x))_{1_G}$. The second claim follows. $\square$

A.4. **Linear actions.** Suppose $X = \mathbb{R}^n$ and the action of G is linear, so that ρ is a group homomorphism $G \to \mathrm{GL}_n(\mathbb{R})$. Then we have:

$$\nabla_{gx} f = \rho(g^{-1})^T (\nabla_x f) \qquad \nabla^2_{gx} f = \rho(g^{-1})^T (\nabla^2_x f) \rho(g^{-1})$$

To be clear, the right-hand side of the first equation is the multiplication of the n by n matrix $\rho(g^{-1})^T$ with the vector $\nabla_x f \in \mathbb{R}^n$, while that of the second equation is the multiplication of three n by n matrices. Two special cases are worth noting. If the action of G is by orthogonal transformations, so that ρ factors through the orthogonal group $O(n)$, then $\rho(g^{-1})^T = \rho(g)$ for any $g \in G$, and we have:

$$\nabla_{gx} f = \rho(g)(\nabla_x f) \qquad \nabla^2_{gx} f = \rho(g)(\nabla^2_x f) \rho(g)^{-1}$$

Another special case is when the action of G is symmetric, so that $\rho(g)$ is a symmetric matrix for every $g \in G$. Then $\rho(g)^T = \rho(g)$ and we have:

$$\nabla_{gx} f = \rho(g^{-1})(\nabla_x f) \qquad \nabla^2_{gx} f = \rho(g^{-1})(\nabla^2_x f) \rho(g^{-1})$$