Dans ce document, je veux compiler de façon tantôt brève, tantôt plus détaillée, diverses notions utiles à notre étude.

Contents

1 Théorie de l'information et statistiques **3**
- 1.1 L'information de Fisher et la borne de Cramér-Rao . . . 3
- 1.2 Entropie et statistique suffisante . . . 5
- 1.3 Deux théorèmes . . . 6
  - 1.3.1 Théorème de Rao-Blackwell . . . 6
  - 1.3.2 Théorème de Fisher-Darmois-Pitman-Koopman . . . 7
  - 1.3.3 Aparté . . . 7

2 Integration on a manifold **7**
- 2.1 Why differential forms? . . . 7
- 2.2 k-forms, wedge product, exterior derivative . . . 12
  - 2.2.1 The wedge product . . . 12
  - 2.2.2 The exterior derivative . . . 13
- 2.3 Stokes' theorem . . . 14

3 Integration theorems of distributions **14**
- 3.1 Differential Equations . . . 15
- 3.2 Vector fields, integral curves, and flows . . . 15
  - 3.2.1 Integral curves . . . 16
  - 3.2.2 Flows . . . 16
- 3.3 Lie Derivatives . . . 17
- 3.4 Commuting Vector Fields . . . 18
- 3.5 Involutivity . . . 19
  - 3.5.1 Integral manifold and involutivity . . . 19
  - 3.5.2 Involutivity and Differential Forms . . . 20
- 3.6 The Frobenius Theorem . . . 21
  - 3.6.1 Proof of the Frobenius Theorem . . . 21
  - 3.6.2 Summary of the proof of the Frobenius Theorem . . . 23
  - 3.6.3 Method to find integral manifolds . . . 23
- 3.7 Foliations . . . 24

4 Curvature **24**
- 4.1 Tensors . . . 24
  - 4.1.1 Aside: definition of the tensor product . . . 24
  - 4.1.2 Tensors on a vector space . . . 25
  - 4.1.3 Tensor bundles . . . 27
- 4.2 Metrics . . . 27
- 4.3 Affine connections . . . 28
  - 4.3.1 Intuition . . . 28
  - 4.3.2 Coordinate change . . . 28
  - 4.3.3 Parallel translation . . . 29
  - 4.3.4 Covariant derivative along a curve . . . 29
  - 4.3.5 Covariant derivative along a tangent vector . . . 30
  - 4.3.6 Comparison with the directional derivative in $\mathbb{R}^n$ and with the Lie derivative . . . 31
  - 4.3.7 Metric connection . . . 32
- 4.4 Flatness . . . 32

5 Geometrical Structures of a Family of Probability Distributions **33**
5.1 Part I: Differential Geometry of Statistical Models . . . 33
5.1.1 Manifold of statistical model . . . 33
5.1.2 Tangent space . . . 34
5.1.3 Riemannian metric and Fisher information . . . 34
5.1.4 Les connexions affines expliquées intuitivement . . . 34
5.1.5 Statistical α-connection . . . 36
5.1.6 Curvature and torsion . . . 38
5.1.7 Imbedding and submanifold . . . 39
5.1.8 Family of ancillary submanifolds . . . 40
5.2 Part II: α-Divergence and α-Projection in Statistical Manifold . . . 41
5.2.1 α-representation . . . 41
5.2.2 Dual affine connections . . . 42
5.2.3 α-family of distributions . . . 44
5.2.4 Duality in α-flat manifolds . . . 45

6 Methods of Information Geometry **45**
6.1 The geometric structure of statistical models . . . 46
6.1.1 The Fisher metric . . . 46
6.1.2 The α-connection . . . 48
6.1.3 Chentsov's theorem . . . 49
6.1.4 The geometry of $\mathcal{P}(\mathcal{X})$ . . . 49
6.2 Dual connections . . . 51
6.2.1 Contrast functions . . . 51
6.2.2 Dually flat spaces . . . 53
6.2.3 Canonical divergence . . . 54
6.2.4 The dualistic structure of exponential families . . . 55
6.2.5 Mutually dual foliations . . . 57
6.2.6 The triangular relation . . . 58
6.3 Statistical inference and differential geometry . . . 58
6.3.1 Estimation based on independent observations . . . 58
6.3.2 Exponential families and observed points . . . 59
6.3.3 Consistency and first-order efficiency . . . 60
6.3.4 Higher-order asymptotic theory of estimation . . . 62

7 Information Geometry and its Applications **64**

8 Graphical Models, Exponential Families, and Variational Inference **65**
8.1 Basics of convex sets and functions . . . 65
8.2 Exponential families . . . 66
8.2.1 Motivation: Maximum Entropy . . . 66
8.2.2 Basics of exponential families . . . 67
8.2.3 Mean parameterization and inference problems . . . 67

9 Sensitivity to initial conditions of multidimensional linear stochastic differential equation **68**
9.1 Multidimensional linear SDE . . . 68
9.2 Geometry of Poincaré half-plane . . . 68
9.3 Geometry of Siegel half-plane from an algebraic perspective . . . 68
9.3.1 Introduction . . . 68
9.3.2 Symplectic and orthogonal groups . . . 68
9.4 Geometry of Siegel half-plane from an algebraic perspective (simplified) . . . 69
9.4.1 Basic definitions and results . . . 69
9.4.2 Symplectic linear algebra . . . 70
9.5 Proof that multidimensional Gaussian laws are the Siegel half-plane . . . 71
9.6 Calculus of variations . . . 71

9.7 Bounds . . . 71

10 Uncertainty in Machine Learning **71**
10.1 Robust Optimization . . . 71
10.2 Bayesian Neural Nets (BNN) . . . 71
10.2.1 Variational Inference (VI) . . . 71
10.2.2 Laplace approximation . . . 72
10.2.3 Stochastic Gradient Markov Chain Monte Carlo . . . 72
10.2.4 MC-Dropout . . . 72
10.2.5 Bayes By Backpropagation (BBP) . . . 72
10.2.6 Preconditioned Stochastic Gradient Langevin Dynamics (p-SGLD) . . . 72
10.3 Non-Bayesian Approaches . . . 72
10.3.1 Threshold . . . 72
10.3.2 Model Ensembling . . . 72
10.3.3 Prior Network (PN) . . . 72
10.3.4 Stochastic Differential Equation Network (SDE-Net) . . . 72
10.3.5 Particle optimization . . . 73
10.3.6 Information Geometry . . . 73

11 Miscellaneous **73**
11.1 Generalization of the Mean Value Theorem . . . 73
11.2 l_p spaces are not Hilbert spaces unless $p = 2$ . . . 74
11.3 Immersed and embedded submanifolds . . . 74
11.4 Pushforward and Pullback . . . 75
11.4.1 Pushforward of Vectors and Vector Fields . . . 75
11.4.2 Pullback of Covectors, Differential Forms, and Functions . . . 75
11.5 The Uncertainty Principle . . . 76
11.5.1 Gabor's Uncertainty Principle . . . 76
11.6 The curvature of a curve in the 3-dimensional Euclidean space . . . 77
11.7 Differential Geometry in Deep Learning . . . 77
11.8 Finally the Truth about Fisher Information? . . . 78
11.9 Is Machine Learning a Pseudoscience? . . . 78
11.9.1 Programme de recherche en IA et critiques de Lê. . . . 78
11.10The Likelihood principle. . . . 80

1 Théorie de l'information et statistiques

Je m'appuie notamment sur le livre de Cover et Thomas [1].

1.1 L'information de Fisher et la borne de Cramér-Rao

Soit $\{f(x,\theta)\}$ une famille de densités de probabilités unidimensionnelles ($x \in \mathcal{S} \subset \mathbb{R}$) paramétrée par un paramètre unidimensionnel ($\theta \in \Theta \subset \mathbb{R}$). Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire i.i.d. tel que $X_1 \sim f(x,\theta)$ où θ est un paramètre inconnu. Soit $\hat{\theta} = t(X)$ un estimateur de θ que l'on suppose non-biaisé i.e., $\text{Biais(t(X))} = \mathbb{E}_\theta[t(X) - \theta] = 0$.

Afin de comparer **l'efficacité** de plusieurs estimateurs non-biaisés, on introduit l'erreur quadratique moyenne $\text{MSE}(t(X)) = \mathbb{E}_\theta[(t(X) - \theta)^2]$. Dans le cas général (estimateur biaisé), on a $\text{MSE}(t(X)) = \text{Biais}(t(X))^2 + \text{Var}(t(X))$. On dira qu'un estimateur $t_1(X)$ est plus efficace que $t_2(X)$ si, pour tout $\theta \in \Theta$, $\text{MSE}(t_1(X)) \leq \text{MSE}(t_2(X))$. La question est de savoir quel est l'estimateur le plus efficace possible.

On introduit le score $v(X,\theta) = \frac{\partial \ln p(X,\theta)}{\partial \theta}$ où $p((X_1, \ldots, X_n), \theta) = \prod_{i=1}^n f(X_i, \theta)$. Le score est la dérivée de la log-vraissemblance de l'échantillon X par rapport au paramètre. On fait l'hypothèse de pouvoir échanger la dérivation par θ avec le signe intégrale. Notez que c'est toujours possible si le support $\mathcal{S}$ ne dépend pas

de θ. En effet, en toute généralité, on a $\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} g(x,\theta)dx = g(b(\theta),\theta)\frac{\partial b(\theta)}{\partial\theta} - g(a(\theta),\theta)\frac{\partial a(\theta)}{\partial\theta} + \int_{a(\theta)}^{b(\theta)} \frac{\partial g(x,\theta)}{\partial\theta}$. Sous cette hypothèse, l'espérance du score est nulle :

$$\begin{aligned}
\mathbb{E}_\theta[v(X,\theta)] &= \int_{\mathcal{S}^n} \frac{\partial \ln p(x,\theta)}{\partial\theta} p(x,\theta)dx, \\
&= \int_{\mathcal{S}^n} \frac{1}{p(x,\theta)} \frac{\partial p(x,\theta)}{\partial\theta} p(x,\theta)dx, \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{S}^n} p(x,\theta)dx, \\
&= \frac{\partial}{\partial\theta}1, \\
&= 0.
\end{aligned}$$

D'après l'inégalité de Cauchy-Schwarz :

$$\mathbb{E}_\theta[(t(X) - \mathbb{E}_\theta[t(X)])(v(X,\theta) - \mathbb{E}_\theta[v(X,\theta)])]^2 \leq \mathbb{E}_\theta[(t(X) - \mathbb{E}_\theta[t(X)])^2]\mathbb{E}_\theta[(v(X,\theta) - \mathbb{E}_\theta[v(X,\theta)])^2].$$

En utilisant le fait que l'estimateur soit non-biaisé et que le score ait une espérance nulle, on obtient :

$$\mathbb{E}_\theta[t(X)v(X,\theta)]^2 \leq \mathrm{Var}[t(X)]\mathrm{Var}[v(X,\theta)].$$

Cependant, on a :

$$\begin{aligned}
\mathbb{E}_\theta[t(X)v(X,\theta)] &= \int_{\mathcal{S}^n} t(x)\frac{\partial \ln p(x,\theta)}{\partial\theta} p(x,\theta)dx, \\
&= \int_{\mathcal{S}^n} t(x)\frac{\partial p(x,\theta)}{\partial\theta} dx, \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{S}^n} t(x)p(x,\theta)dx, \\
&= \frac{\partial}{\partial\theta}\mathbb{E}_\theta[t(X)], \\
&= \frac{\partial}{\partial\theta}\theta, \\
&= 1.
\end{aligned}$$

On a donc :

$$\mathrm{Var}[t(X)] \geq \frac{1}{\mathrm{Var}[v(X,\theta)]}.$$

On note $I(\theta) = \mathrm{Var}[v(X,\theta)]$ la variance du score que l'on appelle **information de Fisher**. On peut donc réécrire l'inégalité précédente :

$$\mathrm{Var}[t(X)] \geq \frac{1}{I(\theta)},$$

que l'on appelle inégalité de Cramér-Rao. On dira qu'un estimateur $t(X)$ est efficace si sa variance est égale à la borne de Cramér-Rao i.e., $\mathrm{Var}[t(X)] = \frac{1}{I(\theta)}$. On dira aussi que $t(X)$ est un MVUE (minimum-variance unbiased estimator). On peut interpréter $I(\theta)$ comme la quantité d'information contenue dans l'échantillon X pour estimer θ. Plus X contient d'information sur θ, plus la variance d'un estimateur efficace (i.e., son erreur moyenne quadratique s'il est non-biaisé) sera proche de zéro. Notez qu'il est possible qu'aucun estimateur efficace n'existe.

Il est possible d'exprimer l'information de Fisher comme une dérivée seconde. En effet :

$$\begin{aligned} I(\theta) &= \int \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx \\ &= \int \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} dx \\ &= \int \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx - \int \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} p(x,\theta) dx \text{ (intégration par parties)} \\ &= -\int \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} p(x,\theta) dx \text{ (l'espérance du score est nulle)} \end{aligned}$$

1.2 Entropie et statistique suffisante

Soit μ et ν deux mesures de probabilités telles que $\nu << \mu$. L'*entropie relative* ou la *divergence de Kullback-Leibler* de ν par rapport à μ est définie par :

$$D(\mu||\nu) = -\mathbb{E}_\mu \left[\ln \frac{d\nu}{d\mu} \right],$$

où $\frac{d\nu}{d\mu}$ est la densité ou dérivée de Radon-Nikodym de ν par rapport à μ.

Dans la suite, on se placera dans le cas de deux variables aléatoires réelles X et Y dont les lois $\mathbb{P}_X$ et $\mathbb{P}_Y$ sont absolument continues par rapport à la mesure de Lebesgue λ. On note p et q les densités de X et Y par rapport à λ. En supposant que p et q ont **le même support**, on a $\mathbb{P}_Y << \mathbb{P}_X$ avec $\frac{d\mathbb{P}_Y}{d\mathbb{P}_X} = \frac{q}{p}$ (et réciproquement). Dans ce cas on notera $D(\mathbb{P}_X||\mathbb{P}_Y) = D(p||q)$ et on a donc :

$$D(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx,$$

qui n'est autre que :

$$D(p||q) = \mathbb{E}_p \left[\ln \frac{p(X)}{q(X)} \right].$$

On a $D(p||q) \geq 0$ et $D(p||q) = 0 \iff p = q$. Cependant, l'entropie relative n'est pas symétrique et ne respecte pas l'inégalité triangulaire. L'entropie relative peut être interprétée comme le nombre moyen de bits[1] supplémentaires qui seront nécessaires si on code des messages selon la loi p alors qu'ils sont en fait tirés selon la loi q.

Soient X et Y deux variables aléatoires de densités marginales $p(x)$ et $p(y)$, et de densité jointe $p(x,y)$. L'information mutuelle de X et Y est :

$$i(p;q) = D(p(x,y)||p(x)p(y)).$$

Il s'agit du nombre moyen de bits redondants qui seront nécessaires pour envoyer des messages selon $p(x,y)$ en supposant que X et Y sont indépendants alors qu'ils ne le sont pas.

L'entropie[2] d'une variable aléatoire X de densité p est définie par :

$$h(p) = -\mathbb{E}_p[\ln p(X)].$$

Il s'agit de (l'opposé de) l'espérance de la log-vraisemblance. L'entropie est parfois appelée "entropie différentielle" pour la différencier de l'entropie définie sur des variables aléatoires discrètes.

On peut maintenant définir l'entropie croisée par :

$$h_p(q) = h(p) + D(p||q),$$

[1] ou de "nats" dans ce cas puisqu'on utilise ln et pas $\log_2$.

[2] A mon grand désespoir, je n'ai pas réussi à définir l'entropie à partir de l'entropie relative ... L'entropie est-elle égale à la "self-information" ? Peut-on dire que $h(p) = i(p;p)$?

qui n'est autre que :

$$h_p(q) = -\mathbb{E}_p[\ln q(X)].$$

Attention à ne pas confondre l'entropie croisée avec l'entropie jointe qui est simplement l'entropie du vecteur aléatoire (X, Y) et que l'on notera $h(p, q)$.

Soit $\{f(x, \Theta)\}$ une famille de densités de probabilités unidimensionnelles ($x \in \mathcal{S} \subset \mathbb{R}$) paramétrée par un paramètre unidimensionnel ($\Theta \in \mathbb{R}$) que l'on suppose constant presque sûrement, $\Theta = \theta$ p.s.. Le support $\mathcal{S}$ est supposé indépendant de θ. Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire i.i.d. tel que $X_1 \sim f(x, \Theta)$ où $\Theta = \theta$ (p.s.) est un paramètre fixé mais inconnu. Soit $u(X)$ une statistique. Remarquez que la loi de $u(X)$ sachant X est indépendante de Θ. Autrement dit, $\Theta \longrightarrow X \longrightarrow u(X)$ forme une chaîne de Markov. On peut alors montrer l'inégalité suivante :

$$i(\Theta; u(X)) \leq i(\Theta; X),$$

qui s'interprète en disant que tout traitement des données X par une fonction U ne peut que diminuer la quantité d'information. Une statistique est dite **suffisante** si il y a égalité i.e., $i(\Theta; u(X)) = i(\Theta; X)$. Dans ce cas, aucune information n'est perdue. On peut aussi dire qu'une statistique est suffisante si et seulement si $\Theta \longrightarrow u(X) \longrightarrow X$ est une chaîne de Markov, c'est à dire que la loi de X est entièrement déterminée par $u(X)$, donc $u(X)$ contient toute l'information de X à propos de la valeur de Θ. Une condition nécessaire et suffisante pour qu'une statistique $u(X)$ soit suffisante est donnée par le théorème de factorisation de Fisher-Neyman : il faut et il suffit de pouvoir factoriser la densité sous la forme $f(x, \Theta) = g(x)k(u(x), \theta)$ où g ne dépend pas de θ et k ne dépend de x qu'à travers $u(x)$.

Maintenant, soit $\theta_0 \in \Theta$ et considérons la fonction $f : \theta \mapsto D(p(\cdot, \theta_0)||p(\cdot, \theta))$ que l'on notera $f(\theta) = D(\theta_0||\theta)$. On a :

$$\frac{\partial}{\partial\theta} f(\theta) = -\int p(x, \theta_0) \frac{\partial \ln p(x, \theta)}{\partial\theta} dx.$$

En $\theta = \theta_0$, la dérivée de l'entropie relative n'est autre que l'opposé de l'espérance du score $\frac{\partial}{\partial\theta} f(\theta_0) = -\mathbb{E}_{\theta_0}[v(X, \theta_0)]$ qui est nulle d'après le paragraphe précédent. Donc :

$$\frac{\partial}{\partial\theta} f(\theta_0) = 0.$$

La dérivée seconde de f est :

$$\frac{\partial^2}{\partial\theta^2} f(\theta) = -\int p(x, \theta_0) \frac{\partial^2 \ln p(x, \theta)}{\partial\theta^2} dx.$$

En $\theta = \theta_0$, on retrouve l'information de Fisher :

$$\frac{\partial^2}{\partial\theta^2} f(\theta_0) = I(\theta_0).$$

Ainsi, le développement de Taylor à l'ordre 2 de l'entropie relative est :

$$D(\theta||\theta + \epsilon) = I(\theta)\epsilon^2 + o(\epsilon^2).$$

1.3 Deux théorèmes

1.3.1 Théorème de Rao-Blackwell

Soit $t(X)$ un estimateur non biaisé de θ. Soit $u(X)$ une statistique suffisante pour θ. Alors $t^*(X) = \mathbb{E}_\theta[t(X)|u(X)]$ (c'est une espérance conditionnelle) est aussi non biaisé et on a $\mathrm{Var}[t^*(X)] \leq \mathrm{Var}[t(X)]$.

$t^*(X)$ est bien un estimateur (ne dépend pas de θ) car $u(X)$ est une statistique suffisante. Le théorème de l'espérance totale montre que $t^*(X)$ est non biaisé (car $t(X)$ est non biaisé). Enfin, l'inégalité de Jensen appliquée à la fonction convexe $x \mapsto x^2$ permet de montrer l'inégalité sur les variances.

Ce théorème permet de construire des estimateurs plus efficaces à partir d'estimateurs peu efficaces en utilisant des statistiques suffisantes.

1.3.2 Théorème de Fisher-Darmois-Pitman-Koopman

On peut se demander à quelles conditions il existe une statistique suffisante de dimension fixe (dont la dimension n'augmente pas quand la taille de l'échantillon augmente). Par exemple, si on cherche à estimer la moyenne d'une loi normale, alors $\frac{1}{n}\sum_i X_i$ est une statistique suffisante qui est toujours de dimension 1 quel que soit n.

Le théorème de Fisher-Darmois-Pitman-Koopman affirme que parmi les familles de distributions à support constant (i.e., le support ne dépend pas des paramètres), seules les familles exponentielles admettent des statistiques suffisantes de dimension constante.

Une famille exponentielle est une famille de distributions de probabilités dont les densités par rapport à une mesure (généralement la mesure de Lebesgue mais pas que, puisqu'avec la mesure de comptage, on obtient des lois discrètes) sont de la forme :

$$\exp(\langle x, \theta\rangle - \psi(\theta))$$

où ψ est une fonction de normalisation permettant d'avoir une probabilité égale à 1 lorsqu'on intègre sur le domaine de x. Les paramètres θ sont appelés "paramètres canoniques". Les familles exponentielles regroupent de nombreuses familles de distributions classiques : gaussiennes, gammas, bêtas, exponentielles, de Poisson ...

1.3.3 Aparté

L'intérêt de la géométrie de l'information repose sur un résultat pour les familles exponentielles. L'espace des θ noté Θ peut être muni de la métrique de l'information de Fisher g et d'une connexion euclidienne (plate) ∇_m qu'on appelle connexion de mélange car $\langle x, \theta\rangle$ est un mélange dans le vocabulaire probabiliste. Le triplet (Θ, g, ∇_m) est appelé un espace de jauge. On note X la variable aléatoire de paramètre θ et on considère $\eta(\theta) = \partial_\theta \mathbb{E}[X]$. Un résultat fondamental de la géométrie de l'information nous dit que η est un difféomorphisme. On obtient alors de nouvelles coordonnées appelées "paramètres naturels". Par exemple, si la famille est celle des gaussiennes unidimensionnelles, on a $\eta = (\mu, \sigma^2)$. Si on considère ∇_e la connexion euclidienne sur les η alors ∇_m et ∇_e sont des connexions conjuguées ou duales. ∇_e est appelé la connexion d'espérance car elle est obtenue par un calcul d'espérance. La théorie des connexions duales nous dit que la connexion de Levi-Civita s'obtient par $\nabla^{LC} = (\nabla_m + \nabla_e)/2$.

2 Integration on a manifold

In this section, I want to answer four questions.

- Why are differential forms (of top degree) the correct object to integrate over a manifold?
- What is the impact of the orientation on integration? Is it only the sign of the integral? In particular, is it possible to define the absolute value of an integral over a non-orientable manifold (but not its sign), or is it impossible to even define integration over a non-orientable manifold?
- Is it possible to integrate of top form with non-compact support?
- What is the equivalent of Fubini's theorem for differential forms? It seems that the antisymmetry of the wedge product is canceled out by the change in orientation in the domain of integration (when we flipped the two integral signs).

2.1 Why differential forms?

Let $\mathcal{M}$ be a differentiable manifold of dimension n. Let $(U, \phi = (x^1, \dots, x^n))$ be a chart. Let $f : U \to \mathbb{R}$ be a function. Assume that f is sufficiently regular, for example it is continuous with compact support. Recall the definition of the support.

Definition 2.1 (Support)**.** The support of a function $f : U \to \mathbb{R}$ is the closure in its domain (i.e., in U) of the subset where f is not zero: $\mathrm{supp} f := \mathrm{cl}_U(\{x \in U : f(x) \neq 0\})$.

We would like to define the integral of f over U. As for many things in differential geometry, we would like to use the chart to move to $\mathbb{R}^n$, use a definition that already exists on $\mathbb{R}^n$, then pull it back to $\mathcal{M}$. When we are on $\mathbb{R}^n$, we use the canonical coordinates, but there are no such canonical coordinates on $\mathcal{M}$, so our definition must be independent of the choice of a chart.

Let $U' = \phi(U) \subset \mathbb{R}^n$. Consider the function $g : U' \subset \mathbb{R}^n \to \mathbb{R}$ defined by $g = f \circ \phi^{-1}$ (which is the pullback $(\phi^{-1})^* f$ but we assume that we don't know the pullback). The integral of g can be defined using the Riemann integral. Intuitively, the Riemann integral is defined by approximating the function with rectangles where the function has constant value. If the upper and lower approximations coincide, then the Riemann integral is defined to be this common value. Let $R = [a^1, b^1] \times \ldots \times [a^n, b^n]$ be a closed rectangle of $\mathbb{R}^n$. Its volume is $\text{vol}(R) = \prod_{i=1}^n (b^i - a^i)$. We can enclose U' into a rectangle R and extend g to be zero on $R \setminus U'$. Then, we can partition each interval of R such that R is divided into closed subrectangles R_j dependent on the chosen partition P. Define the lower and upper sums $l(g, P) = \sum \inf_{R_j}(g)\text{vol}(R_j)$ and $u(g, P) = \sum \sup_{R_j}(g)\text{vol}(R_j)$. If $\sup_P(l(g, P)) = \inf_P(u(g, P))$ then g is Riemann integrable (which is the case thanks to our regularity assumptions) and we define:

Definition 2.2 (Riemann integral)**.**

$$\int_{U'} g := \sup_P(l(g, P)) = \inf_P(u(g, P)).$$

Now, we are tempted to define the integral of f by $\int_U f = \int_{U'} g = \int_{\phi(U)} f \circ \phi^{-1}$. But is this definition independent of the choice of coordinates? Consider another chart $(U, \psi = (y^1, \ldots, y^n))$ on the same coordinate open set U but with a different coordinate map. Let $V' = \psi(U)$ and $F : V' \to U'$ defined by $F = \phi \circ \psi^{-1}$. Since $\mathcal{M}$ is a differentiable manifold, F is a diffeomorphism. According to our definition, $\int_U f = \int_{\psi(U)} f \circ \psi^{-1} = \int_{V'} f \circ \phi^{-1} \circ \phi \circ \psi^{-1} = \int_{V'} g \circ F$. For our definition to be consistent, we need to have $\int_{U'} g = \int_{V'} g \circ F$ but this is not true! This is why it is impossible to define the integral of a function over a manifold (i.e., in a coordinate independent way).

So, what kind of objects can we integrate over a manifold? The answer to this question lies in the theorem of change of variables.

Theorem 2.3 (Change of variables)**.** *Let U' and V' be two open subsets of $\mathbb{R}^n$. Let $g : U' \to \mathbb{R}$ be continuous with compact support. Let $F : V' \to U'$ be a diffeomorphism. Then:*

$$\int_{U'} g = \int_{V'} (g \circ F)|\det J_F|,$$

where $J_F = \left(\frac{\partial(x^i \circ F)}{\partial y^j}\right)_{i,j}$ is the Jacobian matrix of F.

I won't provide a rigorous proof of this result since it is overly technical. I will simply provide an intuition. Everything rests on the following fact.

Fact 2.4. *Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible linear map and R a rectangle of $\mathbb{R}^n$.*
Then $\text{vol}(A(R)) = |\det A|\text{vol}(R)$.

Proof. First, we consider the case of the unit cube $C = [0, 1] \times \ldots \times [0, 1]$. We have $\text{vol}(C) = 1$. We want to show that $\text{vol}(A(C)) = |\det A|$. We identify A with its matrix in the canonical basis (we know that the matrix has the same determinant as its corresponding linear map). The columns of A are the coordinates in the canonical basis of the images by A of the canonical basis vectors. Denote these columns by $v_1, \ldots, v_n$. The volume $\text{vol}(A(C))$ is the volume of the parallelogram spanned by $v_1, \ldots, v_n$. In order to prove that this volume is $|\det A|$, we exhibit several properties of the volume and show that the (absolute value of the) determinant is the only function that satisfies these properties.

- **The volume is invariant by adding to any column a linear combination of the other columns**. For example, we have $\text{vol}(v_1 + \sum_{i>1} \lambda_i v_i, v_2, \ldots, v_n) = \text{vol}(v_1, v_2, \ldots, v_n)$. To show this, recall that the volume of a parallelogram is the product of the basis and the height. The basis is the volume spanned by $(v_2, \ldots, v_n)$. The height is the orthogonal distance between v_1 and the subspace spanned by $(v_2, \ldots, v_n)$. If we add $\sum_{i>1} \lambda_i v_i$ to v_1, we move in the subspace parallel to the basis, hence the height does not change. Obviously, the basis does not change either, so the volume is invariant. The determinant verifies this property thanks to the multilinearity and the alternate property.

- **If we scale a vector by a constant, then the volume is scaled by the absolute value of this constant**. For example $\text{vol}(\lambda v_1, v_2, \ldots, v_n) = |\lambda|\text{vol}(v_1, v_2, \ldots, v_n)$. If $\lambda < 0$ the orientation changes, but the scaling effect on the volume is similar to the case $\lambda > 0$, so we can assume that $\lambda > 0$. If v_1 is replaced by λv_1 then the height is also sclaled by λ. The basis does not change, so the volume is scaled by λ. The determinant verifies this property thanks to the multilinearity.
- **If we interchange two vectors, the volume does not change**. For example $\text{vol}(v_2, v_1, \ldots, v_n) = \text{vol}(v_1, v_2, \ldots, v_n)$. By swapping two vectors, we obtain the same parallelogram, so the volume is the same. If we interchange two vectors, the determinant is multiplied by -1 due to the antisymmetry property. Because of the absolute value, $|\det A|$ will not change, and thus will verify the property.
- The three previous property show that we can use any antisymmetric multilinear function in the absolute value. This is a very restricted set of functions but still infinite. In order to select a unique function, we need to remember that the determinant is the unique multilinear antisymmetric function that is equal to 1 when A is the identity. If A is the identity then $\text{vol}(A(C)) = \text{vol}(C) = 1$, hence the only possible function for the volume is $|\det A|$.

Now, consider an arbitrary rectangle R. There exist two vectors a and b such that $R = aC + b$ where the i-th vector of C is scaled by the i-th component of a. The volume is invariant by translation and the second property of the previous list tells us that $\text{vol}(R) = \prod_i a_i \text{vol}(C) = \prod_i a_i$. By linearity, $A(R) = aA(C) + b$, hence $\text{vol}(A(R)) = \prod_i a_i \text{vol}(A(C)) = \prod_i a_i |\det A| = |\det A|\text{vol}(R)$. □

Note that Fact 2.4 is a special case of Theorem 2.3 when $V' = R$, $U' = A(R)$, $F = A$, $g \equiv 1$. The general case relies on the differentiability of F (i.e., F is locally linear).

Assume we have a integrable function g. Intuitively, the integral of g over U' is the sum of the volume of "infinitesimal" rectangles multiplied by the value of g over theses rectangles (the rectangles are so small that g is constant over each one of them). If R_x is an infinitesimal rectangle at x, then

$$\int_{U'} g := \sum_{x \in U'} g(x)\text{vol}(R_x).$$

A more rigorous statement would say that this is true in the limit, when the volumes of the R_x's go to zero. We can take any value of g over R_x in place of $g(x)$.

Now, we make the "change of variables" $x = F(y)$ and we want to compute the integral of g using rectangles along the y-coordinates instead of x-coordinates. That means that we want to sum over $y \in V'$ instead of $x \in U'$, and multiply the "appropriate value of g" by $\text{vol}(R_y)$ instead of $\text{vol}(R_x)$. Let R_y be a rectangle in y-coordinates. The volume is invariant by translation, thus $\text{vol}(F(R_y)) = \text{vol}(F(R_y) - F(y))$, where we use the notation $S - p_0 = \{p - p_0 : p \in S\}$. The rectangle R_y is so small that we can use the first order Taylor approximation $F(R_y) - F(y) = J_F(y)(R_y - y)$ where $J_F(y)$ is the differential of F at y (that we identify with the Jacobian matrix of F at y). Hence

$$\begin{aligned}
\int_{U'} g &= \sum_{y \in V'} g(F(y))\text{vol}(F(R_y)), \\
&= \sum_{y \in V'} g(F(y))\text{vol}(J_F(y)(R_y - y)), && (\text{since } \text{vol}(F(R_y)) \approx \text{vol}(J_F(y)(R_y - y)) \text{ when } R_y \to 0) \\
&= \sum_{y \in V'} g(F(y))|\det J_F(y)|\text{vol}(R_y - y), && (\text{Fact 2.4}) \\
&= \sum_{y \in V'} g(F(y))|\det J_F(y)|\text{vol}(R_y), && (\text{the volume is invariant by translation}) \\
&= \int_{V'} (g \circ F)|\det J_F|, && (\text{definition of the Riemann integral})
\end{aligned}$$

which is Theorem 2.3.

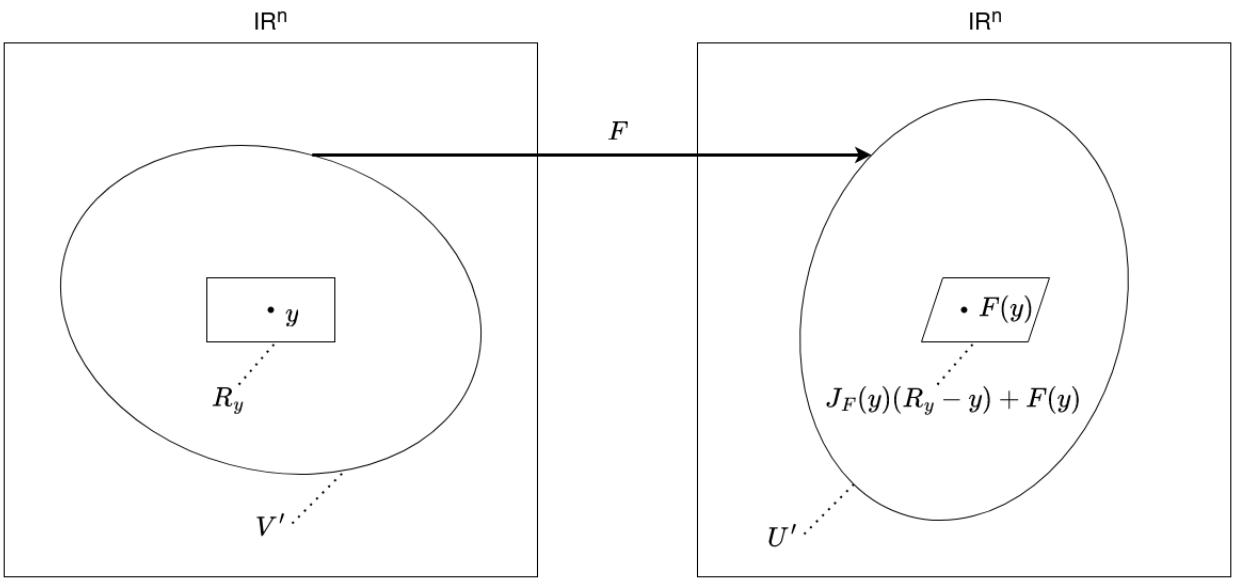

Figure 1: Change of variables

A question that may arise is: why have we used $\text{vol}(F(R_y))$ instead of $\text{vol}(R_y)$ is the first equality? Please look at Figure 1. To answer this question, we need to think of x and y as living on different copies of $\mathbb{R}^n$. Let us call them $\mathbb{R}^n_x$ and $\mathbb{R}^n_y$.

What we want is to compute the integral of g in $\mathbb{R}^n_x$ using the variable y living on $\mathbb{R}^n_y$, but what we do *not* want is to compute the integral of $g \circ F$ in $\mathbb{R}^n_y$. Why are these two methods not identical? This is because there is no reason for g to be spread across the volume of U' in $\mathbb{R}^n_x$ in the same way $g \circ F$ is spread over the volume of V'. The question we are interested in is: if g is a mass density over U', what is the total mass of U'. But we are not interested in: if $g \circ F$ is a mass density over V', what is the total mass of V'? We want to know the mass of U', not the mass of V'. There is no reason for them to have the same mass, even if their densities both rely on g. This is because F has no reason to preserve the volume.

The map F may send a small volume in V' to a large volume in U', such that this volume in U' will have a higher mass than its corresponding volume in V' (since the same mass density will be spread over a larger volume in U' than in V'). The absolute value of the Jacobian is here to correct these deformations of the volume by F. We use the absolute value of the Jacobian (and not the Jacobian itself) because we are interested in the (non-signed) volume, which is a nonnegative quantity, and we do not care if F reverses the orientation or not.

Note that the integral of g over U' *is* equal to the integral of $g \circ F$ over V' if and only if F is a volume-preserving map, because the Jacobian of a volume-preserving map is ± 1. For example, it is true if F is an isometry (but this is not necessary).

Now, we are in position to say what kind of object we can integrate on a manifold. Let ω be such an object defined on U. Let $p \in \mathcal{M}$ and let $x = \phi(p)$ and $y = \psi(p)$ such that $x = F(y)$. Denote by $f_\phi(p)$ (resp. $f_\psi(p)$) the representation of ω_p in the chart (U, ϕ) (resp. (U, ψ)) at p. We would like to have $f_\phi(p) = |\det J_F(y)| f_\psi(p)$.

The map $(v_1, \ldots, v_n) \mapsto \det(v_1, \ldots, v_n)$ is multilinear and antisymmetric. In particular, the map $\det J_F(y)$ is multilinear and antisymmetric in the columns of $J_F(y)$, but it is not the case of $|\det J_F(y)|$. Thus, we would rather have $f_\phi(p) = (\det J_F(y)) f_\psi(p)$ to take advantage of the multilinear and antisymmetric properties of the determinant. The absolute value will be dealt with using the notion of orientation.

First, we consider the simpler case of defining a notion of volume over the manifold $\mathcal{M}$. This will serve as a basis for the general case. The notion of volume is dependent on the choice of a coordinate system (or coordinate atlas is the set spans across several charts).

Definition 2.5 (Volume)**.** Let (U, ϕ) be a chart. Let S be a subset of U such that $\phi(S)$ is bounded and its

indicator function $\mathbb{1}_S$ is integrable on $\mathbb{R}^n$.
The volume of S in the chart (U, ϕ) is

$$\mathrm{vol}_\phi(S) = \int_{\mathbb{R}^n} \mathbb{1}_{\phi(S)} = \int_{\phi(S)} 1.$$

The object ω^ϕ used as integrand to compute the volume according to the chart (U, ϕ) verifies the following relation: $\int_U \omega^\phi = \int_{\phi(U)} f_\phi \circ \phi^{-1} = \int_{\phi(U)} 1 = \mathrm{vol}_\phi(U)$. That means that the representation of ω^ϕ in the chart (U, ϕ) is $f_\phi \equiv 1$. According to the theorem of change of variables, we have $\int_{\phi(U)} 1 = \int_{\psi(U)} |\det J_F|$. Moreover, assume that F has been chosen such that its Jacobian is everywhere positive (i.e., we are on an oriented atlas). Then $\int_{\phi(U)} 1 = \int_{\psi(U)} \det J_F$, which means that $f_\psi \circ \psi^{-1} = \det J_F$. Once again, consider a point $p \in \mathcal{M}$ such that $x = \phi(p)$, $y = \psi(p)$ and $x = F(y)$. We have $f_\phi(p) = 1$ and $f_\psi(p) = \det J_F(\psi(p))$. This two last equations means that when we feed ω_p^ϕ with a chart, it spits a real numbers which depends multilinearly and antisymmetrically on the columns of the Jacobian matrix of the transition map at p from the given chart to ϕ. Since the destination chart is already given (it is ϕ for ω^ϕ), we don't need the entire Jacobian matrix but *only the partial derivative operators along the components of the starting chart* ψ. This works nicely because the directional derivatives at a point p form a n-dimensional vector space called the *tangent space* $T_p\mathcal{M}$. If $\psi = (y^1, \ldots, y^n)$ is a coordinate map at p, the set $\{\partial/\partial y^1|_p, \ldots, \partial/\partial y^n|_p\}$ of partial derivative operators along the components of ψ is a basis of $T_p\mathcal{M}$. Hence, ω_p^ϕ **is a n-linear antisymmetric form on** $T_p\mathcal{M}$. The object ω^ϕ that assigns a n-linear antisymmetric form on $T_p\mathcal{M}$ for each $p \in U$ is called a ***differential form*** on U. More precisely, this is the *volume form* associated with the chart (U, ϕ).

Since the n-linear antisymmetric forms on $T_p\mathcal{M}$ constitute a 1-dimensional vector space called $A_n(T_p\mathcal{M})$, we can use the image at p of any volume form as a basis on $A_n(T_p\mathcal{M})$ (note that a volume form is non-vanishing). Given a volume form (i.e., a basis of $A_n(T_p\mathcal{M})$ for every $p \in U$), we can integrate any function which can be interpreted as a mass density *corresponding to the chosen volume form.* If we choose another chart (in the same oriented atlas), then the volume form will spit a Jacobian which will modify the mass density to get another function whose integral has the same value. The object that can be integrated on a manifold is thus a differential form that can be written as $\omega = f\omega^\phi$ for some function f in the basis $\{\omega^\phi\}$. If we choose another coordinate system ψ, then $\omega = f\omega^\phi(\partial/\partial y^1, \ldots, \partial/\partial y^n)\omega^\psi = f \cdot (\det J_F \circ \psi)\omega^\psi$. This gives us the image of a volume form over any basis of vector fields induced by a chart: $\omega^\phi(\partial/\partial y^1, \ldots, \partial/\partial y^n) = \det J_F \circ \psi$ which is a function from $\mathcal{M}$ to $\mathbb{R}$.

We have already implicitly defined the integral of a volume form by $\int_U \omega^\phi = \int_{\phi(U)} 1 = \mathrm{vol}(\phi(U))$. Now, we give the definition using an arbitrary chart. Let $(U, \psi = (y^1, \ldots, y^n))$ be a chart in the same oriented atlas as (U, ϕ). Then $\int_U \omega^\phi := \int_{\psi(U)} \omega^\phi(\partial/\partial y^1, \ldots, \partial/\partial y^n) \circ \psi^{-1} = \int_{\psi(U)} \det J_F$ where F is the transition map from ψ to ϕ. If $\psi = \phi$, the transition map is the identity so the Jacobian is identically equal to 1, and we recover the intuitive definition of the integral of a volume form.

Now, we can define the integration of arbitrary differential forms.

Definition 2.6 (Integration of differential forms)**.** Let ω be a differential form on a open subset U of $\mathcal{M}$ and let $(U, \phi = (x^1, \ldots, x^n))$ be a chart.

$$\int_U \omega := \int_{\phi(U)} f_{(\phi^{-1})^*\omega},$$

where $f_{(\phi^{-1})^*\omega} : \mathbb{R}^n \to \mathbb{R}$ is defined for any $t = (t^1, \ldots, t^n) \in \mathbb{R}^n$ by

$$\begin{aligned} f_{(\phi^{-1})^*\omega}(t) &:= ((\phi^{-1})^*\omega)_t \left(\left.\frac{\partial}{\partial r^1}\right|_t, \ldots, \left.\frac{\partial}{\partial r^n}\right|_t \right), \\ &= \omega_{\phi^{-1}(t)} \left(\phi_*^{-1} \left.\frac{\partial}{\partial r^1}\right|_t, \ldots, \phi_*^{-1} \left.\frac{\partial}{\partial r^n}\right|_t \right), \\ &= \omega_{\phi^{-1}(t)} \left(\left.\frac{\partial}{\partial x^1}\right|_{\phi^{-1}(t)}, \ldots, \left.\frac{\partial}{\partial x^n}\right|_{\phi^{-1}(t)} \right). \end{aligned}$$

Here, we have implicitly defined two new operators:

- The pushforward of a smooth map between manifold, also called the total derivative, that sends vector fields from one manifold to vector fields of another manifold. Given a smooth map F, a vector field X, and a function f, it is formally defined as $(F_*X)f := X(f \circ F)$. For our purpose, it is simply used to define the directional derivative operator induced by a chart: $\partial f/\partial x^i := \partial(f \circ \phi^{-1})/\partial r^i = (\phi_*^{-1}(\partial/\partial r^i))f$.
- The pullback of a smooth map between manifold, that sends differential forms from one manifold to differential forms of another manifold. It is defined using the pushforward. Given a smooth map F, a n-form ω, n vector fields $X^1, \dots, X^n$, and a point p, we have $(F^*\omega)_p(X_p^1, \dots, X_p^n) := \omega_p(F_*X_p^1, \dots, F_*X_p^n)$.

The definitions of the pushforward and the pullback may seem poorly motivated. However, one may convinced himself that there are natural ways of defining new vector fields or differential forms by using an existing object on one manifold and transporting it to another manifold with a map. For our purpose, recall that an arbitrary form can be written $\omega = f\omega^\psi$ using some function f and a volume form basis. Given a chart $(U, \phi = (x^1, \dots, x^n))$, we have $\omega = f\omega^\psi(\partial/\partial x^1, \dots, \partial/\partial x^n)\omega^\phi = \omega(\partial/\partial x^1, \dots, \partial/\partial x^n)\omega^\phi$. Thus, similarly to the definition of the integral of a volume form, the integrand over $\mathbb{R}^n$ is $\omega(\partial/\partial x^1, \dots, \partial/\partial x^n) \circ \phi^{-1}$ which is precisely $f_{(\phi^{-1})^*\omega}$.

Strictly speaking, in Definition 2.6, we are integrating a n-form on $\mathbb{R}^n$ which is the pullback $(\phi^{-1})^*\omega$. But the integration of a n-form on $\mathbb{R}^n$ is defined as the integration of the function representing the n-form in the basis of the canonical volume form (the volume form induced by the canonical coordinates), so a n-form on $\mathbb{R}^n$ corresponds canonically to a unique function which we denoted $f_{(\phi^{-1})^*\omega}$. This is only possible in $\mathbb{R}^n$ because $\mathbb{R}^n$ has a canonical coordinate system.

For simplicity, we have defined the integral on a single coordinate chart but the definition can easily be extended to a differential form defined across several charts using a partition of unity (and showing that the definition does not depend on the choice of a partition of unity). This is where the assumption that ω has compact support is needed.

2.2 k-forms, wedge product, exterior derivative

In the previous paragraph, we show that a n-form ω is a map that assigns a volume density to each point p. This volume density depends on the choice of coordinate system. For example, if we choose the coordinates $(x^1, \dots, x^n)$, then ω assigns the volume density $\omega_p(\partial/\partial x^1|_p, \dots, \partial/\partial x^n|_p)$ at p. When we change the coordinates, the volume density changes in such a way that the integration of this density gives the same value. Note that the volume density can be negative or zero.

2.2.1 The wedge product

Now, let $0 \leq k < n$. Similarly to a n-form, a k-form assigns a "hypersurface" density to each point p such that the integration over any k-dimensional hypersurface does not depend on the choice of coordinates. A 0-form is a function. In other words, functions assign a density to a zero-dimensional space i.e., a point. Let us consider the more interesting case of 1-forms. A 1-dimensional space is a curve. In a n-dimensional manifold, there are n independent directions for a curve. So, the line density of a 1-form can be decomposed as a sum of line densities over each independent directions. If $\phi = (x^1, \dots, x^n)$ is a coordinate system, then $(dx^1, \dots, dx^n)$ are a basis for 1-forms. The 1-form dx^i is defined as assigning a line density of 1 at any point in the x^i direction, and 0 in all other directions, *in the ϕ coordinate system*. Formally, we have

$$dx_p^i\left(\left.\frac{\partial}{\partial x^i}\right|_p\right) := 1$$
$$dx_p^i\left(\left.\frac{\partial}{\partial x^j}\right|_p\right) := 0 \text{ for } j \neq i.$$

An arbitrary 1-form ω can be decomposed as $\omega = a_i dx^i$ where $a_i : \mathcal{M} \to \mathbb{R}$ are functions (which are smooth is ω is smooth). The value $a_i(p)$ is the line density that ω assign to p in the x^i direction and in the ϕ coordinate system.

Now we consider 2-forms. Similarly to 1-forms, we would like to decompose the surface density of 2-forms along independent "surface directions". The question is : how much independent surface directions do we need to span all possible 2-forms and how can we obtain them? Given two different line directions dx^i and dx^j, we can span one surface direction. The order does not matter i.e., the surface spanned by dx^i and dx^j is the same as the surface spanned by dx^j and dx^i. So, we have $\binom{n}{2}$ independent surface directions. Let us denote this basis by $\{dx^i \wedge dy^j\}_{1\leq i<j\leq n}$. For $i < j$, we would like to define the 2-form $dx^i \wedge dx^j$ similarly as for 1-forms, i.e., a surface density of 1 along the surface spanned by x^i and x^j and 0 in all other directions (in the coordinates ϕ). We can see $dx^i \wedge dx^j$ as the (signed) area (in the coordinates ϕ) of the small parallelogram spanned by dx^i and dx^j. In order to respect the multilinear alternate transformation rule when we change the coordinate system (i.e., $(dx^i \wedge dx^j)_p$ is a bilinear alternate form), we must have $dx^i \wedge dx^j = -dx^j \wedge dx^i$ and of course $dx^i \wedge dx^i = 0$. This construction can be extended to any $k \leq n$: we obtain $\binom{n}{k}$ independent "hypersurface" directions which are given in the coordinate system ϕ by $\{dx^{i_1} \wedge \ldots \wedge dx^{i_k}\}_{1\leq i_1<\cdots<i_k\leq n}$. Once again, we want to preserve the multilinear and alternate property, so this leads naturally to the definition of the wedge product :

Definition 2.7 (Wedge Product)**.** Let ω be a k-form and τ be a l-form. Let $p \in \mathcal{M}$ and let $X_1, \ldots, X_{k+l}$ be tangent vectors at p. Then :

$$(\omega \wedge \tau)_p(X_1, \ldots X_{k+l}) = \frac{1}{k+l} \sum_{\sigma \in \mathfrak{S}_{k+l}} (-1)^{\epsilon(\sigma)} \omega(X_{\sigma(1)}, \ldots X_{\sigma(k)}) \tau(X_{\sigma(k+1)}, \ldots X_{\sigma(k+l)}).$$

2.2.2 The exterior derivative

We want to define a differential operator that takes a k-form and spits a $k+1$-form. Why would that make any sense? If ω is a smooth k-form, and given a coordinate system $\phi = (x^1, \ldots, x^n)$, we have $\omega = a_I dx^I$ where I is a multi-index $I = (i_1, \ldots, i_k)$ with $1 \leq i_1 < \cdots < i_k \leq n$ and $dx^I = dx^{i_1} \wedge \ldots \wedge d^{i_k}$. The smooth function a_I gives the density at each point over the hypersurface direction dx^I. However, since a_I is a function, this density can change when the point changes. So we can compute the partial derivative of each coordinate functions in each direction $\partial a_I / \partial x^j$.

Fix a multi-index I and consider one component $\tau = a dx^I$. The partial derivatives of a provides n functions associated with the n independent directions $x^1, \ldots, x^n$. So, we can naturally define a 1-form $da = \partial a / \partial x^i dx^i$. If the function a assigns a certain mass in kilograms at each point, then we can use the mass difference between two nearby points to define a line density (in kg/m) along the line connecting these two points. As the distance between the points tends to zero, the line density tends to the directional derivative of a along this direction. Now, coming back to the elementary k-form $\tau = a dx^I$, let dx^j be a line direction that does not appear in dx^I. Let p and $q =$ "$p + \epsilon dx^j$" be two points along the x^j direction. At point p, the density assigned by τ in the ϕ coordinates is $a(p)$ kg/m^k along the dx^I hypersurface. At point q, it is $a(q)$ kg/m^k along the same hypersurface. Hence, we can define a density in the volume $dx^j_p \wedge dx^I_p$ as $(a(q) - a(p))/\epsilon$ kg/m^{k+1}. As $\epsilon \to 0$, we obtain the following $k+1$-form : $\partial a / \partial x^j|_p dx^j_p \wedge dx^I_p$ (we are not using the summation convention here). We can do the same thing with every directions dx^j, even if the direction is in dx^I because in this case, we get $dx^j_p \wedge dx^I_p = 0$ (the $k+1$-parallelogram is "flat"). By summing the contributions of all partial derivatives $\partial a / \partial x^j$, we are motivated to define $d\tau = \sum_j \partial a / \partial x^j|_p dx^j_p \wedge dx^I_p$.

Now, we can consider the general case $\omega = a_I dx^I$ by summing the contributions of each component:

$$d\omega_p = \left.\frac{\partial a^I}{\partial x^j}\right|_p dx^j_p \wedge dx^I_p.$$

La propriété importante de la dérivée extérieure est $d^2 = 0$. Malheureusement, je n'arrive pas (pour l'instant) à trouver l'intuition derrière cette propriété. J'ai lu que cela signifiait que "the boundary of a boundary is empty". Cependant, d permet de passer du "bord" à la "variation dans le volume délimité par le bord", ce qui est l'inverse de cette intuition ... Une autre intuition, qui est assez facile à vérifier par le calcul, est que $d^2 = 0$ est une reformulation du théorème de Schwarz, qui affirme que la dérivée seconde est symétrique. Cependant, c'est une intuition géométrique que je recherche.

2.3 Stokes' theorem

Assume that $\mathcal{M}$ is an n-dimensional orientable manifold (such that the integration is well defined) and that $\mathcal{M}$ has a boundary denoted $\partial\mathcal{M}$ with the induced orientation (such that integration is well defined on $\partial\mathcal{M}$). Let ω be a smooth $n-1$ form defined on $\mathcal{M}$ with compact support (such that ω is integrable on $\partial\mathcal{M}$).

In order to give an intuitive explanation of Stokes' theorem, we illustrate a simple case. Let $\phi = (x^1, \dots, x^n)$ be a global coordinate system. We choose the orientation defined by $dx^1 \wedge \dots \wedge dx^n$. Assume that $\omega = a\, dx^2 \wedge \dots \wedge dx^n$ (it only charges one hypersurface direction) where a is an increasing function of x^1 and constant in every other directions. For example $\omega = x^1\, dx^2 \wedge \dots \wedge dx^n$. Assume that $\phi(\partial\mathcal{M})$ is a $n-1$-sphere. The "hypersurfaces" of ω are straight along $dx^2 \wedge \dots \wedge dx^n$. If we move along x^1 in the increasing direction, we enter $\mathcal{M}$ by crossing $\partial\mathcal{M}$ at a point p_0. The orientation of $\partial\mathcal{M}$ is opposed to the orientation of $dx^2 \wedge \dots \wedge dx^n$ at p_0 so we *subtract* the value $a(p_0)$ in the integral (make a drawing). More formally, if $\psi = (\theta^1, \dots, \theta^{n-1})$ is a coordinate system on $\partial\mathcal{M}$ with the induced orientation, then

$$dx^2_{p_0} \wedge \dots \wedge dx^n_{p_0} \left(\left.\frac{\partial}{\partial\theta^1}\right|_{p_0}, \dots, \left.\frac{\partial}{\partial\theta^{n-1}}\right|_{p_0} \right) < 0.$$

If we continue moving along x^1, we exit $\mathcal{M}$ by crossing $\partial\mathcal{M}$ at a point p_1. This time, the orientation of $\partial\mathcal{M}$ and of $dx^2 \wedge \dots \wedge dx^n$ are aligned at p_1. So we *add* the value of $a(q)$ in the integral. Since a is increasing along x^1 by assumption, we have $a(p_1) - a(p_0) > 0$. The integral $\int_{\partial\mathcal{M}} \omega$ can be seen as the "sum" of all the differences $a(p_1) - a(p_0)$ for every pair of opposed points.

Now, we have $d\omega = \partial a/\partial x^1\, dx^1 \wedge dx^2 \wedge \dots \wedge dx^n$. Hence, the integral $\int_{\mathcal{M}} d\omega$ is the "sum" over all small volumes $dx^1 \wedge \dots \wedge dx^n$ of the variation of a along x^1 which is the partial derivative $\partial a/\partial x^1$. This partial derivative is positive since a is increasing. Moreover, since the orientation of $\mathcal{M}$ is given by $dx^1 \wedge \dots \wedge dx^n$, the value of $\partial a/\partial x^1$ is *added* at every point of $\mathcal{M}$.

Finally, we can see that it is almost tautological to say that the variation $a(p_1) - a(p_0)$ of a along a line parallel to x^1 is equal to the sum of all small variations $\partial a(p)/\partial x^1$ of a for all p along the same line. This is just the fundamental theorem of calculus along x^1:

$$a(p_1) - a(p_0) = \int_{x^1(p_0)}^{x^1(p_1)} \frac{\partial a(p)}{\partial x^1}(x^1(p) - x^1(p_0))dx^1(p).$$

Hence the Stokes' theorem

$$\int_{\partial\mathcal{M}} \omega = \int_{\mathcal{M}} d\omega.$$

If a is not monotonic, this is still the fundamental theorem of calculus along x^1, and we can convince ourselves that any increase in a that rises above $a(p_1)$ is compensated by a similar decrease. If a depends on all variables $x^1, \dots, x^n$, then each line along x^1 corresponds to a different function $x^1 \mapsto a(x^1, x^2, \dots, x^n)$ where $x^2, \dots x^n$ have been fixed and the same reasoning applies independently to each line. Finally, in the general case, the $n-1$-form ω is a sum of $a_I dx^I$ for all dx^I. By linearity of the integral, we can apply our reasoning on each component then sum all integrals to obtain Stokes' theorem. Our intuition is not fully general, because we assumed that $\phi(\partial\mathcal{M})$ is a $n-1$-sphere in $\mathbb{R}^n$. I guess that if $\partial\mathcal{M}$ is compact and simply connected, it is always possible to find a coordinate system ϕ such that $\phi(\partial\mathcal{M})$ is a $n-1$-sphere, but I do not know how to prove it. If $\mathcal{M}$ is compact but not simply connected, I think we can still apply our reasoning on each connected component, and for each connected component, on each line between holes. If $\mathcal{M}$ is not compact, then since ω has compact support, we can restrict ourselves to a compact subset of $\mathcal{M}$.

Intuition for: orientation, closed/exact forms, de Rham cohomology, Poincaré Lemma, interior multiplication and its duality with exterior derivative and Lie derivative (Cartan magic formula), vector bundles, orthogonal complement bundle.

3 Integration theorems of distributions

In this section, I want to clarify what are the conditions for a distribution to be integrable. If the distribution has constant rank, i.e., all the tangent subspaces have the same dimension, then Frobenius Theorem tells

us that the distribution is integrable if and only if it is involutive. I want to get an intuition of why the Frobenius Theorem is true. In particular, I need to understand the intuition behind the Lie derivative of vector fields, and also the various formulations of Frobenius theorem, in particular in terms of 1-forms. I use these references: Lee [2], Molino [3], Lavau [4].

3.1 Differential Equations

Before delving into the core of our subject, we must ensure that our foundations are firm. All the topics of integral curves, flows, distributions, foliations etc., arise from the topic of ordinary differential equations (ODEs). Thus, I review this topic in this paragraph.

Theorem 3.1 (Fundamental Theorem for Autonomous ODEs)**.** *Let $U \subset \mathbb{R}^n$ be open and $V : U \to \mathbb{R}^n$ be a smooth vector-valued function. Let $t_0 \in \mathbb{R}$ and $c = (c^1, \ldots, c^n) \in U$. Consider the initial value problem*

$$\dot{y}^i(t) = V^i(y^1(t), \ldots, y^n(t)), \qquad i = 1, \ldots, n, \tag{1}$$

$$y^i(t_0) = c^i, \qquad i = 1, \ldots, n, \tag{2}$$

1. *EXISTENCE: For any $t_0 \in \mathbb{R}$ and $x_0 \in U$, there exist an open interval J_0 containing t_0 and an open subset $U_0 \subset U$ containing x_0, such that for each $c \in U_0$, there is a C^1 map $y : J_0 \to U$ that solves (1)-(2).*
2. *UNIQUENESS: Any two differentiable solutions to (1)-(2) agree on their common domain.*
3. *SMOOTHNESS: Let J_0 and U_0 be as in 1. Let $\theta : J_0 \times U_0 \to U$ be the map defined by $\theta(t, x) = y(t)$, where $y : J_0 \to U$ is the unique solution to (1) with initial condition $y(t_0) = x$. Then θ is smooth.*

Proof. We sketch the proof given in [2].

First, we consider the existence. Since V is smooth, it is locally Lipschitz continuous. We can see that a continuous function y is a solution to the initial value problem if and only if

$$y^i(t) = c^i + \int_{t_0}^{t} V^i(y(s))ds,$$

and hence y is necessarily C^1. For any continuous map $y : J_0 \to U$, we define a new map $Iy : J_0 \to \mathbb{R}^n$ by

$$Iy(t) = c + \int_{t_0}^{t} V(y(s))ds.$$

Then, we want to use the fact that a contraction on a complete metric space has a unique fixed point. By restricting to a smaller open set U if necessary, we assume that V is Lipschitz continuous on U. Given $x_0 \in U$, we define a $U_0 \subset U$ to be an open ball centered on x_0. Then, given a $t_0 \in \mathbb{R}$, we can use a Lipschitz constant of V to define a small open interval J_0 centered on t_0. This allows us to define the set of all continuous maps $y : J_0 \to \overline{U_0}$ satisfying $y(t_0) = c$. We can show that this set is complete with the uniform metric. Finally, we show that I is a contraction and we can conclude.

The uniqueness (on common domain but with possibly different initial conditions) comes from an application of the ODE comparison theorem which is technical (i.e., boring) so I won't give any detail. I just mention another useful result: any locally Lipschitz map between two metric spaces restricts to a Lipschitz map on any compact set. I think that we cannot use the uniqueness of the fixed point of I because the initial conditions are different.

Finally, the smoothness. The general result is: if V is C^k then θ is C^k. This is proved by induction on k. The proof is long, boring, involved and overly technical, so I skip it. □

3.2 Vector fields, integral curves, and flows

Let M be a smooth manifold (with or without boundary).

3.2.1 Integral curves

Definition 3.2 (Integral curve of a vector field)**.** Let V be a vector field on M. An integral curve of V is a differentiable curve $\gamma : J \to M$ such that, for all $t \in J$, $\gamma'(t) = V_{\gamma(t)}$.

If $0 \in J$, the point $\gamma(0)$ is called the starting point of γ.

Let $(U, x^1, \ldots, x^n)$ be a smooth chart, and define $\gamma^i(t) = x^i \circ \gamma$. Then, γ is an integral curve of V if and only if

$$\dot{\gamma}^i(t) \left.\frac{\partial}{\partial x^i}\right|_{\gamma(t)} = V^i(\gamma(t)) \left.\frac{\partial}{\partial x^i}\right|_{\gamma(t)},$$

using Einstein summation convention. This reduces to an autonomous system of ordinary differential equations

$$\dot{\gamma}^i(t) = V^i(\gamma^1(t), \ldots, \gamma^n(t)),$$

for $i = 1, \ldots, n$.

Proposition 3.3 (Existence of Integral Curve)**.** *Let V be a **smooth** vector field. For each point $p \in M$, there exist $\epsilon > 0$ and a smooth curve $\gamma : (-\epsilon, \epsilon) \to M$ that is an integral curve of V starting at p.*

Proof. This is the existence statement of Theorem 3.1 applied to the coordinate representation of V. □

Lemma 3.4 (Rescaling and Translation)**.** *Let V be a smooth vector field, $J \subset \mathbb{R}$ be an interval, $\gamma : J \to M$ be an integral curve of V.*
For any $a \in \mathbb{R}$, the curve $\tilde{\gamma} : \tilde{J} \to \mathbb{R}$ defined by $\tilde{\gamma}(t) = \gamma(at)$ is an integral curve of the vector field aV, where $\tilde{J} = \{t : at \in J\}$.
For any $b \in \mathbb{R}$, the curve $\hat{\gamma} : \hat{J} \to \mathbb{R}$ defined by $\hat{\gamma}(t) = \gamma(t + b)$ is an integral curve of V, where $\hat{J} = \{t : t + b \in J\}$.

Proposition 3.5 (Naturality of Integral Curves)**.** *Let $X \in \Xi(M)$ and $Y \in \Xi(N)$ be smooth vector fields and $F : M \to N$ be a smooth map.*
X and Y are F-related if and only if F takes integral curves of X to integral curves of Y.

3.2.2 Flows

Definition 3.6 (Global Flow)**.** A global flow, or a one-parameter group action, on M is a continuous left $\mathbb{R}$-action on M.
In other words, it is a continuous map $\theta : \mathbb{R} \times M \to M$ satisfying the group action properties for the additive group $\mathbb{R}$: for all $s, t \in \mathbb{R}$ and $p \in M$

$$\theta(t, \theta(s, p)) = \theta(t + s, p),$$
$$\theta(0, p) = p.$$

For each $t \in \mathbb{R}$, we can define a continuous map $\theta_t : M \to M$ by $\theta_t(p) = \theta(t, p)$. It verifies the group laws: $\theta_t \circ \theta_s = \theta_{t+s}$ and $\theta_0 = \mathrm{Id}_M$. It is a homeomorphism. If the flow is smooth, θ_t is a diffeomorphism.
For each $p \in M$, we can define the curve $\theta^{(p)} : \mathbb{R} \to M$ by $\theta^{(p)}(t) = \theta(t, p)$. The image of this curve is the orbit of p under the group action.

Definition 3.7 (Infinitesimal Generator)**.** Let $\theta : \mathbb{R} \times M \to M$ be a smooth global flow. The infinitesimal generator of θ is the (rough) vector field V defined by $V_p = \dot{\theta}^{(p)}(0)$ for each $p \in M$.

Proposition 3.8. *Let $\theta : \mathbb{R} \times M \to M$ be a smooth global flow. The infinitesimal generator V of θ is a smooth vector field on M and each curve $\theta^{(p)}$ is an integral curve of V.*

The last proposition says that every smooth global flow gives rise to a smooth vector field whose integral curves are the curves defined by the flow. However, not every smooth vector field is the infinitesimal generator of a smooth global flow. This is because there exist smooth vector fields whose integral curves are not defined for all $t \in R$. For example, consider the smooth vector field $V = x^2 \partial/\partial x$ on $\mathbb{R}^2$. In the canonical coordinates (x, y), if the starting point p verifies $x(p) > 0$, then the integral curve diverges in finite time because $\gamma(t) = (1/(u - t), 0)$.

Definition 3.9 (Flow Domain)**.** A flow domain for M is an open subset $D \subset \mathbb{R} \times M$ such that for each $p \in M$, the set $D^{(p)} = \{t \in \mathbb{R} : (t, p) \in D\}$ is an open interval containing 0.

Definition 3.10 (Local Flow)**.** A flow, or a local flow, or a local one-parameter group action, on M is a continuous map $\theta : D \to M$ where D is a flow domain, that satisfies the group laws: for all $p \in M$, $\theta(0, p) = p$, and for all $s \in D^{(p)}, t \in D^{(\theta(s,p))}$ such that $t + s \in D^{(p)}$, $\theta(t, \theta(s, p)) = \theta(t + s, p)$.

As for smooth global flow, the infinitesimal generator V of a smooth local flow is a smooth vector field and each curve $\theta^{(p)}$ is an integral curve of V.

Theorem 3.11 (Fundamental Theorem on Flows)**.** *Let V be a smooth vector field. There is a unique smooth maximal flow $\theta : D \to M$ whose infinitesimal generator is V.*

Proposition 3.12 (Diffeomorphism Invariance of Flows)**.** *Let $F : M \to N$ be a diffeomorphism. If θ is the flow of $X \in \Xi(M)$, then the flow of F_*X is $\eta_t = F \circ \theta_t \circ F^{-1}$.*

Definition 3.13 (Complete Vector Fields)**.** A smooth vector field is complete if it generates a global flow, i.e., if each of its maximal integral curves is defined for all $t \in \mathbb{R}$.

Theorem 3.14. *Every compactly supported smooth vector field is complete. In particular, every smooth vector field on a compact manifold is complete.*

Proof. This is a direct application of the following lemma:

Lemma 3.15 (Uniform Time Lemma)**.** *Let V be a smooth vector field and let θ be its flow. If there exists a positive number ϵ such that for every $p \in M$, the domain of $\theta^{(p)}$ contains $(-\epsilon, \epsilon)$, then V is complete.*

□

Another useful lemma.

Lemma 3.16 (Escape Lemma)**.** *Let $\gamma : J \to M$ be a maximal integral curve of V. If J has a finite least upper bound b, then for any $t_0 \in J$, $\gamma([t_0, b))$ is not contained in any compact subset of M.*

3.3 Lie Derivatives

We want to define the directional derivative of a vector field W in the direction v. This makes perfect sense in $\mathbb{R}^n$:

$$D_v W(p) = \lim_{t \to 0} \frac{W_{p+tv} - W_p}{t}.$$

However, this doesn't work on a manifold. We replace $p + tv$ by a curve $\gamma(t)$ that starts at p with initial velocity $v \in T_pM$. So we obtain

$$D_v W(p) = \lim_{t \to 0} \frac{W_{\gamma(t)} - W_{\gamma(0)}}{t}.$$

This makes no sense because $W_{\gamma(t)} \in T_{\gamma(t)}M$ and $W_{\gamma(0)} \in T_pM$ do not belong to the same vector space. It works in $\mathbb{R}^n$ because each tangent space is canonically identified with $\mathbb{R}^n$ itself.

To avoid this problem, we do not seek the directional derivative of a vector field, but the derivative of a vector field *with respect to another vector field* V. Indeed, we can use the flow of V to push values of W back to T_pM. Let θ be the flow of V. We replace the curve $\gamma(t)$ by the flow $\theta(t, p) = \theta_t(p)$. We want to compare $W_{\theta_t(p)}$ and W_p. We can use the total derivative (i.e., the push forward) of θ_{-t} to send $W_{\theta_t(p)} \in T_{\theta_t(p)}M$ back to T_pM.

Definition 3.17 (Lie derivative)**.** Let W and V be smooth vector fields. The Lie derivative of W with respect to V is a rough vector field defined by

$$\begin{aligned}(\mathcal{L}_V W)_p &= \left.\frac{d}{dt}\right|_{t=0} d(\theta_{-t})_{\theta_t(p)}(W_{\theta_t(p)}) \\ &= \lim_{t \to 0} \frac{d(\theta_{-t})_{\theta_t(p)}(W_{\theta_t(p)}) - W_p}{t},\end{aligned}$$

when the derivative exists.

The flow is typically difficult or impossible to write down explicitly, so this definition is not very useful. Remember the expression of Lie bracket in coordinate:

Proposition 3.18. *Let X, Y be smooth vector fields and (x^i) a coordinate system.*

$$\begin{aligned}[X,Y] &= \left(X^i \frac{\partial Y^j}{\partial x^i} - Y^i \frac{\partial X^j}{\partial x^i}\right) \frac{\partial}{\partial x^j} \\ &= (XY^j - YX^j)\frac{\partial}{\partial x^j}.\end{aligned}$$

Proof.

$$\begin{aligned}[X,Y]f &= X^i \frac{\partial}{\partial x^i}\left(Y^j \frac{\partial f}{\partial x^j}\right) - Y^i \frac{\partial}{\partial x^i}\left(X^j \frac{\partial f}{\partial x^j}\right) \\ &= X^i \frac{\partial Y^j}{\partial x^i}\frac{\partial f}{\partial x^j} + X^i Y^j \frac{\partial^2 f}{\partial x^j \partial x^i} - Y^i \frac{\partial X^j}{\partial x^i}\frac{\partial f}{\partial x^j} - Y^i X^j \frac{\partial^2 f}{\partial x^j \partial x^i} \\ &= X^i \frac{\partial Y^j}{\partial x^i}\frac{\partial f}{\partial x^j} - Y^i \frac{\partial X^j}{\partial x^i}\frac{\partial f}{\partial x^j}.\end{aligned}$$

□

Note that this formula is more useful than $[X,Y]_p = X_p Y - Y_p X$ because the second derivative terms have already been canceled out.

Theorem 3.19. *Let $V, W \in \Xi(M)$. Then $\mathcal{L}_V W = [V,W]$.*

Proof. This is the proof from [2]. Let $\mathcal{R}(V)$ be the set of regular points of V, i.e., $\mathcal{R}(V) = \{p \in M : V_p \neq 0\}$. Since V is continuous, $\mathcal{R}(V)$ is open in M. Its closure is the support of V: $\overline{\mathcal{R}(V)} = \operatorname{supp}(V)$. There are three cases.

1. $p \in \mathcal{R}(V)$. First, note that in any coordinate system, the matrix of $d(\theta_{-t})_{\theta_t(x)}$ (i.e., the Jacobian matrix of θ_{-t} at $\theta_t(x)$) is the identity for any $x \in M$ and any fixed t.

   Since p is regular, there is a theorem that prove the existence of coordinates (u^i) on a neighborhood of p such that $V = \partial/\partial u^1$. Hence

   $$(\mathcal{L}_V W)_p = \frac{\partial W^j}{\partial u^1}(p)\left.\frac{\partial}{\partial u^j}\right|_p,$$

   witch is equal to $[V,W]_p$ using Proposition 3.18.

2. $p \in \operatorname{supp}(V)$. We use the fact that if two continuous functions agree on a open set and if the codomain is Hausdorff then they agree on the closure of this open set. Here, the codomain is the tangent bundle which is Hausdorff, and $\overline{\mathcal{R}(V)} = \operatorname{supp}(V)$.

3. $p \in M \setminus \operatorname{supp}(V)$. In this case, there is a neighborhood of p such that $V \equiv 0$. So, for all t, θ_t is the identity map in this neighborhood. Hence $(\mathcal{L}_V W)_p = 0$. Using Proposition 3.18, $[V,W]_p = 0$.

□

3.4 Commuting Vector Fields

Definition 3.20. Let $V, W \in \Xi(M)$. We say that V and W commute if $VWf = WVf$ for every smooth function f, or equivalently if $[V,W] \equiv 0$.

Definition 3.21. Let θ be a smooth flow. A vector field W is invariant under θ if W is θ_t-related to itself for each t, i.e., for all (t,p) in the domain of θ, we have $d(\theta_t)_p(W_p) = W_{\theta_t(p)}$.

Theorem 3.22. *Let V, W be smooth vector fields. V and W commute if and only if W is invariant under the flow of V if and only if V is invariant under the flow of W.*

Since $[V, V] = 0$, every smooth vector field is invariant under its own flow.

Definition 3.23. Let (E_i) be a smooth local frame defined on U, i.e., $(E_i|_p)$ forms a basis for T_pM at each $p \in U$. We call (E_i) a commuting frame, or holonomic frame, if $[E_i, E_j] = 0$ for all i and j.

Every coordinate frame is a commuting frame. Hence, a necessary condition for a smooth frame to be expressible as a coordinate frame in some chart is that it be a commuting frame. It turns out that it is also a sufficient condition.

Theorem 3.24 (Canonical Form for Commuting Vector Fields)**.** *Let $(V_1, \dots, V_k)$ be a linearly independent k-tuple of smooth commuting vector fields on an open subset W. For each $p \in W$, there exists a coordinate chart $(U, (s^i))$ centered at p such that $V_i = \partial/\partial s^i$.*

If $S \subset W$ is an embedded codimension-k (dimension $n - k$) submanifold and $p \in S$ such that T_pS is complementary to $\mathrm{Span}(V_1|_p, \dots V_k|_p)$*, then the coordinates can also be chosen such that $S \cap U$ is the slice defined by $s^1 = \cdots = s^k = 0$.*

3.5 Involutivity

Let M be a n-dimensional differentiable manifold. Let V be a nonvanishing vector field. In other words, all point of M are regular points of V. Hence, the integral curves of V are smooth immersions, since there velocity never vanishes. Moreover, there exist a coordinate system (u^i) on a neighborhood of any point p such that $V = \partial/\partial u^1$. This means that the images of the integral curves can locally be seen as parallel lines in Euclidean space. These curves are entirely determined by the knowledge of their velocity vectors, as stated by Theorem 3.11. Foliations are a generalization of this idea to higher-dimensional submanifolds.

3.5.1 Integral manifold and involutivity

We start with several definitions. Let $k \leq n$.

Definition 3.25 (Distribution)**.** A distribution D is the assignment to each point $p \in M$ of a subspace D_p of the tangent space T_pM.

Definition 3.26 (Constant rank distribution)**.** A constant rank distribution D of dimension k (i.e., rank-k distribution) is a distribution where, for each point $p \in M$, D_p has dimension k.
A constant rank distribution of rank-k can also be defined as a rank-k subbundle of TM.

Definition 3.27 (Smooth distribution)**.** A distribution D is smooth if for any $p \in M$ there exists a neighborhood U and k smooth vector fields $X_1, \dots, X_k$ on U such that $X_1|_q, \dots, X_k|_q$ is a basis of D_q for any $q \in U$.

Let D be a rank-k smooth distribution on M. A smooth vector field X on M is *tangent* to D if, for each point $p \in M$, $X_p \in D_p$.

Definition 3.28 (Submodule associated to a distribution)**.** Let Ξ_D be the set of smooth global vector fields tangent to D. We call Ξ_D the submodule associated to D. It is a submodule of the module $\Xi(M)$ of smooth global vector fields over the ring $\Omega^0(M)$ of smooth functions.

Remember that a module is like a vector space, but over a ring instead of a field.

Definition 3.29 (Integral manifold)**.** Let D be a smooth rank-k distribution. A nonempty immersed submanifold N is called an integral manifold of D if $T_pN = D_p$ at each point $p \in N$.

Example 3.30.

- If V is a nowhere-vanishing smooth vector field, then V spans a smooth rank-1 distribution. The image of any integral curve of V is an integral manifold of this distribution.
- Let D be the smooth distribution on $\mathbb{R}^3$ spanned by $X = \partial/\partial x + y\partial/\partial z$ and $Y = \partial/\partial y$. D has no integral manifolds.

Definition 3.31 (Involution)**.** Let D be a smooth rank-k distribution. Let X and Y be two smooth local sections of D, i.e., X and Y are defined on a open subset of M and $X_p, Y_p \in D_p$ for each p.
D is involutive if $[X, Y]$ is also a local section of D.

Proposition 3.32. *Let D be a smooth rank-k distribution. Let $\Xi_D \subset \Xi(M)$ be the space of smooth global sections of D. Then D is involutive if and only if Ξ_D is a Lie subalgebra of $\Xi(M)$.*

Proof. Assume D is involutive. Since any smooth global section is also a smooth local section, then Ξ_D is closed under Lie brackets. Since Ξ_D is a vector subspace of $\Xi(M)$, it is a Lie subalgebra.

Now, assume Ξ_D is a Lie subalgebra of $\Xi(M)$. Any local section on an open set U can be extended to a global section using a bump function whose support is in U. Since Ξ_D is a Lie subalgebra, it is closed under Lie brackets for global sections. For any $p \in U$, there is a neighborhood of p where the local and global sections are equal, so their Lie bracket is equal. Hence, D is involutive. □

Definition 3.33 (Integrable distribution)**.** A smooth rank-k distribution D is integrable if each point of M is contained in an integral manifold of D.

Proposition 3.34. *Every integrable distribution is involutive.*

Proof. Let D be an integrable distribution and let X and Y be smooth local sections of D defined on some open set U. Let $p \in U$ and let N be an integral manifold of D containing p. Since X and Y are sections of D, then X and Y are tangent to N. Thus, there exist X' and Y' that are vector fields on N and that are i-related to X and Y, where $i : N \mapsto M$ is the inclusion map. Then, by a well-known property of the Lie bracket, $[X', Y']$ is i-related to $[X, Y]$, so $[X, Y]$ is also tangent to N. Hence, $[X, Y]_p \in D_p$ and D is involutive. □

To check the involutivity of a distribution, it is sufficient to check the involutivity of a smooth local frame.

3.5.2 Involutivity and Differential Forms

Lemma 3.35. *Let D be a rank-k distribution. Then D is smooth if and only if each point $p \in M$ has a neighborhood U on which there are smooth 1-forms $\omega^1, \ldots, \omega^{n-k}$ such that for each $q \in U$,*

$$D_q = \ker \omega^1\big|_q \cap \ldots \cap \ker \omega^{n-k}\big|_q .$$

For dimensional reasons, these 1-forms must be linearly independent. They are called local defining forms for D.

Definition 3.36. Let $0 \leq p \leq n$. A p-form $\omega \in \Omega^p(M)$ annihilates D if $\omega(X_1, \ldots, X_p) = 0$ whenever $X_1, \ldots, X_p$ are local sections of D.

Lemma 3.37. *Let D be a smooth rank-k distribution. Let $\omega^1, \ldots \omega^{n-k}$ be smooth local defining forms for D on an open set U. A smooth p-form η defined on U annihilates D if and only if*

$$\eta = \sum_{i=1}^{n-k} \omega^i \wedge \beta^i,$$

for some smooth $(p-1)$-forms $\beta^1, \ldots, \beta^{n-k}$ on U.

Proposition 3.38. *Let ω be a smooth 1-form. Let X, Y be smooth vector fields. Then*

$$d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]).$$

Proof. Consider the case $\omega = udv$ for smooth functions u and v (the general case is a sum of such 1-forms). The left-hand side is

$$\begin{aligned} d(udv)(X, Y) &= du \wedge dv(X, Y) \\ &= du(X)dv(Y) - dv(X)du(Y) \\ &= XuYv - XvYu. \end{aligned}$$

The right hand-side is

$$\begin{aligned} X(udv(Y)) - Y(udv(X)) - udv([X,Y]) &= X(uYv) - Y(uXv) - u[X,Y]v \\ &= (XuYv + uXYv) - (YuXv + uYXv) - u(XYv - YXv) \\ &= XuYv - YuXv = XuYv - XvYu. \end{aligned}$$

□

Theorem 3.39 (1-Form Criterion for Involutivity)**.** *Let D be a smooth rank-k distribution. Then D is involutive if and only if for any smooth 1-form η that annihilates D on an open subset U, then $d\eta$ also annihilates D on U.*

Proof. Assume that D is involutive. Let η be a smooth 1-form that annihilates D on U. Let X, Y be smooth local sections of D. We have

$$d\eta(X,Y) = X(\eta(Y)) - Y(\eta(X)) - \eta([X,Y]).$$

By assumption, $\eta(Y) = \eta(X) = 0$ and since $[X,Y]$ is also a smooth local section of D by involutivity, $\eta([X,Y]) = 0$. Hence, $d\eta(X,Y) = 0$.

Now, assume that for any smooth 1-form η that annihilates D on U, then $d\eta$ also annihilates D on U. Let X, Y be smooth local sections of D. Let $\omega^1, \ldots, \omega^{n-k}$ be smooth local defining forms for D. Then, for each $i = 1, \ldots, n-k$,

$$\omega^i([X,Y]) = X(\omega^i(Y)) - Y(\omega^i(X)) - d\omega^i(X,Y) = 0.$$

Hence, $[X,Y]$ takes its values in D. □

As for the Lie bracket condition, this condition needs only be checked for a particular set of smooth defining forms in a neighborhood of each point.

Definition 3.40. Let $\Omega^*(M) = \Omega^0(M) \oplus \ldots \oplus \Omega^n(M)$ be the graded algebra of smooth differential forms on M. An ideal in $\Omega^*(M)$ is a linear subspace $\mathfrak{I} \subset \Omega^*(M)$ such that for any $\omega \in \mathfrak{I}$ and for any $\eta \in \Omega^*(M)$, we have $\eta \wedge \omega \in \mathfrak{I}$.

Proposition 3.41. *Let D be a smooth rank-k distribution. Let $\mathfrak{I}^p(D) \subset \Omega^p(M)$ be the space of smooth p-form that annihilate D. Let $\mathfrak{I}(D) = \mathfrak{I}^0(D) \oplus \ldots \oplus \mathfrak{I}^n(D) \subset \Omega^*(M)$. Then, $\mathfrak{I}(D)$ is an ideal in $\Omega^*(M)$. We call $\mathfrak{I}(D)$ a* ***Pfaffian system****.*

Definition 3.42. Let $\mathfrak{I}$ be an ideal in $\Omega^*(M)$. We said that $\mathfrak{I}$ is a differential ideal if $d(\mathfrak{I}) \subset \mathfrak{I}$.

Proposition 3.43. *A smooth rank-k distribution D is involutive if and only if $\mathfrak{I}(D)$ is a differential ideal in $\Omega^*(M)$.*

3.6 The Frobenius Theorem

3.6.1 Proof of the Frobenius Theorem

Definition 3.44. Let D be a rank-k distribution. A coordinate chart (U, ϕ) is flat for D if $\phi(U)$ is a cube in $\mathbb{R}^n$ and at points of U, D is spanned by the first k coordinate vector fields $\partial/\partial x^1, \ldots, \partial/\partial x^k$.

In a flat chart, each slice of the form $x^{k+1} = c^{k+1}, \ldots, x^n = c^n$ is an integral manifold of D.

Definition 3.45. A rank-k distribution D is completely integrable if there exists a flat chart for D.

For now, we have the following implications:

$$\text{completely integrable} \Rightarrow \text{integrable} \Rightarrow \text{involutive}.$$

The Frobenius theorem shows that these implications are actually equivalences.

Theorem 3.46 (Frobenius)**.** *Every involutive distribution is completely integrable.*

Proof. According to Theorem 3.24, any distribution locally spanned by independent smooth *commuting* vector fields is completely integrable. This is because the coordinate chart whose existence is guaranteed by that theorem is flat (after shrinking the domain if necessary so the image is a cube). Thus, **it suffices to show that every involutive distribution is locally spanned by independent smooth commuting vector fields.**

Once again, I rewrite the proof of [2]. There are lots of identifications between objects of homeomorphic spaces in the proof, but I will distinguish them in my proof in order to be very clear, even if the notations will be cumbersome.

Let D' be an involutive smooth rank-k distribution. Let $p' \in M$. Since complete integrability is a local question, we consider a coordinate chart (U', ϕ) such that there exists a smooth local frame $X'_1, \dots, X'_k$ for D' on U' (we can always choose a smaller U' for which such local frame exists). For the following, we will work on $U = \phi(U') \subset \mathbb{R}^n$. So we denote $p = \phi(p')$ and $X_i = d\phi(X'_i)$ for $i = 1, \dots, k$ which are also smooth independent vector fields on U because ϕ is a diffeomorphism. We denote by D the distribution in $\mathbb{R}^n$ spanned by $X_1, \dots, X_k$. Obviously, we have $D_{\phi(q')} = d\phi(D'_{q'})$ for every $q' \in U'$. Let $r^1, \dots, r^n$ be the standard coordinates on $\mathbb{R}^n$. By reordering the coordinates if necessary, we can assume that D_p is complementary to $\mathrm{Span}\{\partial/\partial r^{k+1}|_p, \dots, \partial/\partial r^n|_p\} \subset T_pU \sim T_p\mathbb{R}^n$, i.e.,

$$D_p \oplus \mathrm{Span}\left\{ \left.\frac{\partial}{\partial r^{k+1}}\right|_p, \dots, \left.\frac{\partial}{\partial r^n}\right|_p \right\} = T_p\mathbb{R}^n.$$

Let $\pi : U \subset \mathbb{R}^n \to \mathbb{R}^k$ be the projection onto the first k coordinates, i.e., $\pi(r^1, \dots, r^n) = (r^1, \dots, r^k)$. Let $\tilde{r}^i = \pi \circ r^i$ for $i = 1, \dots, k$. The projection π induces a smooth bundle homomorphism $d\pi : TU \to T\mathbb{R}^k$ such that, for every $q \in U$,

$$d\pi \left(\sum_{i=1}^n v^i \left.\frac{\partial}{\partial r^i}\right|_q \right) = \sum_{i=1}^k v^i \left.\frac{\partial}{\partial \tilde{r}^i}\right|_{\pi(q)},$$

such that

$$\ker d\pi_q = \mathrm{Span}\{\partial/\partial r^{k+1}|_q, \dots, \partial/\partial r^n|_q\}.$$

In particular, we have $D_p \oplus \ker d\pi_p = T_p\mathbb{R}^n$.

Let $I : D \hookrightarrow TU$ be the inclusion map. We consider $d\pi|_D \colon D \to T\mathbb{R}^k$ defined as $d\pi|_D = d\pi \circ I$. It is the restriction of $d\pi$ on D. As a composition of smooth bundle homomorphisms, $d\pi|_D$ is a smooth bundle homomorphism. Thus, for every $q \in U$, the matrix entries of $d\pi|_{D_q}$ with respect to the frames $(X_i|_q)$ and $(\partial/\partial \tilde{r}^j|_{\pi(q)})$ are smooth functions of q.

Since $D_p \oplus \ker d\pi_p = T_p\mathbb{R}^n$, the restriction of $d\pi_p$ to D_p, denoted $d\pi|_{D_p}$, is bijective from D_p to $T_{\pi(p)}\mathbb{R}^k$ (i.e., it is an isomorphism of vector spaces). By continuity (of the entries of the matrix of $d\pi|_{D_q}$ with respect to q), there exists a neighborhood V of p such that, for every $q \in V$, $d\pi|_{D_q}$ is bijective. Hence, the matrix entries of $(d\pi|_{D_q})^{-1} : T_{\pi(q)}\mathbb{R}^k \to D_q$ are also smooth functions of q. At last, we can **define a new smooth local frame $V_1, \dots, V_k$ for D in V** by

$$V_i|_q = (d\pi|_{D_q})^{-1} \left.\frac{\partial}{\partial \tilde{r}^i}\right|_{\pi(q)},$$

for every $q \in V$. Note that we are identifying V_i with $dI(V_i)$.

Since $(\partial/\partial \tilde{r}^j|_{\pi(q)})$ is a basis for $T_{\pi(q)}\mathbb{R}^k$ and $(d\pi|_{D_q})^{-1}$ is bijective, $(V_i|_q)$ is also a basis for D_q. Since $(d\pi|_D)^{-1}$ is a smooth bundle homomorphism, the vector fields $V_1, \dots, V_k$ are smooth. Thus, it remains only to show that $[V_i, V_j] = 0$ for all i, j. Since

$$\left.\frac{\partial}{\partial \tilde{r}^i}\right|_{\pi(q)} = (d\pi|_{D_q})V_i|_q = d\pi_q(V_i|q),$$

V_i and $\partial/\partial \tilde{r}^i$ are π-related. By the naturality of Lie brackets, for every $q \in V$

$$d\pi_q([V_i, V_j]_q) = \left[\frac{\partial}{\partial \tilde{r}^i}, \frac{\partial}{\partial \tilde{r}^j}\right] = 0,$$

because $(\tilde{r}^i)$ is a coordinate map. **Since D is involutive**, $[V_i, V_j]$ takes its values in D. For every $q \in V$, $d\pi|_{D_q}$ is bijective, thus it is injective. In other word, $d\pi$ is injective on each fiber of D. Hence, for every $q \in V$, we have $[V_i, V_j]_q = 0$. □

Now, we give one of the most important consequences of the Frobenius Theorem.

Proposition 3.47 (Local Structure of Integral Manifolds)**.** *Let D be an involutive smooth rank-k distribution. Let $(U, (x^i))$ be a flat chart for D. Let H be an integral manifold of D. Then $H \cap U$ is a union of countably many disjoint open subsets of **parallel** k-dimensional slices of U (the plaques?), each of which is open **in** H and embedded in M.*

3.6.2 Summary of the proof of the Frobenius Theorem

The proof can be decomposed into four parts:

1. **completely integrable ⇒ integrable.** This is obvious because for any coordinate chart $(U, (x^i))$, the set of points $\{p \in U : x^{k+1}(p) = c^{k+1}, \ldots, x^n(p) = c^n\}$ where $c^{k+1}, \ldots, c^n$ are constants, is a regular submanifold of U by definition. Recall that any regular submanifold is an embedded submanifold (the converse is also true).

2. **integrable ⇒ involutive.** This is a consequence of the naturality of the Lie bracket applied to the inclusion map of an integral manifold.

3. **locally spanned by independent smooth commuting vector fields ⇒ completely integrable.** This is because a set of independent smooth commuting vector fields are elements of the basis associated to some coordinate chart.

4. **involutive ⇒ locally spanned by independent smooth commuting vector fields.**

3.6.3 Method to find integral manifolds

The proof of the Frobenius theorem gives a technique for finding the integral manifolds in a neighborhood U of a point p. The goal is to find a flat chart in U. Let (x^i) be a coordinate system in U and denote by $(q^1, \ldots, q^n)$ the coordinates of any $q \in U$ in this coordinate system. Given a smooth rank-k distribution D which is **involutive**, the idea is to:

1. Find $n-k$ coordinate vector fields $\partial/\partial x^{k+1}, \ldots, \partial/\partial x^n$ that span a complementary subspace of D_q for any $q \in U$. If U is small enough, this is always possible (up to reordering the coordinate vector fields).

2. Find smooth commuting vector fields $V_1, \ldots, V_k$ spanning D. According to the proof, the coordinate projection π onto the first k coordinates induces an isomorphism when $d\pi$ is restricted to D. Hence, to find $V_1, \ldots, V_k$, it suffices to find smooth vector fields of D that are π-related to $\partial/\partial x^1, \ldots, \partial/\partial x^k$ (they will automatically be commuting and spanning D).

3. Find the flows $\theta^1_{t^1}(q), \ldots, \theta^k_{t^k}(q)$ of $V_1, \ldots, V_k$ for $q \in U$.

4. Define $\phi(t^1, \ldots, t^n) = \theta^1_{t^1} \circ \ldots \circ \theta^k_{t^k}(q)$ where $(q^1, \ldots, q^n) = (p^1, \ldots, p^k, p^{k+1} + t^{k+1}, \ldots, p^n + t^n)$. In other words, q belongs to the submanifold containing p and whose tangent spaces are spanned by $\partial/\partial x^{k+1}, \ldots, \partial/\partial x^n$. This submanifold is transverse to D in U. Since the flows commute, it does not matter in which order they are applied.

5. Solve for $(t^1, \ldots, t^n) = \phi^{-1}(q^1, \ldots, q^n)$ for $q \in U$. We obtain a flat chart $(U, (t^1, \ldots, t^n))$. The integral manifolds are defined by $t^{k+1} = c^{k+1}, \ldots, t^n = c^n$.

3.7 Foliations

4 Curvature

In this section, I want to provide a good intuition for the definition and main properties of Riemannian curvature. I start with intuitions for affine connections and go all the way down to the second fundamental form (and maybe the Gauss-Bonnet formula and Jacobi fields if I am motivated).

I rely on Lee's book about Riemannian manifold.

How does the coordinate expression of the metric change when we change the coordinate system. In particular for the Euclidean metric: is it always the identity matrix?

4.1 Tensors

4.1.1 Aside: definition of the tensor product

We will define rigorously the tensor product of two real vector spaces V and W. First, it is important to note that the tensor product $V \otimes W$ is defined *up to an isomorphism*. That means that we are considering the quotient of the category of vector spaces with respect to the isomorphism equivalence relation (if that makes any sense). When we talk about the tensor product $V \otimes W$ as a vector space, what we really mean is the *equivalence class* of $V \otimes W$.

That being said, there are three different ways to define the tensor product $V \otimes W$ (according to Wikipedia ...).

From bases Let B_V and B_W be bases of V and W respectively.

Definition 4.1. $V \otimes W$ is the set of functions from the Cartesian product $B_V \times B_W$ to $\mathbb{R}$ that have a finite number of nonzero values.

Note that if V and W are both finite-dimensional then all such functions have a finite number of nonzero values. $V \otimes W$ becomes a vector space using the pointwise operations, i.e., $(f+g)(v,w) = f(v,w) + g(v,w)$ and $(\lambda f)(v,w) = \lambda f(v,w)$. Let $v \in B_V$ and $w \in B_W$. Define $v \otimes w \in V \otimes W$ by:

$$\begin{aligned} v \otimes w(v,w) &= 1, \\ v \otimes w(v',w') &= 0, v' \neq v, w' \neq w. \end{aligned}$$

The set $\{v \otimes w : v \in B_V, w \in B_W\}$ is called the tensor product of the bases B_V and B_W. It is a basis of $V \otimes W$. Let $x = x^i v_i \in V$ and $y = y^j w_j \in W$. The tensor product of x and y is

$$x \otimes y = x^i y^j v_i \otimes w_j$$

Remark 4.2.

- The coordinates of $x \otimes y$ is the outer product of the coordinates of x with the coordinates of y. Therefore, the tensor product is a generalization of the outer product.
- The map $(x,y) \mapsto x \otimes y$ from $V \times W$ to $V \otimes W$ is a bilinear map.
- If one chooses another bases for V and W then a different tensor product is defined. However, the change of basis isomorphism defines a canonical isomorphism between the two tensor products, so we can identify them using the remark made in introduction.
- Do not confuse the tensor product with the Cartesian product seen as a vector space. The tensor product has dimension $\dim V \times \dim W$ since a basis element $v \otimes w$ is obtained by choosing a element $v \in B_V$ *and* an element $w \in B_W$. The Cartesian product has dimension $\dim V + \dim W$ since a basis element is obtained by choosing a element $v \in B_V$ (i.e., $(v,0)$) *or* an element $w \in B_W$ (i.e., $(0,w)$).

As a quotient space

Definition 4.3 (Free vector space)**.** The *free vector space* L is the vector space of functions $V \times W \to \mathbb{R}$ that have a finite number of nonzero values, and where $(v, w) \in V \times W$ is identified with the function that takes the value 1 on (v, w) and 0 otherwise.

The Cartesian product $V \times W$ is a basis of L. The free vector space L is infinite dimensional.

Definition 4.4. Let R be the linear subspace of L spanned by:

- $(v_1 + v_2, w) - (v_1, w) - (v_2, w)$,
- $(v, w_1 + w_2) - (v, w_1) - (v, w_2)$,
- $(\lambda v, w) - \lambda(v, w)$,
- $(v, \lambda w) - \lambda(v, w)$.

Then the tensor product is $V \otimes W = L/R$.

If $\pi : V \times W \to V \otimes W$ is the quotient map, then we denote $v \otimes w = \pi(v, w)$.

Remark 4.5.

- This definition is basis independent.
- This is the definition used in Willmore's book.

From an universal property This is the purest definition of the tensor product.

Definition 4.6. Let $V \otimes W$ be a vector space, $\otimes : (v, w) \mapsto v \otimes w$ be a bilinear map from $V \times W$ to $V \otimes W$, and Z be any vector space.

Then $V \otimes W$ is the tensor product of V and W if, for any bilinear map $h : V \times W \to Z$, there is a *unique* linear map $\tilde{h} : V \otimes W \to Z$ such that $h = \tilde{h} \circ \otimes$, i.e., $h(v, w) = \tilde{h}(v \otimes w)$ for every $v \in V$ and $w \in W$.

Remark 4.7.

- Category theory teaches us that two objects satisfying an universal property are related by a unique isomorphism.
- This definition is non-constructive. The two preceding definitions can be viewed as proofs of existence of the tensor product (provided that we prove that these definitions satisfy the universal property).
- Every property of the tensor product can be deduced from the universal property. So, in practice, one may forget the method that has been used to prove its existence.

4.1.2 Tensors on a vector space

Let V be a n-dimensional real vector space.

Definition 4.8 (Vector and Covector)**.**

- The elements of V are called *vectors*. A family of vectors uses *subscripts* while the components of a vector use *superscripts*.

  For example, (E_j) is a basis of V. Let $X \in V$. Then, using *Einstein summation convention*, we have $X = X^j E_j$.

- The elements of the dual $V^* = \mathrm{End}(V, \mathbb{R})$ are called *covectors*. A family of covectors uses *superscripts* while the components of a covector use *subscripts*.

  For example, (ϕ^i) is a basis of V^*. Let $\omega \in V^*$. Then $\omega = \omega_i \phi^i$.

With these notations, and using a basis (E_j) and its dual basis (ϕ^i), the natural pairing can be written:

$$\langle \omega, X \rangle = \omega(X) = \omega_i \phi^i(X^j E_j) = \omega_i X^i.$$

Definition 4.9 (Tensor)**.** The space $T^k_l(V)$ of tensors of rank $k+l$ that are k-covariant and l-contravariant is the space of *multilinear maps*

$$F : \underbrace{V^* \times \ldots \times V^*}_{l \text{ copies}} \times \underbrace{V \times \ldots \times V}_{k \text{ copies}} \to \mathbb{R}.$$

Remark 4.10.

- *Covariant* means using *vectors* as arguments. It is indicated as *subscripts* in a tensor component.
- *Contravariant* means using *covectors* as arguments. It is indicated as *superscripts* in a tensor component.

Definition 4.11 (Tensor Product)**.** The *tensor product* of $F \in T^k_l(V)$ and $G \in T^p_q(V)$ can be defined (in a not-at-all algebraic way) by $F \otimes G \in T^{k+p}_{l+q}(V)$:

$$F \otimes G(\omega^1, \ldots, \omega^{l+q}, X_1, \ldots, X_{k+p}) = F(\omega^1, \ldots, \omega^l, X_1, \ldots, X_k) G(\omega^{l+1}, \ldots, \omega^{l+q}, X_{k+1}, \ldots, X_{k+p}).$$

Lemma 4.12. *There is a natural (i.e., basis-independent) isomorphism between $T^k_{l+1}(V)$ and the space of multilinear maps*

$$\underbrace{V^* \times \ldots \times V^*}_{l} \times \underbrace{V \times \ldots \times V}_{k} \to V$$

In particular, we have the following natural identifications:

- $T^0_0(V) = T^0(V) = \mathbb{R}$.
- $T^0_1(V) = T_1(V) = \{V^* \to \mathbb{R}\} = V^{**} = V$.
- $T^1_0(V) = T^1(V) = \{V \to \mathbb{R}\} = V^*$.
- $T^1_1(V) = \{V^* \times V \to \mathbb{R}\} = \{V \to V\} = \mathrm{End}(V)$.

Proof. Define the map $\Phi : \{\underbrace{V^* \times \ldots \times V^*}_{l} \times \underbrace{V \times \ldots \times V}_{k} \to V\} \to T^k_{l+1}(V)$ by

$$\Phi(A)(\omega^1, \ldots, \omega^{l+1}, X_1, \ldots, X_k) = \omega^1(A(\omega^2, \ldots, \omega^{l+1}, X_1, \ldots, X_k)).$$

Φ is a basis-independent linear map. It remains to show that it is an isomorphism. If $\Phi(A) \equiv 0$, then by choosing $\omega^1 \neq 0$ we see that $A \equiv 0$, so Φ is injective. The dimension of the space of multilinear maps $\underbrace{V^* \times \ldots \times V^*}_{l} \times \underbrace{V \times \ldots \times V}_{k} \to V$ is $n^{k+l} \times n = n^{k+l+1} = \dim T^k_{l+1}(V)$. Hence, Φ is an isomorphism. □

Let (E_j) be a basis of $V = T^0_1(V)$, (ϕ^i) its dual basis for $V^* = T^1_0(V)$, and $F \in T^k_l(V)$. The set $\{E_{j_1} \otimes \ldots \otimes E_{j_l} \otimes \phi^{i_1} \otimes \ldots \otimes \phi^{i_k}\}$ is a basis of $T^k_l(V)$, where the indices i_p, j_q range from 1 to n. This basis acts on basis elements by

$$E_{j_1} \otimes \ldots \otimes E_{j_l} \otimes \phi^{i_1} \otimes \ldots \otimes \phi^{i_k}(\phi^{s_1}, \ldots, \phi^{s_l}, E_{r_1}, \ldots, E_{r_k}) = \delta^{s_1}_{j_1} \ldots \delta^{s_l}_{j_l} \delta^{i_1}_{r_1} \ldots \delta^{i_k}_{r_k}.$$

Note that the elements E_{j_p} act on ϕ^{s_q} as elements of V^{**} by

$$E_{j_p}(\phi^{s_q}) = \mathrm{ev}_{E_{j_p}}(\phi^{s_q}) = \langle \phi^{s_q}, E_{j_p} \rangle = \phi^{s_q}(E_{j_p}).$$

4.1.3 Tensor bundles

Let M be a n-dimensional manifold.

Definition 4.13 (Tensor Bundle)**.**
The *bundle of* $\binom{k}{l}$*-tensors* on M is $T^k_l M := \bigsqcup_{p\in M} T^k_l(T_pM)$.
The *bundle of* k*-forms* is $\Lambda^k M := \bigsqcup_{p\in M} \Lambda^k(T_pM)$.

Remark 4.14. We have the following identifications:

- $T^0_0 M = T^0 M = C^\infty(M)$ is the space of smooth real-valued functions on M.
- $T^0_1 M = T_1 M = TM = \bigsqcup_{p\in M} T_pM$ is the *tangent bundle*.
- $T^1_0 M = T^1 M = \Lambda^1 M = T^*M = \bigsqcup_{p\in M} T^*_p M = \bigsqcup_{p\in M} (T_pM)^*$ is the *cotangent bundle*.

Let $(U, (x^i))$ be a local coordinate chart and $p \in U$. Let ∂_i be the corresponding basis of T_pM and (dx^i) its dual basis. Let $F \in T^k_l(T_pM)$. We have

$$F = F^{j_1\dots j_l}_{i_1\dots i_k} \partial_{j_1} \otimes \dots \otimes \partial_{j_l} \otimes dx^{i_1} \otimes \dots \otimes dx^{i_k},$$

where

$$F^{j_1\dots j_l}_{i_1\dots i_k} = F(dx^{j_1}, \dots, dx^{j_l}, \partial_{i_1}, \dots, \partial_{i_k}).$$

Definition 4.15 (Tensor Field)**.** A *tensor field* on M is a smooth section of some tensor bundle $T^k_l M$.
A *differential* k*-form* is a smooth section of $\Lambda^k M$.

We denote by $\mathcal{T}^k_l(M)$ the space of $\binom{k}{l}$-tensor fields. It is an infinite dimensional vector space under pointwise addition and multiplication by constants. Be careful not to confuse a tensor field F and the tensor F_p obtained by applying F at $p \in M$. We use the following notations:

- $\mathcal{T}^1_0(M) = \mathcal{T}^1(M)$ the space of 1-form (smooth sections of the cotangent bundle).
- $\mathcal{T}^0_1(M) = \mathcal{T}(M)$ the space of smooth vector fields (smooth sections of the tangent bundle).

Lemma 4.16 (Tensor Characterization Lemma)**.** *Any* $\binom{k}{l}$*-tensor field* F *induces a multilinear map over the module* $C^\infty(M)$*:*

$$F : \underbrace{\mathcal{T}^1(M) \times \dots \times \mathcal{T}^1(M)}_{l} \times \underbrace{\mathcal{T}(M) \times \dots \times \mathcal{T}(M)}_{k} \to C^\infty(M).$$

Conversely, any multilinear map over $C^\infty(M)$*:*

$$\tau : \underbrace{\mathcal{T}^1(M) \times \dots \times \mathcal{T}^1(M)}_{l} \times \underbrace{\mathcal{T}(M) \times \dots \times \mathcal{T}(M)}_{k} \to C^\infty(M),$$

is induced by a $\binom{k}{l}$*-tensor field.*

Similarly to Lemma 4.12, a map

$$\tau : \underbrace{\mathcal{T}^1(M) \times \dots \times \mathcal{T}^1(M)}_{l} \times \underbrace{\mathcal{T}(M) \times \dots \times \mathcal{T}(M)}_{k} \to \mathcal{T}(M),$$

is multilinear over $C^\infty(M)$ if and only if *it is induced by a* $\binom{k}{l+1}$*-tensor field.*

4.2 Metrics

Definition 4.17. A *Riemannian metric* is a 2-tensor field $g \in \mathcal{T}^2_0(M)$ that is symmetric and positive definite.

Let $(E_1, \dots, E_n)$ be a local frame for TM and $(\phi^1, \dots, \phi^n)$ its dual coframe. A Riemannian metric can be written locally as:

$$g = g_{ij} \phi^i \otimes \phi^j.$$

Definition 4.18. The *symmetric product* of two 1-forms is $\omega\eta := \frac{1}{2}(\omega \otimes \eta + \eta \otimes \omega)$.

Since g_{ij} is symmetric, we can write in a coordinate frame:

$$g = g_{ij} dx^i dx^j.$$

4.3 Affine connections

Once again, I try to understand the link between the formal definition of affine connections and the intuition behind affine connections. I will rely on Amari & Nagaoka [5]. The idea is to do the reverse of what is done in Lee [6]. Instead of starting from the definition, we begin with the intuition and derive the definition from it.
I also want to compare with the directional derivative of vector fields in $\mathbb{R}^n$ as exposed by Tu [7].

4.3.1 Intuition

Let $\mathcal{S}$ be a manifold. Let $[x^i]$ be a global coordinate system on $\mathcal{S}$. Let p and q be two "neighboring" points, i.e., if $dx^i = x^i(q) - x^i(p)$ then we assume that the dx^i are sufficiently small such that $dx^i dx^j \approx 0$.
We are looking for a linear mapping $\Pi_{p,q}$ between the tangent spaces T_p and T_q. Since it is a linear mapping, it is sufficient to specify the image $\Pi_{p,q}\left((\partial_j)_p\right)$ of the basis vectors as a linear combination of the $(\partial_k)_q$.
In fact, we won't look for the image $\Pi_{p,q}\left((\partial_j)_p\right)$ but for the *difference between the "identity" and the image* $(\partial_j)_q - \Pi_{p,q}\left((\partial_j)_p\right)$. We assume that this difference can be expressed as a linear combination of the $(\partial_k)_q$ with *the coefficients themselves being a linear combination of the* dx^i:

$$(\partial_j)_q - \Pi_{p,q}\left((\partial_j)_p\right) = dx^i \left(\Gamma_{ij}^k\right)_p (\partial_k)_q.$$

For each original basis vector $(\partial_j)_p$, there is a $n \times n$ matrix $\left[\left(\Gamma_{ij}^k\right)_p\right]_{ik}$ which depends on p such that the k-coefficient of the difference is $\sum_{i=1}^n dx^i \left(\Gamma_{ij}^k\right)_p$. Moreover, we assume that the n^3 functions $\Gamma_{ij}^k : p \mapsto \left(\Gamma_{ij}^k\right)_p$ are all smooth. These functions define an **affine connection** and are called the **connection coefficients** of the affine connection with respect to the coordinate system $[x^i]$.
We can write the mapping $\Pi_{p,q}$ as:

$$\Pi_{p,q}\left((\partial_j)_p\right) = (\partial_j)_q - dx^i \left(\Gamma_{ij}^k\right)_p (\partial_k)_q.$$

4.3.2 Coordinate change

Let $[y^r]$ be another coordinate system with basis vector fields $\widetilde{\partial}_r$. We want to find the connection coeffcients in this new coordinate system. We have:

$$\begin{aligned}
\Pi_{p,q}\left((\widetilde{\partial}_s)_p\right) &= \Pi_{p,q}\left(\left(\frac{\partial x^j}{\partial y^s}\right)_p (\partial_j)_p\right), \\
&= \left(\frac{\partial x^j}{\partial y^s}\right)_p \Pi_{p,q}\left((\partial_j)_p\right), \\
&= \left(\frac{\partial x^j}{\partial y^s}\right)_p \left((\partial_j)_q - dx^i \left(\Gamma_{ij}^k\right)_p (\partial_k)_q\right).
\end{aligned}$$

Now, using[3] $dx^i = \left(\frac{\partial x^i}{\partial y^r}\right)_p dy^r$ and $(\partial_k)_q = \left(\frac{\partial y^t}{\partial x^k}\right)_q (\widetilde{\partial}_t)_q$, we get:

$$\Pi_{p,q}\left((\widetilde{\partial}_s)_p\right) = \left(\frac{\partial x^j}{\partial y^s}\right)_p \left((\partial_j)_q - \left(\frac{\partial x^i}{\partial y^r}\right)_p dy^r \left(\Gamma_{ij}^k\right)_p \left(\frac{\partial y^t}{\partial x^k}\right)_q (\widetilde{\partial}_t)_q\right).$$

Similarly, we have $\left(\frac{\partial y^t}{\partial x^k}\right)_q = \left(\frac{\partial y^t}{\partial x^k}\right)_p + \left(\frac{\partial^2 y^t}{\partial x^k \partial x^h}\right)_p dx^h$ and $\left(\frac{\partial x^j}{\partial y^s}\right)_p = \left(\frac{\partial x^j}{\partial y^s}\right)_q - \left(\frac{\partial^2 x^j}{\partial y^s \partial y^u}\right)_p dy^u$. Developing every term and using the facts that $dy^r dx^h \approx 0$ and $dy^u dy^r \approx 0$, we obtain:

$$\Pi_{p,q}\left((\widetilde{\partial}_s)_p\right) = \left(\frac{\partial x^j}{\partial y^s}\right)_q (\partial_j)_q - \left(\frac{\partial x^j}{\partial y^s}\right)_q \left(\frac{\partial x^i}{\partial y^r}\right)_p dy^r \left(\Gamma_{ij}^k\right)_p \left(\frac{\partial y^t}{\partial x^k}\right)_p (\widetilde{\partial}_t)_q - \left(\frac{\partial^2 x^j}{\partial y^s \partial y^u}\right)_p dy^u (\partial_j)_q.$$

[3] This is an informal statement. I guess it could be made rigorous using 1-forms and/or fiber bundle.

Using once again $(\widetilde{\partial}_s)_q = \left(\frac{\partial x^j}{\partial y^s}\right)_q (\partial_j)_q$, $\left(\frac{\partial x^j}{\partial y^s}\right)_q = \left(\frac{\partial x^j}{\partial y^s}\right)_p + \left(\frac{\partial^2 x^j}{\partial y^s \partial y^u}\right)_p dy^u$ and $(\partial_j)_q = \left(\frac{\partial y^t}{\partial x^j}\right)_q (\widetilde{\partial}_t)_q$, we obtain:

$$\Pi_{p,q}\left((\widetilde{\partial}_s)_p\right) = (\widetilde{\partial}_s)_q - \left(\frac{\partial x^j}{\partial y^s}\right)_p \left(\frac{\partial x^i}{\partial y^r}\right)_p dy^r \left(\Gamma_{ij}^k\right)_p \left(\frac{\partial y^t}{\partial x^k}\right)_p (\widetilde{\partial}_t)_q - \left(\frac{\partial^2 x^j}{\partial y^s \partial y^u}\right)_p dy^u \left(\frac{\partial y^t}{\partial x^j}\right)_q (\widetilde{\partial}_t)_q.$$

In the right-most term, we rename the index u by r and the index j by k. Then, we can factor:

$$\Pi_{p,q}\left((\widetilde{\partial}_s)_p\right) = (\widetilde{\partial}_s)_q - dy^r \left(\left(\frac{\partial x^j}{\partial y^s}\right)_p \left(\frac{\partial x^i}{\partial y^r}\right)_p \left(\Gamma_{ij}^k\right)_p + \left(\frac{\partial^2 x^k}{\partial y^s \partial y^r}\right)_p\right) \left(\frac{\partial y^t}{\partial x^k}\right)_p (\widetilde{\partial}_t)_q.$$

Thus, we have:

$$\widetilde{\Gamma}_{rs}^t = \left(\frac{\partial x^j}{\partial y^s}\frac{\partial x^i}{\partial y^r}\Gamma_{ij}^k + \frac{\partial^2 x^k}{\partial y^s \partial y^r}\right)\frac{\partial y^t}{\partial x^k}.$$

4.3.3 Parallel translation

Now, we assume that p and q are *not* neighboring points. Let γ be a curve such that $\gamma(0) = p$ and $\gamma(1) = q$. By connecting the tangent spaces of a sequence of neighboring points using Π, we can find a correspondence between T_p and T_q. However, this correspondence depends on the curve γ connecting p and q.

Define a **vector field along the curve** γ to be a mapping $X : t \mapsto X(t)$ such that $X(t) \in T_{\gamma(t)}$. If the tangent vectors are related by:

$$X(t + dt) = \Pi_{\gamma(t),\gamma(t+dt)}\left(X(t)\right),$$

where dt is an "infinitesimal", then we say that X is **parallel along** γ. In coordinates:

$$\begin{aligned} X(t + dt) &= \Pi_{\gamma(t),\gamma(t+dt)}\left(X^j(t)(\partial_j)_{\gamma(t)}\right), \\ &= X^j(t)\Pi_{\gamma(t),\gamma(t+dt)}\left((\partial_j)_{\gamma(t)}\right), \\ &= X^j(t)\left((\partial_j)_{\gamma(t+dt)} - dt\dot{\gamma}^i(t)\left(\Gamma_{ij}^k\right)_{\gamma(t)}(\partial_k)_{\gamma(t+dt)}\right), \end{aligned}$$

where we use the fact that $d\gamma^i = \gamma^i(t + dt) - \gamma^i(t) = dt\dot{\gamma}^i(t)$. By renaming the index j by k in the left term and by writing $X(t + dt)$ in coordinates, we obtain:

$$X^k(t + dt)(\partial_k)_{\gamma(t+dt)} = \Pi_{\gamma(t),\gamma(t+dt)}\left(X^j(t)(\partial_j)_{\gamma(t)}\right) = \left(X^k(t) - dt\dot{\gamma}^i(t)X^j(t)\left(\Gamma_{ij}^k\right)_{\gamma(t)}\right)(\partial_k)_{\gamma(t+dt)}. \quad (3)$$

Thus, for each k:

$$\frac{X^k(t + dt) - X^k(t)}{dt} + \dot{\gamma}^i(t)X^j(t)\left(\Gamma_{ij}^k\right)_{\gamma(t)} = 0,$$

which can be rewritten:

$$\dot{X}^k(t) + \dot{\gamma}^i(t)X^j(t)\left(\Gamma_{ij}^k\right)_{\gamma(t)} = 0. \quad (4)$$

This is an ordinary linear differential equation. Hence, for any $u \in T_p$, there exists a unique parallel vector field X along γ such that $X(0) = u$ (initial condition). Define the mapping Π_γ by $\Pi_\gamma(u) = X(1) \in T_q$. The mapping Π_γ is a linear isomorphism from T_p to T_q, called the **parallel translation along** γ.

4.3.4 Covariant derivative along a curve

Now, we can define the **covariant derivative** of a vector field X along a curve γ:

$$\frac{DX(t)}{dt} = \lim_{h\to 0}\frac{\Pi_{\gamma(t+h),\gamma(t)}\left(X(t + h)\right) - X(t)}{h}. \quad (5)$$

Informally, we can write:

$$DX(t) = \Pi_{\gamma(t+dt),\gamma(t)}\left(X(t + dt)\right) - X(t).$$

Using Equation 3, we have in coordinates:

$$\begin{aligned}\Pi_{\gamma(t+dt),\gamma(t)}\left(X(t+dt)\right) &= X^j(t+dt)\Pi_{\gamma(t+dt),\gamma(t)}\left((\partial_j)_{\gamma(t+dt)}\right),\\ &= \left(X^k(t+dt) + dt\dot{\gamma}^i(t)X^j(t+dt)\left(\Gamma^k_{ij}\right)_{\gamma(t+dt)}\right)(\partial_k)_{\gamma(t)},\end{aligned}$$

where we used $\gamma^i(t) - \gamma^i(t+dt) = -dt\dot{\gamma}^i(t)$. Using $X^j(t+dt) = X^j(t) + \dot{X}^j(t)dt$, $\left(\Gamma^k_{ij}\right)_{\gamma(t+dt)} = \left(\Gamma^k_{ij}\right)_{\gamma(t)} + dt\left(\frac{d\Gamma^k_{ij}}{dt}\right)_{\gamma(t)}$ and $(dt)^2 = 0$, we obtain:

$$\Pi_{\gamma(t+dt),\gamma(t)}\left(X(t+dt)\right) = \left(X^k(t+dt) + dt\dot{\gamma}^i(t)X^j(t)\left(\Gamma^k_{ij}\right)_{\gamma(t)}\right)(\partial_k)_{\gamma(t)}.$$

Thus, replacing in Equation 5:

$$\frac{DX(t)}{dt} = \left(\dot{X}^k(t) + \dot{\gamma}^i(t)X^j(t)\left(\Gamma^k_{ij}\right)_{\gamma(t)}\right)(\partial_k)_{\gamma(t)}. \tag{6}$$

The parallel translation condition in Equation 5 can be rewritten $\frac{DX}{dt} = 0$.

4.3.5 Covariant derivative along a tangent vector

Let X be a smooth vector field. Let Y_p be a tangent vector in T_p. Let γ be a curve such that $\gamma(0) = p$ and $\dot{\gamma}(0) = Y_p$. Let $Z(t)$ be a vector field along γ such that $Z(t) = X_{\gamma(t)}$. Using Equation 6, we define the **covariant derivative** of X along Y_p by:

$$\begin{aligned}\nabla_{Y_p}X &= \left(\dot{Z}^k(0) + \dot{\gamma}^i(0)Z^j(0)\left(\Gamma^k_{ij}\right)_{\gamma(0)}\right)(\partial_k)_{\gamma(0)},\\ &= \left(Y_pX^k + Y^i_pX^j_p\left(\Gamma^k_{ij}\right)_p\right)(\partial_k)_p,\end{aligned}$$

where we used the facts that $\dot{\gamma}^i(0) = Y^i_p$, $\dot{Z}^k(0) = \dot{\gamma}(0)X^k = Y_pX^k$ and $Z^j(0) = X^j_p$. We can rewrite this equation as:

$$\nabla_{Y_p}X = Y^i_p\left((\partial_i)_pX^k + X^j_p\left(\Gamma^k_{ij}\right)_p\right)(\partial_k)_p. \tag{7}$$

More generally, we can write:

$$\frac{DZ(t)}{dt} = \nabla_{\dot{\gamma}(t)}X.$$

Given two vector fields X and Y, we can define a new vector field called the **covariant derivative** of X with respect to Y by $(\nabla_YX)_p = \nabla_{Y_p}X$. In coordinates, we have:

$$\nabla_YX = Y^i\left(\partial_iX^k + X^j\Gamma^k_{ij}\right)\partial_k. \tag{8}$$

We also have:

$$\nabla_{\partial_i}\partial_j = \Gamma^k_{ij}\partial_k.$$

From there, we could check that the operator ∇ defined in Equation 8 verifies the properties that define affine connections, i.e., $\mathcal{F}$-linearity with respect to Y, $\mathbb{R}$-linearity and Leibniz rule with respect to X.
Two remarks:

- Affine connections forms an affine space, i.e., $\alpha\nabla + (1-\alpha)\nabla'$ is an affine connection for any $\alpha \in \mathbb{R}$.
- The difference of two affine connections is a tensor field (2-covariant and 1-contravariant).

4.3.6 Comparison with the directional derivative in $\mathbb{R}^n$ and with the Lie derivative

Affine connections can be seen as a generalization of the directional derivative of vector fields in $\mathbb{R}^n$ along a tangent vector.
Let $[\partial_k]$ be the canonical tangent vectors basis of $\mathbb{R}^n$. Let $X = X^k\partial_k$ be a vector field in $\mathbb{R}^n$ and Y_p be a tangent vector at a point $p \in \mathbb{R}^n$. If $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, then the directional derivative of f along Y_p is simply $D_{Y_p}f = Y_pf$.

Definition 4.19. The **directional derivative** of X along Y_p is:

$$D_{Y_p}X = (Y_pX^k)(\partial_k)_p.$$

Proposition 4.20 (Properties of the directional derivative)**.**

- $D_YX : p \mapsto D_{Y_p}X$ *is a vector field.*
- D_YX *is* $\mathcal{F}(\mathbb{R})$*-linear in* Y*.*
- D_YX *is* $\mathbb{R}$*-linear in* X *and verifies the Leibniz rule.*
- $D_XY - D_YX = [X, Y]$ *thus* ***the directional derivative is torsion-free*** *or* ***symmetric****.*
- $Z\langle X, Y\rangle = \langle D_ZX, Y\rangle + \langle X, D_ZY\rangle$ *thus* ***the directional derivative is compatible with the Euclidean metric on*** $\mathbb{R}^n$ *(Leibniz rule for inner product).*

Be careful! $D_YX \neq YX$. Indeed, we have:

- $D_YX = (YX^k)\partial_k$, which is a tangent vector.
- $YX = Y(X^k\partial_k)$, which is NOT a tangent vector in general.

Let $\mathcal{T}$ be the set of smooth vector fields on $\mathbb{R}^n$. $\mathcal{T}$ is a $\mathbb{R}$-vector space. With the Lie bracket of vector fields, it becomes a Lie albegra. Similarly, the ring of endomorphisms over vector fields $\hom(\mathcal{T})$ becomes a Lie algebra using the Lie bracket $[A, B] = A \circ B - B \circ A$.
Hence, the directional derivative $D : \mathcal{T} \to \hom(\mathcal{T})$ such that $X \mapsto D_X$ is a $\mathbb{R}$-linear map of Lie algebras. **Question: is it a Lie algebra homomorphism? i.e., do we have:**

$$[D_X, D_Y] = D_{[X,Y]}.$$

The answer is yes for the directional derivative in $\mathbb{R}^n$. But for an arbitrary affine connection ∇, the failure of ∇ to be a Lie algebra homomorphism *defines* the curvature of ∇.

Now, compare Equation 8 $\nabla_YX = \left(YX^k + Y^iX^j\Gamma^k_{ij}\right)\partial_k$ with the **Lie derivative**:

$$\mathcal{L}_Y(X) = [Y, X] = (YX^k - XY^k)\partial_k = (Y^i\partial_iX^k - X^j\partial_jY^k)\partial_k.$$

Both derivatives start with the directional derivative $(YX^k)\partial_k$. Then, they use a different second term. Both second terms depends on X^j. In the covariant derivative, X^j is summed with $Y^i\Gamma^k_{ij}\partial_k$. Here, Y^i is linearly transformed by Γ^k_{ij}. In the Lie derivative, X^j is summed with $\partial_jY^k\partial_k$. There, Y^k is differentiated along ∂_j.
Contrary to an affine connection, the Lie derivative $\mathcal{L}_Y(X)$ is *not* $\mathcal{F}(\mathbb{R})$-linear in Y. If ∇ is **symmetric** (such as the directional derivative in $\mathbb{R}^n$), we have:

$$\mathcal{L}_Y(X) = \nabla_YX - \nabla_XY.$$

4.3.7 Metric connection

Let ∇ be an affine connection on a Riemannian manifold with metric $g = \langle , \rangle$. We say that ∇ is a **metric connection** if:

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle.$$

In coordinates, it is simply:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i},$$

where $\Gamma_{ij,k} = \langle \nabla_{\partial_i} \partial_j, \partial_k\rangle = \Gamma^h_{ij} g_{hk}$.
Question: what is the link between this definition and "being metric"? Answer: *under a metric connection, the parallel translation preserves the inner product.* The parallel translation Π_γ along any curve γ becomes a metric isomorphism:

$$\langle \Pi_\gamma(X_p), \Pi_\gamma(Y_p)\rangle_q = \langle X_p, Y_p\rangle_p,$$

where $X_p, Y_p \in T_p$, $p = \gamma(0)$ and $q = \gamma(1)$.

4.4 Flatness

Let $\mathcal{S}$ be a manifold with an affine connection ∇.

Definition 4.21. Let $X \in \mathcal{T}(\mathcal{S})$ be a (smooth) vector field on $\mathcal{S}$.
X is **parallel** on $\mathcal{S}$ with respect to ∇ if it is parallel along any curve γ on $\mathcal{S}$.

We assume that $\mathcal{S}$ has dimension n and that there exists a global coordinate system. We note $[x^i] = (x^1, \ldots, x^n)$ such a coordinate system. We note ∂_i the corresponding basis vector fields. We note Γ^k_{ij} the corresponding connection coefficients.

Definition 4.22. $[x^i]$ is an **affine coordinate system** for ∇ if ∂_i is parallel on $\mathcal{S}$ for all i.

Affine coordinate systems are related by non-singular **affine transformations**.

Proposition 4.23. *$[x^i]$ is an affine coordinate system if and only if all connection coefficients Γ^k_{ij} are identically 0.*

Since Γ is not a tensor field, it can be identically 0 in some coordinate systems (the affine coordinate systems) and not in other coordinate systems.
Now, here is how Amari & Nagaoka [5] define flatness.

Definition 4.24. ∇ is **flat** if there exists an affine coordinate system for ∇.
We also say that $\mathcal{S}$ is flat with respect to ∇.

Let R and T be the curvature and torsion tensor fields.

Proposition 4.25. *If $[x^i]$ is an affine coordinate system, then $R = 0$ and $T = 0$.*
In other words, if ∇ is flat, then $R = 0$ and $T = 0$.

Since R and T are tensor fields, if they are identically 0 in some coordinate system, they must be identically 0 in all coordinate systems.

Definition 4.26. ∇ is **locally flat** if for all $p \in \mathcal{S}$, there exists a neighborhood U of p such that ∇ is flat on U (i.e., there is an affine coordinate system on U).

Proposition 4.27. *If $R = 0$ and $T = 0$, then ∇ is locally flat (but not necessarily flat).*

However, I don't know any example of a non-flat manifold with $R = 0$ and $T = 0$.

Proposition 4.28. *If ∇ is flat, then parallel translation does not depend on the curve connecting the two points.*

If ∇ is flat, then any vector fields with constant components is parallel.

Proposition 4.29. *If parallel translation does not depend on curve choice, then $R = 0$.*
In other words, if there are n linearly independent parallel vector fields on $\mathcal{S}$, then $R = 0$.

This following proposition is sometimes used to define flatness.

Proposition 4.30. *If $\mathcal{S}$ is simply connected, and if $R = 0$, then parallel translation does not depend on curve choice.*

However, there exist connections with $R = 0$ and $T \neq 0$. If $\mathcal{S}$ is simply connected, then parallel translation does not depend on curve choice, *but there is no affine coordinate system.* Such spaces are called *spaces of distant parallelism.*
Because of the symmetries of R and T, if $\mathcal{S}$ is 1-dimensional, then $R = 0$ and $T = 0$, thus $\mathcal{S}$ is (locally?) flat.
Let g be a Riemannian metric on $\mathcal{S}$. Now, we assume that ∇ is the Riemannian connection, i.e., the only connection which is both metric and torsion-free.

Definition 4.31. $[x^i]$ is a **Euclidean coordinate system** with respect to g if $\langle \partial_i, \partial_j \rangle = \delta_{ij}$.

Proposition 4.32. *The Riemannian connection ∇ is flat if and only if there exists a Euclidean coordinate system.*

Indeed, assume that ∇ is flat. We know that any metric connection preserves the inner product. Thus, if $[x^i]$ is an affine coordinate system, then $\langle \partial_i, \partial_j \rangle$ is constant. We can find an affine transformation such that this constant is δ_{ij}.

5 Geometrical Structures of a Family of Probability Distributions

Je m'appuie sur le livre de Amari [8].

5.1 Part I: Differential Geometry of Statistical Models

5.1.1 Manifold of statistical model

A **statistical model** is a parameterized family of probability distributions $S = \{p(x, \theta)\}$ where x is a random variable belonging to sample space X, and $p(x, \theta)$ is the probability density function of x (parameterized by θ) with respect to some common dominating measure $\mathbb{P}$ on X. Assume that $\theta = (\theta^1, \ldots, \theta^n) \in \Theta$ where $\Theta \subset \mathbb{R}^n$ is an open set.

When $p(x, \theta)$is sufficiently smooth in θ, we can introduce in S the structure of an n-dimensional manifold where θ is a coordinate system. In short, a n-dimensional manifold S is a Hausdorff space locally homeomorphic to $\mathbb{R}^n$. A **differentiable structure** is a maximum set of coordinate systems mutually connected by diffeomorphisms. A (local) differentiable structure is introduced (on an open set U) by defining a coordinate system. A global differentiable structure is introduced in S by giving an open cover $\{U_i\}$ and coordinate functions on each U_i such that the coordinate transformations on each $U_i \cap U_j$ are diffeomorphisms. A metrizable Hausdorff space is called a **differentiable manifold** when it has such global differentiable structure.

In the following, we assume that there always exists a global coordinate system in S (if not, we can say that the theory is valid on some open set U of S). We introduce a differentiable structure in a statistical model with $\phi[p(x, \theta)] = \theta$ such that θ are the coordinates (or the "name") of the distribution $p(x, \theta)$.

The following **regularity conditions** are required in this geometrical theory:

- All the distributions $p(x, \theta)$ have a common support i.e., $p(x, \theta) > 0$ for all $x \in X$ where X is the support.
- Let the log-likelihood be $l(x, \theta) = \log p(x, \theta)$. For every fixed θ, the n functions $X \mapsto \frac{\partial}{\partial \theta^i} l(x, \theta)$ are linearly independent.
- The moments of the random variables $\frac{\partial}{\partial \theta^i} l(x, \theta)$ (the scores) exists up to necessary orders.
- The partial derivatives and the integration with respect to the measure P can always be interchanged: $\frac{\partial}{\partial \theta^i} \int f(x, \theta) d\mathbb{P} = \int \frac{\partial}{\partial \theta^i} f(x, \theta) d\mathbb{P}$ for any function $f(x, \theta)$ treated in the following.

5.1.2 Tangent space

If θ is a coordinate system in S, then the $\frac{\partial}{\partial \theta^i}$ denoted by ∂_i are called the natural basis of the tangent space T_θ associated with the coordinate system θ.

There is a familiar representation of tangent vectors for manifold of statistical model S. Consider the log-likelihood $l(x,\theta) = \log p(x,\theta)$. Since the n partial derivatives $\partial_i l(x,\theta)$ are linearly independent (as functions of x with a fixed θ), we can construct a n-dimensional vector space:

$$T_\theta^{(1)} = \{A(x) : A(x) = A^i \partial_i l(x,\theta)\}.$$

Since x is a random variable, $T_\theta^{(1)}$ is the linear space of random variables spanned by $\partial_i l(x,\theta)$.

There is a natural isomorphism between T_θ (derivative operators) and $T_\theta^{(1)}$ (random variables) given by $\partial_i \iff \partial_i l(x,\theta)$. The space $T_\theta^{(1)}$ is called the **1-representation** of the tangent space.

Let $\mathbb{E}[.]$ be the expectation with respect to the distribution $p(x,\theta)$ i.e., $\mathbb{E}[f(x)] = \int f(x)p(x,\theta)d\mathbb{P}$. Since $\int p(x,\theta)d\mathbb{P} = 1$, we have that $0 = \partial_i \int p(x,\theta)d\mathbb{P} = \int \partial_i p(x,\theta)d\mathbb{P} = \int p(x,\theta)\partial_i l(x,\theta)d\mathbb{P} = \mathbb{E}[\partial_i l(x,\theta)]$. Hence, for any random variable $A(x) \in T_\theta^{(1)}$, we have $\mathbb{E}[A(x)] = 0$.

5.1.3 Riemannian metric and Fisher information

In a manifold of a statistical model, given the 1-representations $A(x)$ and $B(x)$, a natural inner product is defined by $\langle A, B\rangle = \mathbb{E}[A(x)B(x)]$. Since $\mathbb{E}[A(x)] = \mathbb{E}[B(x)] = 0$, the inner product is the covariance of $A(x)$ and $B(x)$. The **metric tensor** is defined by $g_{ij}(\theta) = \langle \partial_i, \partial_j \rangle$. It is a covariant tensor of order 2. Two tangent vectors are orthogonal when their 1-representations are uncorrelated. The (squared) length of a tangent vector is the variance of its 1-representation.

The matrix (g_{ij}) is the Fisher information matrix. Let (g^{ij}) be its inverse. Let $\hat{\theta}$ be an unbiased estimator of the parameter θ based on an observation x from the true distribution $p(x,\theta)$: $\mathbb{E}[\hat{\theta}] = \theta$. The **Cramér-Rao Theorem** states that the covariance of $\hat{\theta}$ is bounded by the inverse of the Fisher information matrix: $\mathrm{Cov}[\hat{\theta}^i, \hat{\theta}^j] \geq g^{ij}$ where $\geq$ means that the difference is a positive semi-definite matrix (it is not an inequality component-wise).

This bound is attained asymptotically in the following sense. Let $x_1, \ldots, x_N$ be N i.i.d. observations from $p(x,\theta)$. Then, there exists an estimator $\hat{\theta}_N$ based on these N observations such that $\mathrm{Cov}[\hat{\theta}_N^i, \hat{\theta}_N^j] \to g^{ij}/N$. The maximum likelihood estimator is such one. Moreover, the distribution of $\hat{\theta}_N$ tends to the normal distribution $\mathcal{N}(\theta, g^{ij}/N)$.

The **indistinguishability** or **non-separability** of two nearby distributions $p(x,\theta)$ and $p(x,\theta')$ may be measured by the probability that θ' is obtained as the estimated value $\hat{\theta}_N$. When N is large, this probability of confusion between $p(x,\theta)$ and $p(x,\theta')$ is proportional to their distance $ds^2 = g_{ij}(\theta)d\theta^i d\theta^j$ where $d\theta^i = \theta'^i - \theta^i$ is infinitesimally small as N tends to infinity. Hence, the distance ds^2 is based on the separability of two distributions by a large number of independent observations. This distance is also related to the power of testing one hypothesis $H_0 : p(x,\theta)$ against the other $H_1 : p(x,\theta')$ based on a large number of observations.

The metric tensor can be calculated with: $g_{ij}(\theta) = -\mathbb{E}[\partial_j \partial_j l(x,\theta)]$. This equation gives another interpretation of the metric tensor. The maximum likelihood estimator $\hat{\theta}$ satisfies $\partial_i l(x,\hat{\theta}) = 0$. We can expand the function $l(x,\theta)$ at $\hat{\theta}$:

$$l(x,\theta) = l(x,\hat{\theta}) + \frac{1}{2}\partial_i \partial_j l(x,\hat{\theta})(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j) + o(\|\hat{\theta} - \theta\|^3).$$

The maximum of $l(x,\theta)$ is attained at $\theta = \hat{\theta}$ and the term $\partial_i \partial_j l(x,\hat{\theta})$ shows how sharp is the peak of $l(x,\theta)$ at $\hat{\theta}$. The Fisher information is the negative of the expectation of this second derivative of $l(x,\theta)$.

5.1.4 Les connexions affines expliquées intuitivement

Soit S une variété différentielle.
Pour $\theta \in S$, on note T_θ l'espace tangent en θ.
On souhaite pouvoir comparer les vecteurs tangents appartenant à deux espaces tangents différents T_θ et

$T_{\theta'}$. L'idée consiste à établir une correspondance affine (et bijective) entre deux espaces tangents "proches", c'est à dire entre T_θ et $T_{\theta'}$ avec[4] $\theta' = \theta + d\theta$. Il est ensuite possible d'étendre cette correspondance entre deux espaces tangents éloignés via une courbe entre ces deux espaces. La correspondance dépendra alors de la courbe choisie.
Soit $m : T_{\theta+d\theta} \to T_\theta$ une application linéaire dépendant de $d\theta$, et se réduisant à l'identité quand $d\theta$ tend vers 0.
L'image $m(\partial'_j)$ du vecteur de base $\partial'_j = \partial_j(\theta + d\theta)$ est proche de $\partial_j = \partial_j(\theta)$. On considère la différence entre les deux :

$$\Delta\partial_j = m(\partial'_j) - \partial_j \in T_\theta.$$

$\Delta\partial_j$ est une fonction de $d\theta$. On considère l'expansion de $\Delta\partial_j$ au premier ordre au voisinage de 0 :

$$\Delta\partial_j = \Delta\partial_j(d\theta) = \Delta\partial_j(0) + J_{\Delta\partial_j}(0)d\theta = d\theta^i \Gamma^k_{ij}(\theta)\partial_k.$$

En effet, on a $\Delta\partial_j(0) = 0$. Pour j fixé, il faut voir $\Gamma^{\cdot}_{\cdot j}(\theta)$ comme la matrice jacobienne de $\Delta\partial_j$ en $d\theta = 0$, entre les bases $\{d\Theta^i\}$ et $\{\partial_k\}$. Attention à ne pas confondre $d\Theta^i$ qui le vecteur de la base duale associé à ∂_i, et $d\theta^i$ qui est la i-ème coordonnée de $d\theta$. Donc, si on néglige les termes d'ordre plus grand que 1, $\Delta\partial_j$ se réduit à une application linéaire entre T^*_θ et T_θ dont la matrice dans les bases canoniques du système de coordonnées choisi est $\Gamma^{\cdot}_{\cdot j}(\theta)$. Finalement, l'application m est déterminé par n^3 fonctions $\Gamma^k_{ij}(\theta)$.
Maintenant, on considère un vecteur quelconque $A^i\partial'_i \in T_{\theta+d\theta}$. Son image par m est :

$$\begin{aligned} m(A^i\partial'_i) &= A^i m(\partial'_i) \\ &= A^i(\partial_i + \Delta\partial_i) \\ &= (A^i\partial_i + A^i d\theta^j \Gamma^k_{ji}(\theta)\partial_k) \\ &= (A^k + d\theta^i \Gamma^k_{ij}(\theta)A^k)\partial_k \in T_\theta. \end{aligned}$$

La première égalité est due à la linéarité de m. Dans la dernière égalité, on renomme les indices de la façon suivante : i devient k dans le premier terme et on échange i et j dans le second terme. On obtient ainsi une correspondance entre les vecteurs de T_θ et $T_{\theta+d\theta}$.
On souhaite cependant obtenir une correspondance affine, et pas seulement linéaire (sûrement pour obtenir une application différente pour des $d\theta$ différents ?). Pour cela, on suppose que l'origine de $T_{\theta+d\theta}$ est envoyée sur le point $d\theta^i\partial_i$ de T_θ. Ainsi, notre correspondance affine est :

$$\tilde{m}(A^i\partial'_i) = (A^k + d\theta^k + d\theta^i A^j \Gamma^k_{ij})\partial_k.$$

La différence $\Delta\partial_j$ peut être vue comme le changement intrinsèque du j-ème vecteur de base $\partial_j(\theta)$ quand le point change de θ à $\theta + d\theta$. Si $d\theta = d\Theta^i$, alors cela revient à dire que le point θ change dans la direction de ∂_i. Dans ce cas, on peut appeler $\nabla_{\partial_i}\partial_j$ le taux de changement intrinsèque de ∂_j quand on se déplace dans la direction de ∂_i. On a :

$$\nabla_{\partial_i}\partial_j = \Gamma^k_{ij}(\theta)\partial_k(\theta),$$

qui correspond bien à $\Delta\partial_j$ lorsque $d\theta = d\Theta^i$. Le champ de vecteurs $\nabla_{\partial_i}\partial_j$ n'est autre que la dérivée covariante du champ de vecteurs ∂_j le long de ∂_i. Dans la suite, on pourra définir la connexion affine via l'expression covariante de ses coefficients : $\Gamma_{ijk}(\theta) = \Gamma^m_{ij}(\theta)g_{mk}(\theta)$. On retrouve les coefficients contravariants via : $\Gamma^k_{ij} = g^{km}\Gamma_{ijm}$.

Soit $T(S)$ l'ensemble des champs de vecteurs lisses de S. Voyons maintenant la véritable définition d'une connexion affine. Une connexion affine sur S est une dérivée covariante $\nabla : T(S) \times T(S) \to T(S)$ vérifiant les axiomes suivants. Soient $A, A', B, B' \in T(S)$ et $f : S \to \mathbb{R}$ une fonction lisse.

$$\begin{aligned} \nabla_{fA+A'}B &= f\nabla_A B + \nabla_{A'}B, \\ \nabla_A(B + B') &= \nabla_A B + \nabla_A B', \\ \nabla_A(fB) &= (Af)B + f\nabla_A B. \end{aligned}$$

[4] Peut-on voir $d\theta$ comme une 1-forme ?

Étant donné les coefficients de la connexion, on peut calculer la dérivée covariante $\nabla_A B$, c'est à dire le taux de changement intrinsèque du champ de vecteurs $B^i\partial_i$ lorsqu'on se déplace le long de $A^i\partial_i$:

$$\begin{aligned}\nabla_A B &= \nabla_A(B^j\partial_j),\\ &= (AB^j)\partial_j + B^j\nabla_A\partial_j,\\ &= (A^i\partial_i B^j)\partial_j + B^j\nabla_{A^i\partial_i}\partial_j,\\ &= (A^i\partial_i B^j)\partial_j + B^j A^i\Gamma^k_{ij}\partial_k,\\ &= (A^i\partial_i B^k + A^iB^j\Gamma^k_{ij})\partial_k.\end{aligned}$$

Remarquez qu'on peut faire le lien avec l'explication intuitive présentée ci-dessus :

$$\nabla_A B = \Delta B(A) = \tilde{m}(B(\theta + A)) - B(\theta),$$

où $\theta + A$ signifie qu'on va de θ à n'importe quel autre point en suivant le flux de A. Ainsi, le terme $d\theta^i$ apparaissant dans la définition de $\tilde{m}(\cdot)$ est remplacé par A^i (attention, le A de la définition correspond ici au B), et "l'ordonnée à l'origine" $d\theta^k\partial_k$ est remplacée par $A^i\partial_i B^k\partial_k$ (et je n'ai pas la moindre idée de comment interpréter ça ...).

Notez que Γ^k_{ij} n'est pas un tenseur. Par exemple, si $\Gamma^k_{ij}(\theta) = 0$ pour tout θ, alors il peut y avoir un autre système de coordonnées ξ pour lequel Γ^k_{ij} n'est pas identiquement nul (sauf si θ et ξ sont liés par une relation linéaire).

Remarque importante. Quelle est la différence entre la dérivée covariante et la dérivée de Lie ? La dérivée covariante permet de définir une dérivée directionnelle d'un champ de vecteur, i.e., la valeur de $(\nabla_X Y)_p$ ne dépend que de la valeur de X en p. En revanche, la valeur de $(\mathcal{L}_X Y)_p = [X, Y]_p$ depend des valeurs de X dans un voisinage de p. On ne peut pas définir de dérivée directionnelle avec la dérivée de Lie.

Voici un exemple illustrant ce point. On se place dans $\mathbb{R}^2$. Il existe des champs de vecteurs V et W tels que $V = W = \partial_1$ *sur l'axe* x^1 mais tels que $[V, \partial_2] \neq [W, \partial_2]$ *sur l'axe* x^1.

Pourquoi cette différence entre dérivée covariante et dérivée de Lie ? Cela est dû à la façon dont chacune de ces dérivées résout le problème de additionner/soustraire des vecteurs tangents issus d'espaces tangents différents. La dérivée de Lie résout se problème en utilisant le flot d'un champ de vecteurs pour ramener un des vecteurs dans l'espace tangent de l'autre. La dérivée covariante se contente d'utiliser la connexion affine ∇ (qui n'est pas du tout unique) pour relier les deux espaces tangents.

La dérivée de Lie peut donc être définie directement à partir de la structure différentielle d'une variété différentielle. En revanche, la dérivée covariante a besoin d'un objet supplémentaire, à savoir une connexion affine.

5.1.5 Statistical α-connection

Our goal is to introduce an affine connection in the space S of a statistical model such that it represents the intrinsic properties of the family of probability distributions. Consider the tangent space $T_{\theta+d\theta}$. The 1-representation of its natural basis $\partial_j(\theta + d\theta)$ is $\partial_j l(x, \theta + d\theta) = \partial_j l(x, \theta) + \partial_i\partial_j l(x, \theta)d\theta^i$. In order to obtain an affine connection, it is necessary to find a way of mapping it to the space $T^{(1)}_\theta$. The expectation of $\partial_i\partial_j l(x, \theta)$ does not vanish at θ, so $\partial_j(x, \theta + d\theta)$ does not belong to $T^{(1)}_\theta$.

Let us modify $\partial_j(x, \theta + d\theta) \in T^{(1)}_{\theta+d\theta}$ so that its expectation vanishes at θ, by adding $g_{ij}(\theta)d\theta^i$ (since $g_{ij}(\theta) = -\mathbb{E}[\partial_j\partial_j l(x, \theta)]$). In general, the random variable $\partial_j l(x, \theta) + (\partial_i\partial_j l(x, \theta) + g_{ij}(\theta))d\theta^i$ does not belong to $T^{(1)}_\theta$ so we project it to $T^{(1)}_\theta$. By this projection, a linear correspondence between T_θ and $T_{\theta+d\theta}$ is established:

$$m(\partial_j(\theta + d\theta)) = \pi(\partial_j l(x, \theta) + (\partial_i\partial_j l(x, \theta) + g_{ij}(\theta))d\theta^i) = \partial_j l(x, \theta) + \pi((\partial_i\partial_j l(x, \theta) + g_{ij}(\theta))d\theta^i),$$

where π is the (orthogonal?) projection to $T^{(1)}_\theta \cong T_\theta$ defined for any random variable f (in L^2?) by[5]:

$$\pi(f) = g^{mk}(\theta)\mathbb{E}[f\partial_k l(x, \theta)]\partial_m l(x, \theta),$$

[5]Check that $\pi(\partial_j l(x, \theta)) = g^{mk}(\theta)g_{jk}(\theta)\partial_m l(x, \theta) = \partial_j l(x, \theta)$ so that elements of $T^{(1)}_\theta$ are projected to themselves.

where the expectation is taken according to θ. Using the same notations as in the previous paragraph, we have

$$\begin{aligned}
\Delta\partial_j &= m(\partial_j l(x,\theta+d\theta)) - \partial_j l(x,\theta),\\
&= \pi((\partial_i\partial_j l(x,\theta) + g_{ij}(\theta))d\theta^i),\ \text{(definition of } m(\cdot))\\
&= g^{mk}(\theta)\mathbb{E}[(\partial_i\partial_j l(x,\theta) + g_{ij}(\theta))d\theta^i\partial_k l(x,\theta)]\partial_m l(x,\theta),\ \text{(definition of } \pi(\cdot))\\
&= d\theta^i g^{mk}(\theta)\mathbb{E}[(\partial_i\partial_j l(x,\theta) + g_{ij}(\theta))\partial_k l(x,\theta)]\partial_m l(x,\theta),\ (d\theta \text{ is not random})\\
&= d\theta^i \Gamma^m_{ij}(\theta)\partial_m l(x,\theta),\ \text{(definition of } \Gamma^m_{ij})\\
&= d\theta^i g^{mk}(\theta)\Gamma_{ijk}(\theta)\partial_m l(x,\theta)\ \text{(definition of } \Gamma_{ijk}).
\end{aligned}$$

Hence, the resultant affine connection is:

$$\begin{aligned}
\Gamma_{ijk}(\theta) &= \mathbb{E}[(\partial_i\partial_j l(x,\theta) + g_{ij}(\theta))\partial_k l(x,\theta)],\\
&= \mathbb{E}[\partial_i\partial_j l(x,\theta)\partial_k l(x,\theta)],
\end{aligned}$$

using the relation $\mathbb{E}[\partial_k l(x,\theta)] = 0$. Notice that this connection can be obtained by directly projecting $\partial_j l(x,\theta+d\theta) = \partial_j l(x,\theta) + \partial_i\partial_j l(x,\theta)d\theta^i$ to $T^{(1)}_\theta$ since we have:

$$\begin{aligned}
\pi(\partial_j l(x,\theta+d\theta)) - \partial_j l(x,\theta) &= \pi(\partial_i\partial_j l(x,\theta)d\theta^i),\\
&= g^{mk}(\theta)\mathbb{E}[\partial_i\partial_j l(x,\theta)d\theta^i\partial_k l(x,\theta)]\partial_m l(x,\theta),\\
&= d\theta^i g^{mk}(\theta)\mathbb{E}[\partial_i\partial_j l(x,\theta)\partial_k l(x,\theta)]\partial_m l(x,\theta).
\end{aligned}$$

There is another possibility of modifying the random variable using the fact that the expectation at θ of $\partial_i\partial_j l(x,\theta) + \partial_i l(x,\theta)\partial_j l(x,\theta)$ also vanishes. The resultant affine connection is :

$$\Gamma_{ijk}(\theta) = \mathbb{E}[(\partial_i\partial_j l(x,\theta) + \partial_i l(x,\theta)\partial_j l(x,\theta))\partial_k l(x,\theta)].$$

These two definitions suggest that an infinite number of affine connections can be introduced by using a weighted mean. Let α be a scalar parameter. Then, we can modify $\partial_j l(x,\theta+d\theta) = \partial_j l(x,\theta) + \partial_i\partial_j l(x,\theta)d\theta^i$ into $\partial_i\partial_j l(x,\theta) + \frac{1+\alpha}{2}g_{ij}(\theta) + \frac{1-\alpha}{2}\partial_i l(x,\theta)\partial_j l(x,\theta)$. The corresponding affine connection coefficients are :

$$\Gamma^{(\alpha)}_{ijk}(\theta) = \mathbb{E}[(\partial_i\partial_j l(x,\theta) + \frac{1-\alpha}{2}\partial_i l(x,\theta)\partial_j l(x,\theta))\partial_k l(x,\theta)].$$

This is called the **α-connection**. We obtain the two preceding cases when $\alpha = \pm 1$.

Define the following third-order tensor:

$$T_{ijk}(\theta) = \mathbb{E}[\partial_i l(x,\theta)\partial_j l(x,\theta)\partial_k l(x,\theta)],$$

Then, the α-connection can be written as:

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(1)}_{ijk} + \frac{1-\alpha}{2}T_{ijk}.$$

We have introduced a one-parameter family of affine connections. But which is the true connection in S? This question is meaningless and here is why.

Let $p_1(x)$ and $p_2(x)$ be two density functions, and define

$$p(x,t) = (1-t)p_1(x) + tp_2(x),$$

called the **mixture family**. For $l(x,t) = \log p(x,t)$, we can check that $\partial_t\partial_t l(x,t) + (\partial_t l(x,t))^2 = 0$. If we use $\alpha = -1$, then $\partial_t l(x,t+dt) \in T^{(1)}_{t+dt}$ is mapped to $\partial_t l(x,t) \in T^{(1)}_t$. Hence, the -1-connection manifests the criterion that the mixture families should be understood as straight models. This point of view can be extended to the mixture of $n+1$ distributions, giving a n-dimensional -1-flat manifold.

There is another way to connect two distributions by taking the linear combination of $l_i(x) = \log p_i(x)$:

$$p(x,t) = \exp((1-t)l_1(x) + tl_2(x) - c(t)),$$

where $c(t)$ is the normalization factor. This is the **exponential family**. If we differentiate the relation $\exp(c(t)) = \int \exp((1-t)l_1(x) + tl_2(x))d\mathbb{P}$, we obtain $\partial_t c(t) = \mathbb{E}[l_2(x) - l_1(x)]$ where the expectation is taken with respect to t. Thus, we have that $\partial_t l(x,t) = l_2(x) - l_1(x) - \mathbb{E}[l_2(x) - l_1(x)]$. Then, $\partial_t \partial_t l(x,t) = -\mathbb{E}[(l_2(x) - l_1(x))(l_2(x) - l_1(x) - \mathbb{E}[l_2(x) - l_1(x)])] = -\mathrm{Var}[l_2(x) - l_1(x)] = -\mathrm{Var}[\partial_t l(x,t)] = -g_{tt}$, where g_{tt} is the Fisher information. Hence, if we use $\alpha = 1$, then $\partial_t l(x, t+dt) \in T^{(1)}_{t+dt}$ is mapped to $\partial_t l(x,t) \in T^{(1)}_t$. The 1-connection is based on the criterion which regards the exponential family as straight line.

Moreover, we can check that the Levi-Civita connection is nothing but the 0-connection:

$$\Gamma^{(0)}_{ijk} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}).$$

The α-connection are in general non-metric except for the case $\alpha = 0$. A connection is said to be metric if the parallel shift of a vector does not change its length[6].

5.1.6 Curvature and torsion

Consider a manifold with an affine connection. The tangent space T_θ is mapped by an affine transformation to the tangent space $T_{\theta'}$. This affine transformation depends on the curve connecting the two points. If you consider a loop, then a tangent space T_θ will be mapped to itself along the loop. This map is not necessary the identity map: the origin of T_θ is mapped to another point and the direction of a vector also changes. These discrepancies represent characteristic structures of the manifold and are described by the **torsion** and **curvature** of the manifold.

The torsion is a bilinear mapping from $T(S) \times T(S)$ to $T(S)$ induced by the affine connection. Hence, it is a tensor of order three (covariant order 2 and contravariant order 1). Let $A, B \in T(S)$. The torsion is defined as:

$$S(A,B) = \nabla_A B - \nabla_B A - (AB - BA).$$

What is the intuition for the torsion? Let ϵ be a small quantity. If we shift T_θ in parallel along the parallelogram composed of the infinitesimal vectors ϵA and ϵB, then the change is the position of the origin of T_θ will be $\epsilon^2 S(A,B)$. Given a coordinate system θ, the torsion is represented by the torsion tensor:

$$S_{ijk}(\theta) = \langle S(\partial_i, \partial_j), \partial_k \rangle.$$

Since the partial derivatives commute, we have

$$S_{ijk} = \Gamma_{ijk} - \Gamma_{jik},$$

so S_{ijk} is antisymmetric with respect to i and j. Since the coefficients of the α-connection are symmetric with respect to i and j, the torsion tensor identically vanishes for any α-connection. The manifold of a statistical model is torsion-free.

The **Riemann-Christoffel curvature** R (or curvature endomorphism) is a trilinear mapping from $T(S) \times T(S) \times T(S)$ to $T(S)$ induced by the affine connection. It is a tensor of order four (covariant of order three, contravariant of order one). Let $A, B, C \in T(S)$, the map is defined by:

$$\begin{aligned} R(A,B,C) &= \nabla_A \nabla_B C - \nabla_B \nabla_A C - \nabla_{AB-BA} C, \\ &= [\nabla_A, \nabla_B] C - \nabla_{[A,B]} C. \end{aligned}$$

What is the intuition for the curvature endomorphism? If we shift C in parallel along the parallelogram composed of the infinitesimal vectors ϵA and ϵB, then $\epsilon^2 R(A,B,C)$ represents the change in vector C. Given a coordinate system θ, the curvature is represented by the Riemann-Christoffel curvature tensor

$$\begin{aligned} R_{ijkm} &= \langle R(\partial_i, \partial_j, \partial_k), \partial_m \rangle, \\ &= (\partial_i \Gamma^s_{jk} - \partial_j \Gamma^s_{ik}) g_{sm} + (\Gamma_{irm} \Gamma^r_{jk} - \Gamma_{jrm} \Gamma^r_{ik}). \end{aligned}$$

[6]Does this property corresponds to being compatible with the metric or symmetric (i.e., torsion-free)? I think it corresponds to being compatible with the metric but I cannot link this with the parallel shift ...

A space with an affine connection is said to be **flat** when the Riemann-Christoffel curvature vanishes identically. In a flat space, the affine transformation between two tangent spaces does not depend on the curve (and it is the identity along a loop). Hence, we can construct a vector field A such that, for any θ, $A(\theta)$ is obtained from $A(\theta_0)$ by its parallel shift. Such a vector field is called a **parallel vector field**.

In a torsion-free manifold, $R_{ijkm} = 0$ identically if and only if there exists a coordinate system θ such that $\Gamma_{ijk}(\theta) = 0$ identically. In this case, the basis vector fields of this coordinate system are parallel vector fields (i.e., $\partial_i(\theta) \in T_\theta$ is sent to $\partial_i(\theta') \in T_{\theta'}$). Such a coordinate system is said to be **affine**. In a flat manifold, all the affine coordinate systems are connected by affine transformations. In an affine coordinate system, the coordinate curves are geodesics.

If S is not flat, then there exists no affine coordinate systems. However, for any point θ_0, there exists a coordinate system such that the coefficients of the affine connection as well as their derivatives vanish at θ_0. Such a coordinate system is called a **normal coordinate system** at θ_0.

When the α-Riemann-Christoffel curvature vanishes identically, the space is said to be α-flat. We will see later that a manifold is α-flat if and only if it is $-\alpha$-flat. This is an example of the dualistic structure of a statistical manifold with the α-connection.

An α-flat manifold is not necessarily Euclidean except if $\alpha = 0$. In this later case, there exists an orthonormal Cartesian coordinate system such that $g_{ij}(\theta) = \delta_{ij}$.

The **scalar curvature** κ is defined by[7]

$$\kappa = \frac{1}{n(n-1)} R_{ijkm} g^{im} g^{jk}.$$

It takes the same value in any coordinate systems.

The space S of unidimensional normal distributions has constant curvature for every α (the scalar curvature does not depend of the point θ). Moreover, S is ± 1-flat. Indeed, the coordinate system $\xi^1 = \mu, \xi^2 = \mu^2 + \sigma^2$ is -1-affine, it is called the **expectation** parameter. The coordinate system $\zeta^1 = \mu/\sigma^2, \zeta^2 = -1/(2\sigma^2)$ is 1-affine, it is called the **natural** or **canonical** parameter. We will show later that S belongs to the exponential type family which is ± 1-flat.

5.1.7 Imbedding and submanifold

A **regular submanifold** is a subset M of a manifold S such that M has a structure of a manifold (of dimension m strictly less than the dimension n of S) and whose topology is the relative topology[8] induced in S. Let θ be a coordinate system of S and let u be a coordinate system of M. A point $u \in M$ has coordinates $\theta = \theta(u)$ in S. The equation $\theta = \theta(u)$ gives a parametric representation of M in S. This equation defines an injection from M to S. The mapping $\theta = \theta(u)$ is an **imbedding** of M in S if and only if it is smooth and its Jacobian matrix has full rank. When such an imbedding exists, we simply call M a submanifold of S.

Let $u \in M$. The tangent space $T_u(M)$ of M at u is spanned by m vectors $\partial_a = \partial/\partial u^a$. Since M is imbedded in S, the tangent vector of a curve in M is also tangent to the same curve in S. Therefore, $T_u(M)$ is a subspace of the tangent space $T_{\theta(u)}(S)$ of S at $\theta = \theta(u)$. In the following, we denote $T_u(S) = T_{\theta(u)}(S)$ for short. Denote the Jacobian matrix with n rows and m columns by

$$B_a^i(u) = \frac{\partial \theta^i}{\partial u^a}.$$

Then, we have $\partial_a = B_a^i(u)\partial_i$ (the vectors ∂_a are seen as elements of $T_u(S)$). Since the rank of (B_a^i) is full (since it is an imbedding), the ∂_a are linearly independent and span the subspace $T_u(M)$ in $T_u(S)$.

Now, assume that S is equipped with a Riemannian metric g_{ij} and an affine connection Γ_{ijk}. We can induce these structures in M. The induced metric tensor is[9]:

$$g_{ab}(u) = \langle \partial_a, \partial_b \rangle = B_a^i B_b^j \langle \partial_i, \partial_j \rangle = B_a^i B_b^j g_{ij}(u),$$

[7] How this relates with the definition from Lee's textbook?

[8] Meaning that the topology on M is the coarsest that makes the inclusion map continuous(?). Or more simply, it is the topology induced by S when seeing M as a subset of S.

[9] Notice that we can rewrite this with the pullback operator "$g_{ab} = \theta^* g_{ij}$" where θ is the inclusion map.

where $g_{ij}(u) = g_{ij}(\theta(u))$. The covariant derivative in the enveloping manifold S is[10]:

$$\nabla_{\partial_a}\partial_b = (B_a^i B_b^j \Gamma_{ij}^k + \partial_a B_b^k)\partial_k.$$

Notice that the resultant vector $\nabla_{\partial_a}\partial_b \in T_u(S)$ does not necessarily belong to $T_u(M)$, **because the intrinsic change in the tangent vector ∂_b may have a component orthogonal to $T_u(M)$**. This component shows how M is curved in S. Now, we want to define a proper covariant derivative $\nabla'_{\partial_a}\partial_b$ that belongs to $T_u(M)$. This is simply done by projecting $\nabla_{\partial_a}\partial_b$ to $T_u(M)$ by discarding the component orthogonal to $T_u(M)$. Hence, the induced affine connection in M is:

$$\Gamma_{abc}(u) = \langle \nabla'_{\partial_a}\partial_b, \partial_c\rangle = \langle \nabla_{\partial_a}\partial_b, \partial_c\rangle = B_a^i B_b^j B_c^k \Gamma_{ijk} + (\partial_a B_b^k) B_c^m g_{km}.$$

In order to denote the curvature of M in S, we choose $n-m$ vectors ∂_k in $T_u(S)$ such that:

- The set $\{\partial_a, \partial_k\}$ is a basis of $T_u(S)$ at every $u \in M$.
- The ∂_k are orthogonal to $T_u(M)$ i.e., $\langle \partial_a, \partial_k\rangle = 0$.
- The ∂_k defined on M are smooth vector fields in $T(S)$.

The curvature of M in S is measured by intrinsic changes in the directions of $T_u(M)$ in $T_u(S)$ as the point u moves in M. These intrinsic changes can be measured by the orthogonal components of the covariant derivatives $\nabla_{\partial_a}\partial_b$ in S. Using the orthogonal vectors ∂_k, the curvature of M in S is defined by the **Euler-Schouten curvature tensor** (or **imbedding curvature**):

$$H_{abk}(u) = \langle \nabla_{\partial_a}\partial_b, \partial_k\rangle.$$

The imbedding curvature can equivalently be defined as the bilinear mapping[11] H from $T(M)\times T(M)$ to the orthogonal complement of $T(M)$ in $T(S)$ which assigns to $A, B \in T(M)$ the component $H(A,B)$ of $\nabla_A B$ orthogonal to $T(M)$.

The Riemann-Christoffel curvature of M is related to the Euler-Schouten imbedding curvature of M in S and to the Riemann-Christoffel curvature of S by:

$$R_{abcd}^{(\alpha)} = B_a^i B_b^j B_c^k B_d^m R_{ijkm}^{(\alpha)} + g^{\kappa\lambda}(H_{ad\kappa}^{(-\alpha)} H_{bc\lambda}^{(\alpha)} - H_{ac\kappa}^{(\alpha)} H_{bd\lambda}^{(-\alpha)}.$$

When S is flat ($R_{ijkm} = 0$), the curvature R_{abcd} is determined from the imbedding curvature H. When M is "flat in a flat S", then $H_{ab\kappa} = 0$ and $R_{ijkm} = 0$ so that M is itself flat ($R_{abcd} = 0$). However, the fact that M is flat does not imply that M is a flat submanifold of S (see the cylinder in the Euclidean 3-space for example). The Riemann-Christoffel curvature is an intrinsic characteristic of M (not depending on how it is imbedded in S), while the Euler-Schouten curvature represents the manner how M is imbedded in S. There are different geometrical characteristics (although they are related).

A statiscial model $M = \{q(x,u)\}$ is a submodel of another statistical model $S = \{p(x,\theta)\}$ when there exists an injection $\theta(u)$ such that $q(x,u) = p(x,\theta(u))$. When the mapping $\theta(u)$ is smooth having a full rank Jacobian matrix, the statistical manifold M is a smooth submanifold imbedded in S. We can introduce the metric and the α-connection directly in M (with their definitions) without referring to the geometric structures of S (both approaches coincide).

5.1.8 Family of ancillary submanifolds

Let M be a submanifold of dimension m imbedded in S of dimension n. For each point $u \in M$, attach a $(n-m)$-dimensional submanifold $A(u)$ of S which is transverse to M at $\theta(u)$. Assume that the family $A = \{A(u) : u \in M\}$ fills S up smoothly (or at least a neighborhood of M). What that means is that their exists a coordinate system v on each $A(u)$ such that the pair $\xi = (u,v)$ is a coordinate system of S (or at least of a neighborhood of M). A is called a **family of ancillary submanifolds** rigging M. It is

[10] I skip the computation that only uses the basic properties of ∇.

[11] This is the second fundamental form.

convenient to fix the origin $v = 0$ of each $A(u)$ at the point $\theta(u)$ such that $\theta = \theta(u) = \theta(u, 0)$ is a parametric representation of M and $v = 0$ is another one.

A **foliation** of a manifold S is a partitioning $S = \cup_{u \in M} A(u)$ of the manifold S into submanifolds $A(u)$ of dimension $n - m$. A foliation defines an ancillary family when $A(u)$ is transverse to M at u. When M is a regular submanifold, there always exists a neighborhood U_M of M such that an ancillary family A is defined in U_M. This U_M is called a **tubular neighborhood**. An ancillary family is a local foliation of S (i.e., a foliation of U_M). We have that $T_u(S) = T_u(M) \oplus T_u(A)$.

Denote by (∂_a) the basis of $T_u(M)$ (Latin letters), and (∂_κ) the basis of $T_u(A)$ (Greek letters). If the "mixed" part of the metric is zero on M i.e., $g_{a\kappa}(u) = \langle \partial_a, \partial_\kappa \rangle = 0$, then $T_u(A)$ and $T_u(M)$ are orthogonal complement of each other. In this case, $A = \{A(u)\}$ is called an orthogonal ancillary family. This property is independent of the choice of the coordinate system v in $A(u)$.

The mixed parts of the connection $\Gamma_{\kappa\lambda a}(u) = H_{\kappa\lambda a}(u)$ and $\Gamma_{ab\kappa}(u) = H_{ab\kappa}(u)$ gives respectively the imbedding curvature of $A(u)$ in S at $\theta(u)$ and the imbedding curvature of M in S at $\theta(u)$.

It is always possible to choose a coordinate system v in each $A(u)$ such that $\{\partial_\kappa\}$ forms an orthonormal basis vector in $A(u)$ on M and at the same time v is the normal coordinate[12] of $A(u)$ at the points $v = 0$ on M. However, remark that the α-normal coordinate v is not necessary normal for another $\alpha' \neq \alpha$.

5.2 Part II: α-Divergence and α-Projection in Statistical Manifold

5.2.1 α-representation

Define the one-parameter family of functions $F_\alpha(p)$ by:

$$F_\alpha(p) = \begin{cases} \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}} & , \alpha \neq 1, \\ \log p & , \alpha = 1. \end{cases}$$

The derivative $F'_\alpha(p) = p^{-(1+\alpha)/2}$ is a homogeneous function[13] in p of degree $-(1+\alpha)/2$.

The α**-representation** of the density function $p(x, \theta)$ is:

$$l_\alpha(x, \theta) = F_\alpha(p(x, \theta)).$$

Notice that the 1-representation is the logarithm $l(x, \theta)$ and the -1-representation is $p(x, \theta)$ itself.

Let $T_\theta^{(\alpha)}$ be the vector space spanned by n linearly independent functions $\partial_i l_\alpha(x, \theta)$. We have the natural isomorphism $T_\theta \cong T_\theta^{(\alpha)}$, so $T_\theta^{(\alpha)}$ is called the α-representation of the tangent space T_θ. Denote by $Al_\alpha = A^i \partial_i l_\alpha(x, \theta)$ the α-representation of a vector $A = A^i \partial_i$.

Define the α**-expectation** of a random variable $f(x)$ by:

$$\mathbb{E}_\alpha[f(x)] = \int p(x, \theta)^\alpha f(x) d\mathbb{P}.$$

Since $\partial_i l_\alpha(x, \theta) = p(x, \theta)^{(1-\alpha)/2} \partial_i l(x, \theta)$ (chain rule), then the inner product of two vectors A and B is[14]:

$$\langle A, B \rangle := \mathbb{E}[(Al)(Bl)] = \mathbb{E}_\alpha[(Al_\alpha)(Bl_\alpha)].$$

Moreover, since $(\partial_i l_\alpha)(\partial_j l_{-\alpha}) = p(x, \theta)(\partial_i l)(\partial_j l)$, the inner product has the following dualistic expression for any α:

$$\langle A, B \rangle = \int Al_\alpha(x, \theta) Bl_{-\alpha}(x, \theta) d\mathbb{P}.$$

The vector spaces $T_\theta^{(\alpha)}$ and $T_\theta^{(-\alpha)}$ are dually coupled: the inner product of two vectors is given by the integration of the product of their α and $-\alpha$ representations. After computation, we have:

$$\Gamma_{ijk}^{(\alpha)}(\theta) = \langle \nabla^\alpha_{\partial_i} \partial_j, \partial_k \rangle = \int (\partial_i \partial_j l_\alpha(x, \theta))(\partial_k l_{-\alpha}(x, \theta)) d\mathbb{P},$$

[12] Recall that normal coordinates at a point u mean that the connection coefficient and all their derivatives vanish at this point.

[13] A function f is homogeneous of degree d if $\forall K, \forall p, f(Kp) = K^d f(p)$.

[14] We can check that everything cancels out nicely.

using the $\pm\alpha$-representations. Let A and B be two vector fields. We have:

$$ABl_\alpha(x,\theta) = p(x,\theta)^{\frac{1-\alpha}{2}}(ABl + \frac{1-\alpha}{2}AlBl).$$

Thus, for three vectors fields A, B, C, the α-covariant derivative can be written as[15]:

$$\langle \nabla^\alpha_A B, C\rangle = \mathbb{E}_\alpha[(ABl_\alpha)(Cl_\alpha)] = \int (ABl_\alpha(x,\theta))(Cl_{-\alpha}(x,\theta))d\mathbb{P}.$$

Hence, the α-covariant derivative $\nabla^\alpha_A B$ is given by projecting ABl_α to $T^{(\alpha)}_\theta$ in the α-representation. Therefore, $T^{(\alpha)}_\theta$ provides a natural frame for representing the α-covariant derivative and studying the properties of the α-connection.

Remark that the metric and the α-connections are invariant under changes of the parametrization θ of the family S. They are also invariant under one-to-one transformations of the random variable x to $y = f(x)$, which can be seen as a coordinate transformation of the sample space X. The new family is $S' = \{q(y,\theta)\}$ with:

$$q(y,\theta) = p(f^{-1}(y),\theta)J^{-1}(y),$$

where $J = \det|\partial f/\partial x|$. The invariancy of geometrical structures comes from the fact that:

$$\langle \partial_i, \partial_j\rangle = \mathbb{E}[(\partial_i \log p)(\partial_j \log p)] = \mathbb{E}[(\partial_i \log q)(\partial_j \log q)].$$

When using the function F_α for the α-representation, the α-covariant derivative is naturally defined. But what about the uniqueness of these geometrical structures? Is it possible to introduce another geometrical structure by using another function F instead of F_α? Assume that we introduce another inner product $\langle A, B\rangle'$ using the representation $F(p(x,\theta))$ as:

$$\langle A, B\rangle' = \mathbb{E}'[(AF)(BF)],$$

and another covariant derivative ∇':

$$\langle \nabla'_A B, C\rangle' = \mathbb{E}'[(ABF)(CF)],$$

where $\mathbb{E}'$ is some kind of expectation operator:

$$\mathbb{E}'[f(x)] = \int G(p(x,\theta))f(x)d\mathbb{P},$$

for some function G. It can be proved that, if we want the above definitions to be invariant under coordinate transformations of the sample space X, then the derivative $F'(p)$ must be an homogeneous function in p. Hence, we are naturally led to the class of functions F_α, and no other definition can produce invariant structures. In fact, we can use a slightly more general definition of F_α to define the same geometrical structures:

$$F_\alpha(p) = \begin{cases} \frac{2}{1-\alpha}(p^{\frac{1-\alpha}{2}} - C_\alpha(x)) & , \alpha \neq 1, \\ \log p & , \alpha = 1, \end{cases}$$

where $C_\alpha(x)$ is a function in x (but not in θ). When $\lim_{\alpha\to 1} C_\alpha(x) = 1$ holds, then $F_\alpha(p)$ is continuous with respect to α. However, we use $C_\alpha(x) = 0$ in the definition only for brevity's sake.

5.2.2 Dual affine connections

Let $c : \theta(t)$ be a smooth curve in a statistical manifold S and $B(t)$ be a vector field defined on the curve. The intrinsic change of $B(t)$ along the curve is measured by $\nabla_{\dot\theta}B(t)$. When $\nabla_{\dot\theta}B(t) = 0$, there is no intrinsic change in $B(t)$ along the curve. In this case, the vector $B(t')$ is said to be the **parallel displacement** of vector $B(t)$ from $\theta(t)$ to $\theta(t')$ along the curve. The parallel displacement defines a mapping:

$$\Pi_c : T_\theta \to T_{\theta'}, \Pi_c B(t) \mapsto B(t').$$

[15]Remember that the α-connection coefficients are defined by $\Gamma^{(\alpha)}_{ijk}(\theta) = \mathbb{E}[(\partial_i\partial_j l(x,\theta) + \frac{1-\alpha}{2}\partial_i l(x,\theta)\partial_j l(x,\theta))\partial_k l(x,\theta)]$.

The parallel displacement does not necessarily preserve the metric structure of the tangent space (i.e., the inner product of two vectors is not necessarily equal to the vector product of their parallel displacements). An affine connection is said to be **metric** when the metric (i.e., inner product) is preserved by the parallel displacement.

For a non-metric covariant derivative ∇, there might exist another covariant derivative ∇^* such that the pair (∇, ∇^*) satisfies:

$$\langle A, B\rangle_\theta = \langle \Pi_c A, \Pi_c^* B\rangle_{\theta'}. \tag{9}$$

Such a pair of covariant derivatives are said to be **mutually dual**. When ∇ is metric, it is **self-dual**.

The **dual connections** are formally defined when, for any vector fields A, B, C:

$$A\langle B, C\rangle = \langle \nabla_A B, C\rangle + \langle B, \nabla_A C\rangle.$$

Substituting $A = \partial_i, B = \partial_j, C = \partial_k$, we have:

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma^*_{ikj}.$$

This shows that every affine connection has a unique dual determined by:

$$\Gamma^*_{ijk} = \partial_i g_{jk} - \Gamma_{ikj},$$

and the dual of the dual is the primal one, $\nabla^{**} = \nabla$.

Now, we prove 9. Let $A(t)$ and $B(t)$ be vector fields obtained by the parallel displacements of two vectors $A, B \in T_\theta$ along the curve c with respect to the dual connections respectively. We have that $\nabla_{\dot\theta} A(t) = 0$ and $\nabla^*_{\dot\theta} B(t) = 0$. Hence:

$$\frac{d}{dt}\langle A(t), B(t)\rangle = \dot\theta\langle A(t), B(t)\rangle = \langle \nabla_{\dot\theta} A(t), B\rangle + \langle A, \nabla^*_{\dot\theta} B(t)\rangle = 0.$$

Thus, the inner product is preserved by parallel displacement with respect to the dual connections respectively.

Theorem 5.1 (Dual α-connections)**.** *The α- and $-\alpha$-connections are mutually dual. In particular, the 0-connection is self-dual and hence is metric.*

Proof.

$$\begin{aligned} A\langle B, C\rangle &= A\int (Bl_\alpha)(Cl_{-\alpha})d\mathbb{P}, \\ &= \int (ABl_\alpha)(Cl_{-\alpha})d\mathbb{P} + \int (Bl_\alpha)(ACl_{-\alpha})d\mathbb{P}, \text{ (chain rule for vector fields)} \\ &= \langle \nabla^\alpha_A B, C\rangle + \langle B, \nabla^{-\alpha}_A C\rangle. \end{aligned}$$

□

We have the following result. When S is flat with respect to ∇, it is also flat with respect to its dual ∇^*. In particular, an α-flat manifold is also $-\alpha$-flat. To prove this fact, recall the following property. Let c be a loop passing through θ and $A \in T_\theta$. A manifold S is flat when any vectors do not change by the parallel displacement along any loops i.e., for any A and any c, $\Pi_c A = A$. In this case, the Riemann-Christoffel curvature vanishes identically. Now, for the parallel displacement along a loop c, we have for any A, B: $\langle \Pi_c A, \Pi_c^* B\rangle = \langle A, B\rangle$. Assume S is flat with respect to ∇. Hence, $\Pi_c A = A$ so that, for any A, B, $\langle A, \Pi_c^* B\rangle = \langle A, B\rangle$. This implies $\Pi_c^* B = B$, proving that S is flat with respect to ∇^*.

If c^{-1} is the inverse loop encircling c in the reverse order, we have $\Pi_{c^{-1}} = (\Pi_c)^{-1}$. Hence:

$$\begin{aligned} \langle \Pi_c C, D\rangle &= \langle \Pi_{c^{-1}}\Pi_c C, \Pi^*_{c^{-1}} D\rangle \text{ (dual connections definition)}, \\ &= \langle (\Pi_c)^{-1}\Pi_c C, \Pi^*_{c^{-1}} D\rangle, \\ &= \langle C, \Pi^*_{c^{-1}} D\rangle. \end{aligned}$$

Since c^{-1} is the reverse loop of c, we have:

$$R(A,B,C,D) = R^*(B,A,D,C) = -R^*(A,B,C,D),$$

and

$$R_{ijkm} = -R^*_{ijmk}.$$

In the statistical manifold:

$$R^{(\alpha)}_{ijkm} = -R^{(-\alpha)}_{ijmk}.$$

5.2.3 α-family of distributions

We define the **α-family of distributions** by extending the exponential family and the mixture family. A family $S = \{p(x,\theta)\}$ of distributions is said to be an α-family when their α-representations can be written as $l_\alpha(x,\theta) = \theta^i c_i(x) + k(\theta)c_{n+1}(x)$ by choosing an adequate parametrization $\theta = (\theta^1,\dots,\theta^n)$, where $c_i(x)$ are fixed random variables and $k(\theta)$ is determined from the normalization condition. We can write $\theta^{n+1} = k(\theta)$ and define $\tilde{\theta} = (\theta^1,\dots,\theta^{n+1})$. We call $\tilde{\theta}$ the **natural** or **canonical homogeneous coordinate system** of the α-family. The first n coordinates θ (or any subset of n coordinates) can be adopted as a coordinate system, but the $n+1$ components are not independent.

If $c_{n+1}(x) = 1$ and $k(\theta) = -\psi(\theta)$, the 1-family is $l(x,\theta) = \theta^i c_i(x) - \psi(\theta)$ so that $p(x,\theta) = \exp(\theta^i c_i(x) - \psi(\theta))$, which is the standard form of the exponential family. The function $\psi(\theta)$ is related to the cumulant generating function. Indeed, the characteristic function[16] of the $c_i(x)$'s is:

$$\mathbb{E}_\theta[\exp(is^j c_j(x))] = \int \exp((is^j + \theta^j)c_j(x) - \psi(\theta))d\mathbb{P} = \exp(\psi(is+\theta) - \psi(\theta)).$$

This shows that $\exp(\psi(s+\theta) - \psi(\theta))$ is the moment generating function[17]:

$$\mathbb{E}[c_{i_1}(x)\dots c_{i_k}(x)] = \frac{\partial^p}{\partial s^{i_1}\dots\partial s^{i_k}} \exp(\psi(s+\theta) - \psi(\theta))|_{s=0}.$$

Hence, $\psi(s+\theta) - \psi(\theta)$ or $\psi(s+\theta)$ itself is the cumulant generating function of $c_i(x)$ with respect to the distribution $p(x,\theta)$.

The -1-family can be written as $P(x,\tilde{\theta}) = \tilde{\theta}^i c_i(x)$. When all c_i's satisfy $c_i(x) > 0$ and $\int c_i(x)d\mathbb{P} = 1$, we obtain the mixture family.

The discrete (or multinomial) distributions $p(x,\xi) = \xi^i\delta_i(x), \sum_{i=1}^{n+1}\xi^i = 1$ is an α-family for any α. In particular, their family is an exponential family and also a mixture family.

In order to study the properties of S, we extend it to a manifold of finite measures $\tilde{S} = \{m(x,\theta,c)\}$ with $m(x,\theta,c) = cp(x,\theta), c > 0, p(x,\theta) \in S$. We consider S as a submanifold of $\tilde{S}$, because $\tilde{S}$ has simpler geometrical structures. $\tilde{S}$ is $(n+1)$-dimensional and the pair (θ,c) is an example of coordinate system of $\tilde{S}$. Let $\tilde{\theta}$ be a coordinate system in $\tilde{S}$ such that a member of $\tilde{S}$ is parameterized as $m(x,\tilde{\theta})$. Let $K(\tilde{\theta})$ be the total measure of the distribution $m(x,\tilde{\theta})$: $K(\tilde{\theta}) = \int m(x,\tilde{\theta})d\mathbb{P}$. The original S is a submanifold in $\tilde{S}$ defined by $K(\tilde{\theta}) = 1$.

The geometrical structure can be introduced in $\tilde{S}$ in the same manner as in S, and the geometrical structures of the original S are compatible with those induced from $\tilde{S}$ as a submanifold.

When S is an α-family, the α-representation of the measure in the extended $\tilde{S}$ are $\tilde{l}_\alpha(x,\tilde{\theta}) = \tilde{\theta}^i c_i(x)$ (when $\alpha = 1$, we assume $c_{n+1}(x) = 1$). Since $\tilde{\partial}_i\tilde{\partial}_j\tilde{l}_\alpha(x,\tilde{\theta}) = 0$, the α-covariant derivative of $\tilde{S}$ vanishes. This implies that $\tilde{S}$ is α-flat and that the natural coordinate system $\tilde{\theta}$ is α-affine. The geodesic connecting two points $\tilde{\theta}_1$ and $\tilde{\theta}_2$ in S is: $\tilde{\theta}(t) = (1-t)\tilde{\theta}_1 + t\tilde{\theta}_2$. Remark that even when the two points $\tilde{\theta}_1, \tilde{\theta}_2$ belong to S, the geodesic $\tilde{\theta}(t)$ does not necessarily belong to S, because S is in general curved in $\tilde{S}$.

A submanifold S' in S is said to be **autoparallel** when it has vanishing imbedding curvature (i.e., vanishing Euler-Schouten curvature). Since our manifolds are torsion-free, an autoparallel submanifold is **totally geodesic** i.e., it consists of all the geodesics whose tangent vectors belong to $T_\theta(S')$. For a

[16] Recall that the characteristic function of x is $\phi_x(s) = \mathbb{E}[\exp(is^j x_j)]$.

[17] Recall that the moment generating function of x is $M_x(s) = \mathbb{E}[\exp(s^j x_j)]$. In particular, $\phi_x(s) = M_x(is)$. The cumulant generating function is defined as $K(s) = \log M_x(s)$.

submanifold S' of an α-family S, the extension $\tilde{S}'$ of S' is a submanifold of $\tilde{S}$. We have the following result. A submanifold S' of an α-family S is autoparallel in S if and only if the extended submanifold $\tilde{S}'$ of S' is autoparallel in the extended manifold $\tilde{S}$. Using this result[18], it is possible to obtain the geodesic c in the α-family S from the α-geodesic $\tilde{c}$ in $\tilde{S}$. This is done by "projecting" $\tilde{c}$ to S. More precisely, it is done by adding a normalization constant that depends on t such that $K(\tilde{c}(t)) = 1$ for all t.

The mixture family ($\alpha = -1$) is special since S itself is a -1-flat submanifold in $\tilde{S}$. This is because the constraint $K(\tilde{\theta}) = \sum_{i=1}^{n+1} \tilde{\theta}^i = 1$ which determines S is linear in $\tilde{\theta}$, hence the mixture family S is autoparallel. The exponential family ($\alpha = 1$) is also special. The extended manifold $\tilde{S}$ has the form $\tilde{l}(x, \tilde{\theta}) = \tilde{\theta}^i c_i(x) + \tilde{\theta}^{n+1}$ so that the constraint determining S is $\tilde{\theta}^{n+1} = -\psi(\theta)$ and is not linear in θ, so S is not an autoparallel submanifold in $\tilde{S}$. However, S itself is a 1-flat manifold (having null Riemann-Christoffel curvature since the 1-connection vanishes) but the imbedding curvature of S in $\tilde{S}$ does not vanish.

To summarize, the extended manifold $\tilde{S}$ of any α-family S is α-flat while S itself is in general not so. Both the mixture and the exponential family are 1- and -1-flat by themselves.

5.2.4 Duality in α-flat manifolds

When a manifold is ∇-flat, it is also ∇^*-flat. There are two special coordinate systems in such a dually flat manifold: ∇-affine coordinate system θ and ∇^*-affine coordinate system η.

Let $\theta = (\theta^i)$ and $\eta = (\eta_i)$ be two coordinate systems in an n-dimensional Riemannian manifold S (the lower index is employed to denote the components of η with the intention of constructing a dualistic theory). The natural basis of the tangent space T_P is $\{\partial_i = \partial/\partial\theta^i\}$ for the coordinate system θ and $\{\partial^i = \partial/\partial\eta_i\}$ for the coordinate system η. When we have $\langle \partial_i, \partial^j \rangle = \delta_i^j$, the two bases are said to be biorthogonal.

Two coordinate systems θ and η are said to be mutually dual when their natural bases are biorthogonal. Dual coordinate systems do not necessarily exist in a Riemannian manifold (but they always exist in an α-flat manifold). Here, we assume that a pair of dual systems exist in S. The two natural bases are related by $\partial_i = (\partial\eta_k/\partial\theta^i)\partial^k, \partial^j = (\partial\theta^k/\partial\eta_j)\partial_k$. We have:

$$g_{ij} = \langle \partial_i, \partial_j \rangle = (\partial\eta_k/\partial\theta^i)\langle \partial^k, \partial_j \rangle = \partial\eta_j/\partial\theta^i.$$

Hence, $\partial\theta^j/\partial\eta_k = g^{jk}$ and $\langle \partial^i, \partial^j \rangle = g^{ij}$.

Theorem 5.2 (Potential functions)**.** *When a Riemannian manifold S has a pair of dual coordinate system (θ, η), there exist **potential functions** $\psi(\theta)$ and $\phi(\eta)$ such that*

$$g_{ij}(\theta) = \partial_i\partial_j\psi(\theta), g^{ij}(\eta) = \partial^i\partial^j\phi(\eta).$$

Conversely, when either potential function ψ or ϕ exists from which the metric is derived, there exists a pair of dual coordinate systems.

*The dual coordinate systems are related by the **Legendre transformations***

$$\theta^i = \partial^i\phi(\eta), \eta_i = \partial_i\psi(\theta),$$

where the two potential functions satisfy the identity

$$\psi(\theta) + \phi(\eta) - \theta^i\eta_i = 0.$$

When a Riemannian manifold S is flat with respect to a pair of torsion-free dual affine connections ∇ and ∇^*, there exists a pair (θ, η) of dual coordinate systems such that θ is ∇-affine and η is ∇^*-affine.

6 Methods of Information Geometry

This section relies on Amari & Nagaoka [5].

[18] But I do not understand how ...

6.1 The geometric structure of statistical models

Let $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ be a n-dimensional statistical model. The distributions $p_\theta(x)$ are defined over a sample space $\mathcal{X}$ with $x \in \mathcal{X}$.
In order to be a statistical model, a family of distributions $\mathcal{S}$ must verify some regularity conditions. A simple way to state these conditions is to see $\mathcal{P}(\mathcal{X})$ as a Riemannian manifold with the Fisher metric, then ask $\mathcal{S}$ to be a submanifold of $\mathcal{P}(\mathcal{X})$.
If $\mathcal{X}$ is finite, $\mathcal{P}(\mathcal{X})$ can easily be described as a manifold. However, it is no longer the case when $\mathcal{X}$ is infinite, but the intuition is the same.

6.1.1 The Fisher metric

Let $l_\theta(x) = l(x;\theta) = \log p(x;\theta) = \log p_\theta(x)$.

Definition 6.1 (Fisher information matrix)**.** $G(\theta) = [g_{ij}(\theta)]$ such that:

$$g_{ij}(\theta) = \mathbb{E}_\theta[\partial_i l_\theta \partial_j l_\theta] = \int \partial_i l(x;\theta)\partial_j l(x;\theta)p(x;\theta)dx.$$

Proposition 6.2. *$G(\theta)$ is symmetric positive semidefinite.*
$G(\theta)$ is positive definite if and only if the functions $\{\partial_i p_\theta : \mathcal{X} \to \mathbb{R}\}$ are linearly independent.
Moreover, we have:

$$g_{ij}(\theta) = -\mathbb{E}_\theta[\partial_i \partial_j l_\theta].$$
$$g_{ij}(\theta) = 4\int \partial_i \sqrt{p(x;\theta)}\partial_j \sqrt{p(x;\theta)}dx.$$

Sufficient statistics Let X be a random variable over $\mathcal{X}$ with distribution $p(x;\theta)$.
Let Y a random variable over $\mathcal{Y}$ with distribution $q(y;\theta)$.
Let $\kappa(y|x)$ be the conditional probability distribution of y given x.
The conditional probability distribution of x given y is:

$$p(x|y;\theta) = \frac{p(x;\theta)\kappa(y|x)}{q(y;\theta)}.$$

Now, we focus on the case where Y is a deterministic function of X, i.e., $Y = F(X)$ for some deterministic function F. Then, we have $\kappa(y|x) = \delta_{F(x)}(y)$. Define:

$$r(x;\theta) = \frac{p(x;\theta)}{q(F(x);\theta))}.$$

Definition 6.3 (Sufficient statistic)**.** F is a **sufficient statistic** if $r(x;\theta) = r(x)$ does not depend on θ.
Equivalently, F is a sufficient statistic if the conditional probability distribution $p(x|y;\theta) = r(x;\theta)\delta_{F(x)}(y) = r(x)\delta_{F(x)}(y)$ does not depend on θ.

One-to-one mappings ("reparametrization") are examples of sufficient statistics.

Proposition 6.4 (Fisher-Neyman factorization theorem)**.** *F is a sufficient statistic if and only if there exist some functions $s : \mathcal{Y} \times \Theta \to \mathbb{R}$ and $t : \mathcal{X} \to \mathbb{R}$ such that:*

$$p(x;\theta) = s(F(x);\theta)t(x).$$

In particular, if F is a sufficient statistic, we have:

$$p(x;\theta) = q(F(x);\theta)r(x).$$

Thus, the portion of the distribution $p(x;\theta)$ which depends on θ is entirely contained within the distribution $q(y;\theta)$ of $Y = F(X)$. Hence, in order to estimate the unknown parameter θ, it is *sufficient* to know the value of Y. This is why F is called a *sufficient statistic*. Indeed, given the value Y, we can "simulate" X by using the "random number generator" $p(x|y)$ which does not depend on θ.

Let $G_F(\theta)$ be the Fisher information matrix of $\mathcal{S}_F = \{q(y;\theta)\}$. Let $\Delta G(\theta) = G(\theta) - G_F(\theta)$ (**chain rule** of the Fisher metric).

Theorem 6.5. ***Monotonicity*** *of the Fisher metric:*

$$G_F(\theta) \preceq G(\theta),$$

or equivalently $\Delta G(\theta)$ *is positive semidefinite.*
$G_F(\theta) = G(\theta)$ *if and only if* F *is a sufficient statistic for* $\mathcal{S}$.
The information loss caused by summarizing the data x *into* $y = F(x)$ *is:*

$$\begin{aligned}\Delta g_{ij}(\theta) &= \mathbb{E}_\theta[\partial_i \log r(X;\theta)\partial_j \log r(X;\theta)],\\ &= \mathbb{E}_\theta\left[\mathrm{Cov}_\theta[\partial_i l(X;\theta), \partial_j l(X;\theta)|Y]\right],\\ &= \int_{\mathcal{Y}}\left(\int_{\mathcal{X}} \left(\partial_i l_\theta - \mathbb{E}_\theta[\partial_i l_\theta|y]\right)\left(\partial_j l_\theta - \mathbb{E}_\theta[\partial_j l_\theta|y]\right) p(x|y;\theta)dx\right) q(y;\theta)dy.\end{aligned}$$

Corollary 6.6. ***Additivity*** *of the Fisher metric. If* $p_{12}(x_1, x_2;\theta) = p_1(x_1;\theta)p_2(x_2;\theta)$ *then:*

$$G_{12}(\theta) = G_1(\theta) + G_2(\theta).$$

Convexity *of the Fisher metric. If* $p_\lambda(x;\theta) = \lambda p_1(x;\theta) + (1-\lambda)p_2(x;\theta)$ *with* $0 \le \lambda \le 1$*, then:*

$$G_\lambda(\theta) \le \lambda G_1(\theta) + (1-\lambda)G_2(\theta).$$

Efficient estimator Let X be a random variable with distribution $\mathbb{P}_\theta$ where θ is unknown.

Definition 6.7 (Unbiased estimator)**.** The mapping $\hat{\theta} : \mathcal{X} \to \mathbb{R}^n$ is an **unbiased estimator** if:

$$\mathbb{E}_\theta\left[\hat{\theta}(X)\right] = \theta.$$

The mean squared error of an unbiased estimator $\hat{\theta}$ is the covariance matrix $V_\theta\left[\hat{\theta}\right] = [v_\theta^{ij}]$ such that:

$$v_\theta^{ij} = \mathbb{E}_\theta\left[\left(\hat{\theta}^i(X) - \theta^i\right)\left(\hat{\theta}^j(X) - \theta^j\right)\right].$$

Theorem 6.8 (Cramér-Rao inequality)**.** *If* $\hat{\theta}$ *is an unbiased estimator, then:*

$$V_\theta\left[\hat{\theta}\right] \succeq G(\theta)^{-1}.$$

Definition 6.9 (Efficient estimator)**.** The unbiased estimator $\hat{\theta}$ is an **efficient estimator** if $V_\theta\left[\hat{\theta}\right] = G(\theta)^{-1}$ for all θ.

Efficient estimators may not exist. Furthermore, a biased estimator may have a smaller mean square error than the efficient estimator.

Definition 6.10 (Asymptotically efficient estimator)**.** Let $x_1, \ldots, x_N$ be a sequence of iid observations from p_θ.
A sequence of estimators $\left\{\hat{\theta}_N(x_1, \ldots, x_N)\right\}_{N=1}^{\infty}$ is an **asymptotically efficient estimator** or **first-order efficient estimator** if it achieves equality in the Cramér-Rao inequality as $N \to \infty$.
Such asymptotically efficient estimators always exist.

In order to accurately estimate the parameter θ, the "character" of the data (i.e., p_θ) should change dramatically as θ varies. The Fisher metric can be seen as a geometric expression of the size of this change. The larger $G(\theta)$ (i.e., the smaller $G(\theta)^{-1}$), the less an efficient estimator will fluctuate around the true value of θ.

6.1.2 The α-connection

Definition 6.11. Let $\alpha \in \mathbb{R}$. The α-connection is defined as:

$$\left(\Gamma^{(\alpha)}_{ij,k}\right)_\theta = \mathbb{E}_\theta\left[\left(\partial_i\partial_j l_\theta + \frac{1-\alpha}{2}\partial_i l_\theta \partial_j l_\theta\right)(\partial_k l_\theta)\right]. \tag{10}$$

Define also the 3-covariant symmetric tensor:

$$(T_{ijk})_\theta = \mathbb{E}_\theta[\partial_i l_\theta \partial_j l_\theta \partial_k l_\theta].$$

Definition 6.12 (Exponential family)**.** Let $C, F_1, \dots, F_n$ be functions $\mathcal{X} \to \mathbb{R}$ and $\psi : \Theta \to \mathbb{R}$. Assume that the functions $\{1, F_1, \dots, F_n\}$ are linearly independent. Then define:

$$p(x;\theta) = \exp\left[C(x) + \sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right].$$

The $[\theta^i]$ are called the **natural** or **canonical parameters**. We have:

$$\psi(\theta) = \log \int \exp\left[C(x) + \sum_{i=1}^n \theta^i F_i(x)\right] dx.$$

Proposition 6.13 (Properties of exponential family)**.**

$$\begin{aligned}\partial_i l(x;\theta) &= F_i(x) - \partial_i\psi(\theta),\\ \partial_i\partial_j l(x;\theta) &= -\partial_i\partial_j\psi(\theta).\end{aligned}$$

Then, using the fact that $\mathbb{E}_\theta[\partial_k l_\theta] = 0$, we have:

$$\Gamma^{(1)}_{ij,k} = -\partial_i\partial_j\psi(\theta)\mathbb{E}_\theta[\partial_k l_\theta] = 0.$$

Thus, $[\theta^i]$ ***is a 1-affine coordinate system*** *and $\mathcal{S}$* ***is 1-flat****.*

$\nabla^{(1)} = \nabla^{(e)}$ can be called the **exponential connection** or **e-connection**.

Definition 6.14 (Mixture family)**.** Let $C, F_1, \dots, F_n$ be functions $\mathcal{X} \to \mathbb{R}$. Then define:

$$p(x;\theta) = C(x) + \sum_{i=1}^n \theta^i F_i(x).$$

The $[\theta^i]$ are called the **mixture parameters**.

Proposition 6.15 (Properties of mixture family)**.** *A mixture family is an affine subspace of $\mathcal{P}(\mathcal{X})$. When $\mathcal{X}$ is finite, $\mathcal{P}(\mathcal{X})$ is a mixture family.*

$$\begin{aligned}\partial_i l(x;\theta) &= \frac{F_i(x)}{p(x;\theta)}\\ \partial_i\partial_j l(x;\theta) &= -\frac{F_i(x)F_j(x)}{p(x;\theta)^2}.\end{aligned}$$

Then:

$$\Gamma^{(-1)}_{ij,k} = \mathbb{E}_\theta\left[(\partial_i\partial_j l_\theta + \partial_i l_\theta \partial_j l_\theta)(\partial_k l_\theta)\right] = 0.$$

Thus, $[\theta^i]$ ***is a (-1)-affine coordinate system*** *and $\mathcal{S}$* ***is (-1)-flat****.*

$\nabla^{(-1)} = \nabla^{(m)}$ can be called the **mixture connection** or **m-connection**.
Remark: a statistical model can be α-flat for all α (thus Euclidean for the Fisher metric) while being neither an exponential nor a mixture family.

6.1.3 Chentsov's theorem

If $F : \mathcal{X} \to \mathcal{Y}$ is a sufficient statistic, then $\partial_i l_\theta(x) = \partial_i \log p(x;\theta) = \partial_i \log q(F(x);\theta) = \partial_i l^F_\theta(F(x))$. This is because $p(x;\theta) = q(F(x);\theta)r(x)$. Hence, g_{ij} and $\Gamma^{(\alpha)}_{ij,k}$ are the same on both $\mathcal{S}$ and $\mathcal{S}_F$. This is the **invariance of the Fisher metric and the α-connection with respect to F.**

Theorem 6.16 (Chentsov)**.** *Let $\mathcal{X}$ be finite and let g and ∇ be arbitrary Riemannian metric and affine connection on a statistical model $\mathcal{S}$ defined on $\mathcal{X}$.*
Assume that g and ∇ are invariant with respect to sufficient statistics.
Then there exist $c > 0$ and $\alpha \in \mathbb{R}$ such that g equals c times the Fisher metric and ∇ is the α-connection.

Under certain conditions, Chentsov's theorem can be extended to infinite set $\mathcal{X}$ but this is not easy.

6.1.4 The geometry of $\mathcal{P}(\mathcal{X})$

Since any statistical model is a submanifold of $\mathcal{P} = \mathcal{P}(\mathcal{X})$, we will study the properties of $\mathcal{P}$. We assume that $\mathcal{X}$ is finite.
$\mathcal{P}$ is a subset of $\mathbb{R}^{\mathcal{X}} = \{A : \mathcal{X} \to \mathbb{R}\}$, the set of $\mathbb{R}$-valued functions on $\mathcal{X}$. Remark: since $\mathcal{X}$ is finite, $\mathbb{R}^{\mathcal{X}}$ can be identified with $\mathbb{R}^{|\mathcal{X}|}$.
More precisely, $\mathcal{P}$ is an open set of the affine subspace $\mathcal{A}_1 = \left\{A : \sum_{x\in\mathcal{X}} A(x) = 1\right\}$ of $\mathbb{R}^{\mathcal{X}}$. Hence, the tangent space $T_p(\mathcal{P})$ can be identified with the linear subspace $\mathcal{A}_0 = \left\{A : \sum_{x\in\mathcal{X}} A(x) = 0\right\}$.

Mixture representation or m-representation Let $X \in T_p(\mathcal{P})$. If X is seen as an element of $\mathcal{A}_0$, it is denoted by $X^{(m)}$. Thus:

$$T^{(m)}_p(\mathcal{P}) = \mathcal{A}_0.$$

Let $[\theta^i]$ be a coordinate system. Then, the m-representation of a basis vector ∂_i is $(\partial_i)^{(m)}_\theta = \partial_i p_\theta$.
Since $\mathcal{P}$ is a mixture family, it is m-flat, i.e., its mixture parameters form an m-affine coordinate system. Thus, **the m-connection on $\mathcal{P}$ is the natural connection induced from the affine structure of $\mathcal{A}_1$.** The natural embedding of $\mathcal{P}$ into $\mathbb{R}^{\mathcal{X}}$ makes the meaning of the m-connection clear.

Exponential representation or e-representation Consider the embedding $p \mapsto \log p$. We can identify $\mathcal{P}$ with the subset $\{\log p : p \in \mathcal{P}\} \subset \mathbb{R}^{\mathcal{X}}$.
In a coordinate system $[\theta^i]$, we have $(\partial_i)^{(e)} = \partial_i \log p_\theta$.
If $X \in T_p(\mathcal{P})$, we denote by $X^{(e)}$ its e-representation. We have:

$$X^{(e)}(x) = \frac{X^{(m)}(x)}{p(x)},$$

and:

$$T^{(e)}_p(\mathcal{P}) = \left\{ A \in \mathbb{R}^{\mathcal{X}} : \mathbb{E}_p[A] = \sum_{x\in\mathcal{X}} p(x)A(x) = 0 \right\}.$$

In the e-representation, the Fisher metric can be expressed:

$$\langle X, Y\rangle_p = \mathbb{E}_p\left[X^{(e)}Y^{(e)}\right] = X^i_p Y^j_p \sum_{x\in\mathcal{X}} \left(\partial_i \log p(x)\right)_p \left(\partial_j \log p(x)\right)_p p(x). \tag{11}$$

Unlike $T^{(m)}_p(\mathcal{P})$, the space $T^{(e)}_p(\mathcal{P})$ depends on p.

Theorem 6.17.

$$\Pi^{(e)}_{p,q}(X) = X' \iff X'^{(e)} = X^{(e)} - \mathbb{E}_q\left[X^{(e)}\right]. \tag{12}$$

Proof. Let X be an arbitrary vector field. In the e-representation, we have $X^{(e)} : p \mapsto X^{(e)}_p = X^i_p \left(\partial_i \log p\right) \in \mathbb{R}^{\mathcal{X}}$. According to Equation 10:

$$\left\langle \nabla^{(e)}_{\partial_i} X, \partial_k \right\rangle_p = \mathbb{E}_p\left[\left(\partial_i X^{(e)}\right)_p (\partial_k)^{(e)}_p\right]$$

Assume that there is a function F on $\mathcal{X}$ (which can be seen as a random variable) such that $X_p^{(e)} = F - \mathbb{E}_p[F]$. We have $\left(\partial_i X^{(e)}\right)_p = -\partial_i \mathbb{E}_p[F]$ which does not depend on $x \in \mathcal{X}$. Thus:

$$\left\langle \nabla^{(e)}_{\partial_i} X, \partial_k \right\rangle_p = -\partial_i \mathbb{E}_p[F] \mathbb{E}_p\left[(\partial_k)_p^{(e)}\right] = 0,$$

since $\mathbb{E}_p\left[(\partial_k)_p^{(e)}\right] = 0$. Thus[19], $\nabla^{(e)} X$ is identically 0. Hence, X is e-parallel.
The proof of Amari & Nagaoka ends here, although I feel that we only showed one implication ($\Leftarrow$). □

Remark:

$$\Pi^{(m)}_{p,q}(X) = X' \iff X'^{(e)} = \frac{p}{q} X^{(e)}.$$

Now, we use the e-representation to prove the Cramér-Rao inequality as well as a result that seems to be linked to the Fisher-Darmois-Pitman-Koopman's theorem.
First, remember that an inner product gives rise to a natural isomorphism between $T_p(\mathcal{P})$ and $T_p^*(\mathcal{P})$ using $\omega_X(Y) = \langle X, Y \rangle_p$. Thus, given a function $f \in \mathcal{F}(\mathcal{P})$ and its differential $(df)_p(X) = X(f)$, we can define its gradient as:

$$\langle (\mathrm{grad} f)_p, X \rangle_p = (df)_p(X) = X(f). \tag{13}$$

Let $\|\cdot\|$ be the norm induced from the Fisher metric.

Proposition 6.18 (Properties of the gradient)**.**

$$(\mathrm{grad} f)_p = (\partial_i f)_p g^{ij}(p)(\partial_j)_p,$$
$$\|(df)_p\|_p^2 = \|(\mathrm{grad} f)_p\|_p^2 = (\partial_i f)_p (\partial_j f)_p g^{ij}(p).$$

Let $A \in \mathbb{R}^{\mathcal{X}}$. We see A as the random variable $A(X)$ where X follows the distribution p.
Define the function $\mathbb{E}[A] : \mathcal{P} \to \mathbb{R}$ such that $p \mapsto \mathbb{E}_p[A] = \sum_{x \in \mathcal{X}} p(x) A(x)$.
Define the variance $V_p[A] = \mathbb{E}_p\left[(A - \mathbb{E}_p[A])^2\right]$.

Theorem 6.19.

$$V_p[A] = \|(d\mathbb{E}[A])_p\|_p^2.$$

If $\mathbb{E}[A]$ is restricted to a submanifold $\mathcal{S} \subset \mathcal{P}$, then:

$$V_p[A] \geq \|(d\mathbb{E}[A]|_S)_p\|_p^2,$$

with equality if and only if:

$$A - \mathbb{E}_p[A] \in T_p^{(e)}(\mathcal{S}).$$

Proof. Let X be a tangent vector in $T_p(\mathcal{P})$. We have:

$$\begin{aligned}
X(\mathbb{E}[A]) &= X^i \partial_i \mathbb{E}_p[A], \\
&= \sum_x X^i \partial_i p(x) A(x) = \sum_x X^{(m)}(x) A(x), \\
&= \sum_x \frac{X^{(m)}(x)}{p(x)} p(x) A(x) = \sum_x X^{(e)}(x) p(x) A(x), \\
&= \mathbb{E}_p[X^{(e)} A] = \mathbb{E}_p[X^{(e)}(A - \mathbb{E}_p[A])],
\end{aligned}$$

where we use $\mathbb{E}_p[X^{(e)}] = 0$ in the last equality. We have $A - \mathbb{E}_p[A] \in T_p^{(e)}(\mathcal{P})$. By definition of the gradient (Equation 13) and of the Fisher metric (Equation 11), we have $A - \mathbb{E}_p[A] = (\mathrm{grad}\mathbb{E}[A])_p$. Hence, using Proposition 6.18, we obtain:

$$\|(d\mathbb{E}[A])_p\|_p^2 = \|A - \mathbb{E}_p[A]\|_p^2 = V_p[A].$$

[19] $\nabla^{(e)} X$ is the vector field endomorphism $Y \mapsto \nabla^{(e)}_Y X$.

Now, consider the restricted function $\mathbb{E}[A]|_{\mathcal{S}}$. Its gradient $(\text{grad}\mathbb{E}[A]|_{\mathcal{S}})_p$ is the orthogonal projection of $(\text{grad}\mathbb{E}[A])_p$ onto $T_p(\mathcal{S})$. Thus:

$$\|(d\mathbb{E}[A]|_S)_p\|_p^2 \leq \|(d\mathbb{E}[A])_p\|_p^2 = V_p[A].$$

There is equality if and only if $(\text{grad}\mathbb{E}[A])_p$ is already in $T_p(\mathcal{S})$, i.e., $A - \mathbb{E}_p[A] \in T_p(\mathcal{S})$. □

Let $\hat{\theta}$ be an unbiased estimator and $[c_i] \in \mathbb{R}^n$. Define $A = c_i\hat{\theta}^i$ and apply the previous theorem:

$$V_\theta[c_i\hat{\theta}^i] \geq (\partial_i c_k \theta^k)_\theta (\partial_j c_k \theta^k)_\theta g^{ij}(\theta),$$

where we used $\mathbb{E}_p[c_i\hat{\theta}^i] = c_i\theta^i$ since $\hat{\theta}$ is unbiased. Simplifying the expression, we obtain the Cramér-Rao inequality (for a finite $\mathcal{X}$):

$$c^T V_\theta[\hat{\theta}] \geq c^T G(\theta)^{-1} c.$$

Moreover, the equality condition becomes $c_i(\hat{\theta}^i - \theta^i(p)) \in T_p^{(e)}(\mathcal{S})$ for any $[c_i] \in \mathbb{R}^n$ (using the fact that the estimator is unbiased). Therefore, if $\hat{\theta}$ is an efficient estimator (i.e., verifies the equality condition), there exists n vector fields $X_1, \ldots, X_n$ on $\mathcal{S}$ such that $(X_i^{(e)})_p = \hat{\theta}^i - \theta^i(p)$ for all i and all p. According to the expression of $\Pi^{(e)}$ (Equation 12), the vector fields $X_1, \ldots, X_n$ are parallel with respect to the e-connection on $\mathcal{P}$.
Thus, **if there exists an efficient estimator for $\mathcal{S}$, then $\mathcal{S}$ is e-autoparallel in $\mathcal{P}$, hence $\mathcal{S}$ is an exponential family**. The e-representation is useful for connecting the Fisher metric with statistical notions like expectation and variance. For purely geometrical aspects, see the next subsubsection.

0-representation Consider the embedding of $\mathcal{P}$ into $\mathbb{R}^{\mathcal{X}}$ defined by $p \mapsto 2\sqrt{p}$.
I won't write anything since this is the well-known sphere point-of-view.

6.2 Dual connections

Definition 6.20.

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z^* Y\rangle,$$
$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{ki,j}.$$

Theorem 6.21.

$$\langle \Pi_\gamma(X), \Pi_\gamma^*(Y)\rangle_q = \langle X, Y\rangle_p,$$
$$\langle R(X,Y)Z, W\rangle = -\langle R^*(X,Y)W, Z\rangle,$$
$$R = 0 \iff R^* = 0.$$

There is no similar properties between the torsion tensors T and T^*.

6.2.1 Contrast functions

Definition 6.22 (Contrast function)**.** Let $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ satisfies for any $p, q \in \mathcal{S}$:

$$D(p||q) \geq 0,$$
$$D(p||q) = 0 \iff p = q.$$

Let $[\theta^i]$ be an arbitrary coordinate system. We can see D as a function of $\Theta \times \Theta \subseteq \mathbb{R}^n \times \mathbb{R}^n$. Define:

$$D((\partial_i)_p||p') = \partial_i D(p||p').$$

This can be generalized to any composition of vector fields $D((X_1 \ldots X_l)_p||(Y_1 \ldots Y_m)_{p'})$.
Define the restriction to the diagonal[20] (i.e., $p = p'$):

$$D[X_1 \ldots X_l||.] : p \mapsto D((X_1 \ldots X_l)_p||p).$$

[20] I find this notation confusing. We must distinguish when the derivation is with respect to the first or to the second variable, even if both variables are equal to the same p.

Define the matrix $[g_{ij}^{(D)}]$:

$$g_{ij}^{(D)}(p) = -D[(\partial_i)_p||(\partial_j)_p] = D[(\partial_i\partial_j)_p||p].$$

When $[g_{ij}^{(D)}]$ is positive definite for all $p \in \mathcal{S}$, we say that D is a **contrast function** on $\mathcal{S}$.
Define an affine connection $\nabla^{(D)}$ by:

$$\Gamma_{ij,k}^{(D)} = -D[\partial_i\partial_j||\partial_k]$$

We can write:

$$\langle X, Y\rangle^{(D)} = -D[X||Y],$$
$$\langle \nabla_X^{(D)} Y, Z\rangle^{(D)} = -D[XY||Z].$$

Proposition 6.23. *We have:*

$$D[(\partial_i)_p||p] = D[p||(\partial_i)_p] = 0.$$

If D is a contrast function, then $[g_{ij}^{(D)}]$ is a Riemannian metric on $\mathcal{S}$ and we have:

$$D(p||q) = \frac{1}{2} g_{ij}^{(D)}(q)\Delta\theta^i\Delta\theta^j + o\left(\|\Delta\theta\|^2\right),$$

where $\Delta\theta^i = \theta^i(p) - \theta^i(q)$.
The connection $\nabla^{(D)}$ is always symmetric.
Define:

$$h_{ijk}^{(D)}(p) = D[(\partial_i\partial_j\partial_k)_p||p] = \partial_i g_{jk}^{(D)}(p) + \left(\Gamma_{jk,i}^{(D)}\right)_p.$$

We have the following third-order approximation:

$$D(p||q) = \frac{1}{2} g_{ij}^{(D)}(q)\Delta\theta^i\Delta\theta^j + \frac{1}{6} h_{ijk}^{(D)}(q)\Delta\theta^i\Delta\theta^j\Delta\theta^k + o\left(\|\Delta\theta\|^3\right).$$

We could also define $g^{(D)}$ and $\nabla^{(D)}$ using the third-order approximation.
The **dual** of a contrast function is $D^*(p||q) = D(q||p)$.

Proposition 6.24. *We have:*

$$g^{(D)} = g^{(D^*)}.$$

Moreover, $\nabla^{(D)}$ and $\nabla^{(D^*)}$ are dual with respect to $g^{(D)}$.

Conversely, any (g, ∇, ∇^*) (i.e., a metric with mutually dual *symmetric* connections) is induced from a contrast function. However, this relation is not one-to-one: there are infinitely many contrast functions that induce the same (g, ∇, ∇^*).

f-divergences

Definition 6.25. Let f be a strictly convex and smooth function defined on $\mathbb{R}^{+*}$. Assume that $f(1) = 0$.
Define a special class of contrast functions called **f-divergences**:

$$D_f(p||q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx.$$

Proposition 6.26 (Monotonicity of f-divergence)**.** *Let $\kappa(y|x)$ be a transition probability distribution. Let $p_\kappa(y) = \int \kappa(y|x)p(x)dx$ and $q_\kappa(y) = \int \kappa(y|x)q(x)dx$. Then:*

$$D_f(p||q) \geq D_f(p_\kappa||q_\kappa).$$

There is equality if and only if κ is induced from a sufficient statistic with respect to $\{p, q\}$, i.e[21]*, $p_\kappa(x|y) = q_\kappa(x|y)$.*

[21] $p_\kappa(x|y) = q_\kappa(x|y) \iff \frac{\kappa(y|x)p(x)}{p_\kappa(y)} = \frac{\kappa(y|x)q(x)}{q_\kappa(y)} \iff \frac{q_\kappa(y)}{p_\kappa(y)} = \frac{q(x)}{p(x)} \iff \frac{q(x)}{p(x)}$ is constant $\iff p = q$. WTF??

Since D_f is invariant with respect to sufficient statistics, so are $g^{(D_f)}$ and $\nabla^{(D_f)}$. Thus, according to Chentsov's theorem, $g^{(D_f)}$ must be proportional to the Fisher metric and $\nabla^{(D_f)}$ must be an α-connection.

Example 6.27. If $f(u) = 4(1 - \sqrt{u})$ then:

$$D^f(p||q) = D^{(0)}(p||q) = 2\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

$\nabla^{(0)}$ is the Riemannian connection associated to the Fisher metric.
$\sqrt{D^{(0)}(p||q)}$ is a distance called the **Hellinger distance**. When restricted to a statistical model (i.e., a parameterized family of distributions), the Hellinger distance becomes the Fisher-Rao distance, i.e., the distance induced by the Fisher metric.

If $f(u) = -\log(u)$ then:

$$D^f(p||q) = D^{(-1)}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

$\nabla^{(-1)} = \nabla^{(m)}$ is the mixture connection.
$D^{(-1)}(p||q)$ is called the **relative entropy** or **Kullback-Leibler divergence**.
The relative entropy satisfies a **chain rule**:

$$D^{(-1)}(p||q) = D^{(-1)}(p_\kappa||q_\kappa) + \int D^{(-1)}(p_\kappa(\cdot|y)||q_\kappa(\cdot|y))p_\kappa(y)dy.$$

6.2.2 Dually flat spaces

Let (g, ∇, ∇^*) be a dualistic structure on $\mathcal{S}$. Remember from Proposition 4.25, if ∇ and ∇^* are both symmetric then ∇-flatness and ∇^*-flatness are equivalent. This is the case for α-connections since they are all symmetric.

Definition 6.28. If ∇ and ∇^* are flat, then $(\mathcal{S}, g, \nabla, \nabla^*)$ is called a **dually flat space**.

In particular, for $\mathcal{X}$ finite, $\left(\mathcal{P}(\mathcal{X}), g, \nabla^{(-1)}, \nabla^{(1)}\right)$ is dually flat. Remark-proposition: A mixture family is m-flat. If a submanifold of a mixture family is e-autoparallel, then it may not be an exponential nor a mixture family. However, it must be (± 1)-flat[22].

Dual coordinate systems Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a dually flat space.
Let $[\theta^i]$ be a ∇-affine coordinate system and let $[\eta_j]$ be a ∇^*-affine coordinate system. Let $\partial_i = \partial/\partial\theta^i$ and $\partial^j = \partial/\partial\eta_j$.
We have that $\langle \partial_i, \partial^j \rangle$ is constant on $\mathcal{S}$. Using well-chosen affine transformations, we can chose the constant such that $\langle \partial_i, \partial^j \rangle = \delta_i^j$. Then we say that the coordinate systems are **mutually dual**.
Define the components of the Fisher metric g with respect to $[\theta^i]$ and $[\eta_j]$:

$$g_{ij} = \langle \partial_i, \partial_j \rangle = \frac{\partial \eta_j}{\partial \theta^i},$$
$$g^{ij} = \langle \partial^i, \partial^j \rangle = \frac{\partial \theta^i}{\partial \eta_j}.$$

[22] If a submanifold of a mixture family is (-1)-autoparallel, it must be a mixture family (or is it only true for exponential family? Doesn't matter because the argument works if you take a (1)-autoparallel submanifold of an exponential family). Here, we have a (-1)-flat submanifold of a mixture family, but the authors claim that it may not be a mixture family. Thus, a (-1)-flat submanifold of a (-1)-flat manifold may not be (-1)-autoparallel??

We check that $g_{ij}g^{jk} = \delta_i^k$.
There exist strictly convex functions ψ and ϕ from $\mathcal{S}$ to $\mathbb{R}$ such that:

$$\begin{aligned}\eta_i &= \partial_i\psi = \frac{\partial\psi}{\partial\theta^i},\\ \theta^i &= \partial^i\phi = \frac{\partial\phi}{\partial\eta_i},\\ g_{ij} &= \partial_i\partial_j\psi,\\ g^{ij} &= \partial^i\partial^j\phi,\end{aligned}$$

and we have:

$$\phi = \theta^i\eta_i - \psi. \tag{14}$$

Proposition 6.29. *For all $p \in \mathcal{S}$:*

$$\begin{aligned}\phi(p) &= \max_{q\in\mathcal{S}}\left\{\theta^i(q)\eta_i(p) - \psi(q)\right\},\\ \psi(p) &= \max_{q\in\mathcal{S}}\left\{\theta^i(p)\eta_i(q) - \psi(q)\right\}.\end{aligned}$$

Proof. We have:

$$\phi(p) = \theta^i(p)\eta_i(p) - \psi(p) = \theta^i(p)(\partial_i)_p\psi - \psi(p).$$

Since ψ is convex, it is "above its tangents", hence:

$$\psi(q) \geq \psi(p) + (\partial_i)_p\psi\left(\theta^i(q) - \theta^i(p)\right).$$

Thus:

$$-\psi(p) \geq -\psi(q) + (\partial_i)_p\psi\left(\theta^i(q) - \theta^i(p)\right).$$

Finally:

$$\phi(p) \geq \theta^i(p)(\partial_i)_p\psi - \psi(q) + (\partial_i)_p\psi\left(\theta^i(q) - \theta^i(p)\right) = \theta^i(q)\eta_i(p) - \psi(q),$$

with equality for $q = p$. □

These correspond to **Legendre transformations** with ϕ and ψ the associated **potentials**. Remark that ϕ can be seen as a function of η directly, and ψ can be seen as a function of θ. Moreover, we can define:

$$\begin{aligned}\Gamma^*_{ij,k} &= \langle\nabla^*_{\partial_i}\partial_j, \partial_k\rangle = \partial_i\partial_j\partial_k\psi,\\ \Gamma^{ij,k} &= \langle\nabla_{\partial^i}\partial^j, \partial^k\rangle = \partial^i\partial^j\partial^k\phi.\end{aligned}$$

6.2.3 Canonical divergence

Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a dually flat space and $\{[\theta^i], [\eta_i]\}$ mutually dual affine coordinate systems with potentials $\{\psi, \phi\}$. Define:

$$D(p||q) = \psi(p) + \phi(q) - \theta^i(p)\eta_i(q).$$

This equation uniquely determines a contrast function called the **canonical divergence of** $(\mathcal{S}, g, \nabla, \nabla^*)$ or the (g, ∇)**-divergence** on $\mathcal{S}$. In particular, it is independent of the choice of the mutually dual affine coordinate system.
Remark: if ∇ is the Riemannian connection (and if the manifold is ∇-flat), then $\sqrt{2D(p||q)}$ is the Euclidean distance.

Theorem 6.30 (Triangular relation)**.** *Let D be a contrast function on a dually flat space $(\mathcal{S}, g, \nabla, \nabla^*)$. D is the canonical divergence if and only if $D(p||q) + D(q||r) = D(p||r) + \left(\theta^i(p) - \theta^i(q)\right)(\eta_i(r) - \eta_i(q))$, where $\{[\theta^i], [\eta_i]\}$ are any mutually dual affine coordinate systems.*

Theorem 6.31 (Pythagorean relation)**.** *Let D be the canonical divergence of $(\mathcal{S}, g, \nabla, \nabla^*)$. Let $p, q, r \in \mathcal{S}$. Let γ_1 be the ∇-geodesic connecting p and q, and γ_2 be the ∇^*-geodesic connecting q and r. If γ_1 and γ_2 are orthogonal at q, then:*

$$D(p||r) = D(p||q) + D(q||r).$$

6.2.4 The dualistic structure of exponential families

Consider an exponential family:

$$p(x;\theta) = \exp\left[C(x) + \sum_{i=1}^{c} \theta^i F_i(x) - \psi(\theta)\right].$$

The **natural parameters** $[\theta^i]$ form a 1-affine coordinate system. Define:

$$\eta_i(\theta) = \mathbb{E}_\theta[F_i] = \int F_i(x) p(x;\theta) dx.$$

We have $\eta_i = \partial_i \psi$ and $g_{ij} = \partial_i \partial_j \psi$. Thus, $[\eta_i]$ form a (-1)-affine coordinate system dual to $[\theta^i]$ with potential ψ. The $[\eta_i]$ are called the **expectation parameters** or the **dual parameters**.
The dual potential ϕ is given by:

$$\begin{aligned}\phi(\theta) &= \theta^i \eta_i(\theta) - \psi(\theta),\\ &= \mathbb{E}_\theta[\log p_\theta - C],\\ &= -H(p_\theta) - \mathbb{E}_\theta[C],\end{aligned}$$

where H is the **entropy**.

Example 6.32. We consider $\mathcal{P}(\mathcal{X})$ for $\mathcal{X} = \{x_0, x_1, \ldots, x_c\}$ finite.
$C(x) = 0$, $F_i(x) = \delta_{x_i}(x)$.
We have for $i = 1, \ldots, c$:

$$\eta_i = \mathbb{E}_\theta[F_i] = \mathbb{E}_\theta[\delta_{x_i}(x)] = p(x_i).$$

Let us write $\eta_0 = p(x_0)$. Then:

$$\theta^i = \log \frac{\eta_i}{\eta_0}.$$

Let us write $\theta_0 = \log(1) = 0$. We have for $i = 1, \ldots, c$:

$$\eta_i = \frac{\exp \theta^i}{1 + \sum_{j=1}^{c} \exp \theta^j} = \frac{\exp \theta^i}{\sum_{j=0}^{c} \exp \theta^j}.$$

We recover the softmax function.
We have $\psi(\theta) = -\log p(x_0) = -\log\left(1 - \sum_{i=1}^{c} \eta^i\right) = \log\left(1 + \sum_{j=1}^{c} \exp \theta^j\right)$. The dual potential is:

$$\phi(\theta) = -H(p_\theta) = \sum_{i=0}^{c} \eta_i \log \eta_i = \frac{\sum_{i=1}^{c} \theta^i \exp \theta^i}{1 + \sum_{i=1}^{c} \exp \theta^i} - \psi(\theta).$$

Finally, after simplification:

$$p(x;\theta) = \exp\left[\sum_{i=1}^{c} \theta^i \delta_{x_i}(x) - \log\left(1 + \sum_{j=1}^{c} \exp \theta^j\right)\right] = \prod_{i=0}^{c} \eta_i^{\delta_{x_i}(x)}.$$

Since $\partial_i \log p = F_i$, we have the following expression for the Fisher information matrix:

$$g_{ij}(\theta) = \mathbb{E}_\theta\left[(F_i - \eta_i)(F_j - \eta_j)\right]. \tag{15}$$

The function $\hat{\eta}(x) = [F_1(x), \ldots, F_c(x)]$ can be seen as an unbiased estimator of η. By Equation 15, its covariance matrix is equal to the Fisher information matrix $G = [g_{ij}]$. This G is the Fisher information matrix in the coordinate system θ, thus it is the inverse of the Fisher information matrix in coordinates η. Hence, $\hat{\eta}$ achieves the equality in the Cramér-Rao inequality, i.e., it is an efficient estimator. To summarize: if $\mathcal{S}$ is an exponential family parameterized by a m-affine coordinate system, then it has an efficient estimator. The converse is also true:

Theorem 6.33. *A model $\mathcal{S} = \{p_\eta\}$ has an efficient estimator for the coordinate system $[\eta]$ if and only if $\mathcal{S}$ is an exponential family and η is m-affine.*

Proof. We already show one implication ($\Leftarrow$). Now, we show the other ($\Rightarrow$) on the assumption that $\mathcal{X}$ is finite (the statement is still true if $\mathcal{X}$ is infinite).
We already showed in subsection 6.1.4 than if $\mathcal{S}$ has an efficient estimator for η then $\mathcal{S}$ must be an exponential family. It remains to show that $[\eta]$ is m-affine. In subsection 6.1.4, we also showed that there is c linearly independent e-parallel vector fields $X_1, \dots, X_c$ such that $(X_i)_\eta^{(e)} = \hat{\eta}^i - \eta^i$. Let $\partial^i = \partial/\partial\eta^i$. Using $\mathbb{E}\left[\partial^i \log p_\eta\right] = 0$, we obtain:

$$\begin{aligned}\langle \partial^i, X_j \rangle &= \mathbb{E}_\eta\left[(\partial^i \log p_\eta)(\hat{\eta}_j - \eta_j)\right] = \mathbb{E}\left[(\partial^i \log p_\eta)\hat{\eta}_j\right], \\ &= \sum_x (\partial^i p_\eta)\hat{\eta}_j = \partial^i \sum_x p_\eta \hat{\eta}_j, \\ &= \partial^i \mathbb{E}_\eta\left[\hat{\eta}_j\right] = \partial^i \eta_j = \delta^i_j.\end{aligned}$$

In other words, the inner product between ∂^i (basis vectors of $[\eta]$) and any e-parallel vector field is constant. Thus, by the duality, ∂^i must be m-parallel and consequently $[\eta]$ is m-affine. □

Now, we consider the canonical divergence $D^{(1)}$. It is the dual of the KL divergence $D^{(-1)}$ that we simply denote D. Using the Pythagorean relation, we have that the solution of:

$$\min_{q \in M} D(p||q)$$

is the $\nabla^{(m)}$-projection of p on M. This problem is classical in statistics where it is linked with *maximum likelihood estimation.*
Similarly, the solution of:

$$\min_{q \in M} D(q||p)$$

is the $\nabla^{(e)}$-projection of p on M. This problem is important in the *large variation theory* via Sanov's theorem (I don't know what this is).

Example 6.34. We consider once again the exponential family $\mathcal{S} = \{p_\theta\}$. For any $q \in \mathcal{P}(\mathcal{X})$ and any θ:

$$D(q||p_\theta) = H(p_\theta) + \mathbb{E}_{p_\theta}[C] + \theta^i \mathbb{E}_{p_\theta}[F_i] - H(q) - \mathbb{E}_q[C] - \theta^i \mathbb{E}_q[F_i].$$

Thus:

$$\max_{q \in \mathcal{P}(\mathcal{X})} \left\{H(q) + \mathbb{E}_q[C] + \theta^i \mathbb{E}_q[F_i]\right\} = H(p_\theta) + \mathbb{E}_{p_\theta}[C] + \theta^i \mathbb{E}_{p_\theta}[F_i] = \psi(\theta). \tag{16}$$

Let $\lambda = (\lambda_1, \dots, \lambda_c) \in \mathbb{R}^c$. Consider the submanifold defined by the knowledge of the c first moments:

$$M_\lambda = \left\{q \in \mathcal{P}(\mathcal{X}) : \mathbb{E}_q[F_i] = \lambda_i, i = 1, \dots, c\right\}.$$

M_λ is a mixture family because it is defined by linear constraints (why?). Now, assume that $\mathcal{S} \cap M_\lambda \neq \varnothing$, i.e., there is a θ_λ such that $\eta_i(\theta_\lambda) = \mathbb{E}_{\theta_\lambda}[F_i] = \lambda_i$ for all i. Then, using Equation 16:

$$\begin{aligned}\max_{q \in M_\lambda} \left\{H(q) + \mathbb{E}_q[C]\right\} &= H(p_{\theta_\lambda}) + \mathbb{E}_{p_{\theta_\lambda}}[C], &(17)\\ &= \psi(\theta_\lambda) - \theta^i_\lambda \lambda_i, &(18)\\ &= -\phi(\theta_\lambda), &(19)\\ &= \min_\theta \left\{\psi(\theta) - \theta^i \lambda_i\right\}, &(20)\end{aligned}$$

where we used Equation 14 and Proposition 6.29 giving the duality relation between ϕ and ψ.
When $C = 0$, we get:

$$\max_{q \in M_\lambda} H(q) = H(p_{\theta_\lambda}).$$

This is the **principle of maximum entropy**. In statistical physics, where one moment is known (the average energy), the solution p_{θ_λ} is the Boltzmann-Gibbs distribution. It characterizes the thermal equilibrium state, i.e., the state which maximizes the thermodynamical entropy.
When $C(x) = \log p(x)$ for some distribution $p \in \mathcal{P}(\mathcal{X})$, Equation 17 becomes:

$$\min_{q \in M_\lambda} D(q||p) = D(p_{\theta_\lambda}||p) = \max_\theta \left\{\theta^i \lambda_i - \psi(\theta)\right\}. \tag{21}$$

Then, we have:

$$\psi(\theta) = \log \mathbb{E}_p \left[\exp\left(\theta^i F_i\right)\right].$$

This is the **cumulant generating function** of p w.r.t. the random variables $F_1, \ldots, F_c$.
The mixutre family M_λ and the exponential family $\mathcal{S}$ intersects orthogonally at p_{θ_λ}. Thus, Equation 21 is a consequence of the Pythagorean relation.

Remark. The KL divergence has two mutually dual integral representations. Let p_0 and p_1 be two distributions. We can connect them by two different curves:

$$p_t^{(m)} = (1-t)p_0 + tp_1,$$

which is a mixture family, and:

$$p_t^{(e)} = p_0^{1-t} p_1^t / Z_t,$$

which is an exponential family (Z_t is the normalizing constant). Let $g^{(m)}(t)$ and $g^{(e)}(t)$ be the Fisher informations of $\{p_t^{(m)}\}$ and $\{p_t^{(e)}\}$. Then:

$$D(p_1||p_0) = \iint_{0 \le s \le t \le 1} g^{(m)}(s) ds dt,$$
$$D(p_0||p_1) = \iint_{0 \le s \le t \le 1} g^{(e)}(s) ds dt.$$

6.2.5 Mutually dual foliations

Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a n-dimensional dually flat space, and let $[\theta]$ and $[\eta]$ be mutually dual coordinate systems. Let us divide the range of indexes $i = 1, \ldots, n$ into section I, $i = 1, \ldots, k$ and section II, $i = k+1, \ldots, n$. Given a parameter $c \in \mathbb{R}^k$, define the foliation:

$$M(c) = \{p \in \mathcal{S} : \eta_1(p) = c_1, \ldots, \eta_k(p) = c_k\}.$$

For all c, $M(c)$ is a ∇^*-autoparallel submanifold.
Similarly, given $d \in \mathbb{R}^{n-k}$, define the foliation:

$$E(d) = \left\{p \in \mathcal{S} : \theta^{k+1}(p) = d^{k+1}, \ldots, \theta^n(p) = d^n\right\}.$$

For all d, $E(d)$ is a ∇-autoparallel submanifold.
For any c and d, there is a unique intersection point $p \in M(c) \cap E(d)$. Moreover, $T_p(E(d))$ and $T_p(M(c))$ are orthogonal. Thus, the foliations E and M are said to be **mutually dual**.
We can divide the coordinates as $\eta = (\eta_I, \eta_{II})$ and $\theta = (\theta^I, \theta^{II})$, such that $\{p\} = M(\eta_I) \cap E(\theta^{II})$.
Now, define the **mixed coordinate system** $\xi = (\eta_I, \theta^{II})$.

Proposition 6.35. *Let D be the (g, ∇^*)-divergence. Let p, q, r, r' be points with mixed coordinates $(\eta_I(p), \theta^{II}(p)), (\eta_I(q), \theta^{II}(q)), (\eta_I(p), \theta^{II}(q)), (\eta_I(q), \theta^{II}(p))$.*
Then r is the ∇^*-projection of q onto $M(\eta_I(p))$ and r' is the ∇-projection of q onto $E(\theta^{II}(p))$. In particular:

$$D(p||q) = D(p||r) + D(r||q),$$
$$D(q||p) = D(q||r') + D(r'||p).$$

6.2.6 The triangular relation

6.3 Statistical inference and differential geometry

Assume we have data generated from some unknown probability distribution. **Statistical inference** is the process of extracting information concerning the underlying distribution from this data. If we have prior knowledge about the distribution, we can restrict the candidates to a parameterized family called a **statistical model**. Our goal here is to reformulate classical inference of **estimation** and **hypothesis testing** using geometry.

Let $\mathcal{S} = \{p(x;\eta)\}$ be a family of probability distributions, sufficiently regular to be seen as an n-dimensional manifold.
Let $x^N = (x_1, \dots, x_N)$ be N iid sampled from $p(x;\eta)$.
Statistical inference aims at inferring $p(x;\eta)$ given x^N.
Estimation aims at finding an estimate $\hat{\eta}$ of η.
Testing aims at deciding if the hypothesis $H_0 : \eta = \eta_0$ is accepted against the alternative hypothesis $H_1 : \eta \neq \eta_0$.

6.3.1 Estimation based on independent observations

The distribution of x^N is $p_N\left(x^N;\eta\right) = \prod_{t=1}^N p(x_t;\eta)$. We obtain a manifold $\mathcal{S}_N = \{p_N\left(x^N;\eta\right)\}$ with the same coordinate system.

Proposition 6.36.

$$\begin{aligned} g_{ij}^N(\eta) &= N g_{ij}(\eta), \\ \Gamma_{ij,k}^{(\alpha)\,N}(\eta) &= N\Gamma_{ij,k}^{(\alpha)}(\eta), \\ \partial_i^N &= \sqrt{N}\partial_i. \end{aligned}$$

Proof. I will only prove the first equality. The second one can be obtained similarly from the definition. I am not sure how to obtain the third one (maybe as a consequence of the first one?).

$$\begin{aligned} g_{ij}^N(\eta) &= \mathbb{E}_\eta\left[\sum_{t_1=1}^N \partial_i l(x_{t_1};\eta) \sum_{t_2=1}^N \partial_j l(x_{t_1};\eta)\right] = \sum_{t_1=1}^N \sum_{t_2=1}^N \mathbb{E}_\eta\left[\partial_i l(x_{t_1};\eta)\partial_j l(x_{t_2};\eta)\right], \\ &= \sum_{t=1}^N \mathbb{E}_\eta\left[\partial_i l(x_t;\eta)\partial_j l(x_t;\eta)\right] = N g_{ij}(\eta), \end{aligned}$$

where the first equality of the second line comes from the independence of x_{t_1} and x_{t_2} when $t_1 \neq t_2$:

$$\mathbb{E}_\eta[\partial_i l(x_{t_1};\eta)\partial_j l(x_{t_2};\eta)] = \mathbb{E}_\eta[\partial_i l(x_{t_1};\eta)]\mathbb{E}_\eta[\partial_j l(x_{t_2};\eta)] = 0,$$

using $\mathbb{E}_\eta[\partial_i l(x;\eta)] = 0$. □

Hence, the geometry of $\mathcal{S}_N$ is the same as $\mathcal{S}$ scaled by a factor of N. Thus, we can simply consider the geometry of $\mathcal{S}$.
An *estimator* $\hat{\eta}$ is a function of the N data points: $\hat{\eta} = \hat{\eta}\left(x^N\right)$. Since x^N is a random variable, then so is $\hat{\eta}$.
The estimator $\hat{\eta}$ is unbiased if $\mathbb{E}_\eta\left[\hat{\eta}\right] = \eta$ for all η. Here, the expectation is taken with respect to $p_N\left(x^N;\eta\right)$.
Define the *mean square error* $V_\eta\left[\hat{\eta}\right] = \left[v_\eta^{ij}\right]$:

$$v_\eta^{ij} = \mathbb{E}_\eta\left[\left(\hat{\eta}^i - \eta^i\right)\left(\hat{\eta}^j - \eta^j\right)\right].$$

If $\hat{\eta}$ is unbiased, then $V_\eta\left[\hat{\eta}\right]$ is the covariance matrix. **Neither the unbiasedness nor the mean square error are geometrically invariant criteria. They depend on the choice of coordinates η.**
The Cramér-Rao inequality becomes:

$$\left[v_\eta^{ij}\right] \succeq \frac{1}{N}\left[g^{ij}(\eta)\right].$$

We know that there exists an efficient estimator if and only if $\mathcal{S}_N$ is an exponential family and η is an m-affine coordinate system. In fact, this is also equivalent with $\mathcal{S}$ being an exponential family and η being an m-affine coordinate system.

Asymptotic theory. Now, we are interested in the performance of an estimator when $N \to \infty$. Here, unbiasedness is not as meaningful as when N is fixed. Instead, we say that a sequence of estimators $\hat{\eta}_N$ is **consistent** if for any η, $\hat{\eta}_N$ converges in probability to η when $N \to \infty$, i.e., for all $\epsilon > 0$:

$$\lim_{N\to\infty} \mathbb{P}_\eta \left(|\hat{\eta}_N - \eta| > \epsilon\right) = 0.$$

Under some regularity conditions, a consistent estimator is asymptotically unbiased and we have the **asymptotic Cramér-Rao inequality:**

$$\lim_{N\to\infty} N\left[v_\eta^{ij}\left[\hat{\eta}_N\right]\right] \succeq \left[g^{ij}(\eta)\right].$$

If the equality is achieved for all η, the estimator is called an **asymptotically efficient estimator** or **first-order efficient estimator**. Such an estimator is optimal w.r.t. mean square error with a correction term of order $1/N$ (hence the "first-order" name). Unlike the case for finite N, there always exists an asymptotically efficient estimator for an arbitrary model.

Definition 6.37. For fixed x^N, define the **likelihood function** $\eta \mapsto p_N\left(x^N;\eta\right)$.
The **maximum likelihood estimator** $\hat{\eta}_{mle}$ satisfies:

$$p_N\left(x^N;\hat{\eta}_{mle}\right) = \max_\eta p_N\left(x^N;\eta\right).$$

Theorem 6.38. *The maximum likelihood estimator $\hat{\eta}_{mle}$ is asymptotically efficient, i.e., for all η:*

$$\lim_{N\to\infty} N v_\eta^{ij}\left[\hat{\eta}_{mle}\right] = g^{ij}(\eta).$$

More precisely, $\hat{\eta}_{mle}$ converges in distribution to a Gaussian with mean η and covariance $N^{-1}\left[g^{ij}\right]$.

The mean square error of an asymptotically efficient estimator is:

$$v_\eta^{ij}\left[\hat{\eta}\right] = \frac{1}{N} g^{ij}(\eta) + O\left(\frac{1}{N^2}\right).$$

The study of the order $1/N^2$ term is called **higher-order asymptotic theory** of statistical estimation.

6.3.2 Exponential families and observed points

Consider the exponential family $p(x;\theta) = \exp\left(C(x) + \theta^i F_i(x) - \psi(\theta)\right)$. The n functions $F_i(x)$ can be seen as random variables. Hence, we rename them $x_i = F_i(x)$. We also rename[23] $x = (x_1, \ldots, x_n)$. Moreover, we redefine the pdf using the n-dimensional random variable x (different from the previous x) w.r.t. the dominating measure $d\mu(x) = \exp\left(C(x)\right) dx$ instead of dx. Then, without loss of generality, the exponential family can be written:

$$p(x;\theta) = \exp(\theta^i x_i - \psi(\theta)).$$

Remember that we have a m-affine coordinate system:

$$\eta_i = \mathbb{E}_\theta\left[x_i\right],$$
$$\mathbb{E}_\theta\left[(x_i - \eta_i)(x_j - \eta_j)\right] = g_{ij}(\theta).$$

Consider N iid observations $x^N = x_1, \ldots, x_N$ distributed according to some p_θ in the exponential family. Let[24]:

$$\overline{x} = \frac{1}{N}\sum_{t=1}^N x_t.$$

[23] Be careful of any confusion. With this notation, x is no longer a random variable over the sample space, but a n-dimensional random vectors.

[24] Once again, confusing notations. For each t, $x_t \in \mathbb{R}^n$ is the t-th observed vector. While $\overline{x}_i \in \mathbb{R}$ is the i-th component of $\overline{x} \in \mathbb{R}^n$.

The pdf of x^N is:

$$p_N\left(x^N;\theta\right) = \prod_{t=1}^{N} p(x_t;\theta) = \exp\left(N\left(\theta^i \overline{x}_i - \psi(\theta)\right)\right),$$

which is also an exponential family with natural parameters θ. Here we have $x^N \in \mathbb{R}^{nN}$. However, we were able to express $p_N\left(x^N;\theta\right)$ using only $\overline{x} \in \mathbb{R}^n$. This means that $\overline{x}$ **is a sufficient statistic with respect to the exponential family**. Thus, statistical inference based on x^N can be reduced to inference based on $\overline{x}$ without compromsing the quality of the result. This is a consequence of the **Rao-Blackwell theorem**.

Now, consider the point in $\mathcal{S}$ with coordinates:

$$\hat{\eta} = \overline{x}.$$

This is called the **observed point**. We have[25]:

$$\begin{aligned}\mathbb{E}_\theta[\overline{x}] &= \eta,\\ \mathbb{E}_\theta[(\overline{x}_i - \eta_i)(\overline{x}_j - \eta_j)] &= \frac{1}{N} g_{ij}(\theta),\end{aligned}$$

hence $\hat{\eta}$ is an efficient estimator of η. According to the CLT, $\hat{\eta} = \frac{1}{N}\sum_{t=1}^{N} x_t$ is asymptotically distributed according to a normal distribution. Moreover, the observed point $\hat{\eta}$ is the maximum likelihood estimator of the model. Indeed, denoting $\hat{\theta}$ the natural parameters dual to $\hat{\eta}$, we have for any θ:

$$\begin{aligned}\log p_N\left(x^N;\hat{\theta}\right) - \log p_N\left(x^N;\hat{\theta}\right) &= N\left((\hat{\theta}^i - \theta^i)\hat{\eta}_i - \psi(\hat{\theta}) + \psi(\theta)\right),\\ &= N\left(\phi(\hat{\eta}) + \psi(\theta) - \theta^i\hat{\eta}_i\right),\\ &= N\, D\left(p_{\hat{\theta}} || p_\theta\right) \geq 0,\end{aligned}$$

using $\phi(\hat{\eta}) = \hat{\theta}^i\hat{\eta}_i - \psi(\hat{\theta})$ and the KL divergence $D(p||q) = \phi(p) + \psi(q) - \theta^i(q)\eta_i(p)$.

6.3.3 Consistency and first-order efficiency

Definition 6.39. A (n,m)**-curved exponential family** is a smooth m-dimensional submanifold of an n-dimensional exponential family.

Let M be a curved exponential family with the coordinate system $u = [u^a]$. The pdf in M are:

$$p(x;u) = \exp\left(\theta^i(u)x_i - \psi(\theta(u))\right).$$

$\hat{\eta} = \overline{x}$ is also a sufficient statistic for M, thus we only need functions of $\hat{\eta}$ for the estimators $\hat{u} = f(\hat{\eta})$ of u in $p(x;u)$. In other words, an estimator is a mapping $f : \mathcal{S} \to M$ such that $\hat{\eta} \mapsto \hat{u} = f(\hat{\eta})$. The set:

$$A(u) = f^{-1}(u) = \{\eta : f(\eta) = u\},$$

is in general a $(n-m)$-dimensional submanifold of $\mathcal{S}$. It is called the **estimating submanifold** corresponding to $u \in M$. Selecting an estimator f decomposes the space $\mathcal{S}$ into a collection of estimating submanifolds. The characteristics of an estimator are entirely determined by the set of estimating submanifolds $\{A(u)\}$ and the shape of M.

Geometry of the estimator $\hat{u} = f(\hat{\eta})$. Assume that the true distribution is u. Then, the expected value of the random variable x is $\eta(u)$. By the law of large numbers, $\overline{x}$ converges with probability 1 to $\eta(u)$. Hence:

Theorem 6.40. *f is consistent if and only if $\eta(u) \in A(u)$.*

[25] using the fact that the Fisher information matrix w.r.t. the η-coordinate system is the inverse of the Fisher information matrix w.r.t. the θ-coordinate system denoted $g_{ij}(\theta)$.

Since $\eta(u) \in M$, this is equivalent to $\{\eta(u)\} = M \cap A(u)$.
Remark: it is possible to consider an estimator such that $A(u) = A_N(u)$ depends on N. In this case, consistency is equivalent to $\eta(u) \in \lim_{N\to\infty} A_N(u)$.

Now, we assume that $\hat{u}$ is consistent and we want to study its estimation error.
$\hat{\eta}$ converges to $\eta(u)$, thus we may consider the tangent space of $\eta(u)$ in $\mathcal{S}$ and linearize in this space the possible value of $\hat{\eta}$. Define:

$$\widetilde{x} = \sqrt{N}\left(\overline{x} - \eta(u)\right).$$

By the CLT, $\widetilde{x}$ asymptotically follows a centered normal distribution with covariance matrix $[g_{ij}(\eta(u))]$, which is also the metric in $T_{\eta(u)}(\mathcal{S})$ w.r.t. the θ-coordinate system.
Define a coordinate system v on $A(u)$ such that u is the origin. Then, a point $\eta \in \mathcal{S}$ can be indexed by $w = (u, v)$. We have $M = \{\eta(u, 0)\}$.
Let $\hat{w} = (\hat{u}, \hat{v})$ be the w-coordinates of the observed point $\hat{\eta}$. Since $\hat{u}$ is close to u and $\hat{v}$ is close to 0, let us consider:

$$\widetilde{u} = \sqrt{N}\left(\hat{u} - u\right),\ \widetilde{v} = \sqrt{N}\hat{v},\ \widetilde{w} = (\widetilde{u}, \widetilde{v}).$$

The Taylor expansion of $\hat{\eta} = \eta(\hat{w})$ around $w = (u, 0)$ yields:

$$\widetilde{x}_i = B_{\alpha i}\widetilde{w}^\alpha + \frac{1}{2\sqrt{N}} C_{\alpha\beta i}\widetilde{w}^\alpha\widetilde{w}^\beta + \frac{1}{6N} D_{\alpha\beta\gamma i}\widetilde{w}^\alpha\widetilde{w}^\beta\widetilde{w}^\gamma + O\left(\frac{1}{N\sqrt{N}}\right), \tag{22}$$

where B, C and D are derivatives of η w.r.t. w:

$$B_{\alpha i} = \partial_\alpha \eta_i(u, 0),\ C_{\alpha\beta i} = \partial_\alpha\partial_\beta \eta_i(u, 0),\ D_{\alpha\beta\gamma i} = \partial_\alpha\partial_\beta\partial_\gamma \eta_i(u, 0).$$

In $T_{\eta(u)}(\mathcal{S})$, define:

$$\begin{aligned}
e^i &= \frac{\partial}{\partial \eta_i} = \partial^i, \\
e_\alpha &= \frac{\partial}{\partial w^\alpha} = \partial_\alpha, \\
e_a &= \frac{\partial}{\partial u^a} = B_{ai}e^i,\ a = 1, \ldots, m \\
e_\kappa &= \frac{\partial}{\partial v^\kappa} = B_{\kappa i}e^i,\ \kappa = m+1, \ldots, n, \\
g^{ij} &= \langle e^i, e^j \rangle, \\
g_{\alpha\beta} &= \langle e_\alpha, e_\beta \rangle = B_{\alpha i}B_{\beta j}g^{ij}.
\end{aligned}$$

In the asymptotic theory of estimation, we typically drop the terms smaller than $O\left(1/\sqrt{N}\right)$ keeping only the linear approximation:

$$\widetilde{w}^\alpha = B^{\alpha i}\widetilde{x}_i, \tag{23}$$

where:

$$B^{\alpha i} = g^{\alpha\beta}g^{ij}B_{\beta j}.$$

Be careful, here $[g^{\alpha\beta}]$ is the inverse of $[g_{\alpha\beta}]$ which is different from $[g^{ij}]$.
Since $\widetilde{x}$ asymptotically follows a centered normal distribution with covariance matrix $[g_{ij}]$, we have that $\widetilde{w}^\alpha$ follows a centered normal distribution with covariance matrix $[g^{\alpha\beta}]$ (why?).
The asymptotic mean square error of the estimator $\hat{u}$ scaled by N is:

$$\lim_{N\to\infty} N\mathbb{E}\left[(\hat{u}^a - u^a)(\hat{u}^b - u^b)\right] = \mathbb{E}\left[\widetilde{u}^a\widetilde{u}^b\right] := \overline{g}^{ab},$$

where $[\overline{g}^{ab}]$ is the submatrix of $[g^{\alpha\beta}]$ for $a, b = 1, \ldots, m$, and $\mathbb{E}$ is the expectation with respect to $p(x; u)$.

Proposition 6.41.

$$[\overline{g}^{ab}] = [g_{ab} - g_{a\kappa}g^{\kappa\lambda}g_{b\lambda}]^{-1}$$

Proof. Seems to be a classic formula.
The following is not mathematically correct because the block dimensions don't match, but the intuition might be to apply the inverse of of 2×2 matrix to a block matrix. We get:

$$\begin{aligned}[\overline{g}^{ab}] &= [g_{\kappa\lambda}][g_{ab}g_{\kappa\lambda} - g_{a\kappa}g_{b\lambda}]^{-1}, \\ &= [g_{ab} - g_{a\kappa}g^{\kappa\lambda}g_{b\lambda}]^{-1}.\end{aligned}$$

□

Amari & Nagaoka claim that it is obvious (not for me) than, according to the previous proposition:

$$\overline{g}^{ab} \succeq g^{ab},$$

where g^{ab} is the Fisher matrix in the η-coordinates. This corresponds to the asymptotic Cramér-Rao inequality.
The estimator is asymptotically efficient (i.e., we have equality in the last equation) if and only if $g_{a\kappa} = 0$, which can be restated as:

Theorem 6.42. *A consistent estimator is first-order asymptotically efficient if and only if $A(u)$ and M are orthogonal.*

If the estimator depends on N, the theorem becomes: $\lim_{N\to\infty} A_N(u)$ and M are orthogonal.

Now, we show the properties of the maximum likelihood estimator from a geometric point of view. The KL divergence between the observed point $\hat{\eta}$ and another point $\eta(u) \in M$ is:

$$\begin{aligned}D\left(\hat{\eta}||\eta(u)\right) &= \phi(\hat{\eta}) + \psi(\theta(u)) - \theta^i(u)\hat{\eta}_i, \\ &= \phi(\hat{\eta}) - \frac{1}{N}\log p_N\left(x^N; u\right).\end{aligned}$$

Since $\phi(\hat{\eta})$ does not depend on u, then the point $\hat{u}_{mle} \in M$ which minimizes the divergence with $\hat{\eta}$ is the point which maximizes the likelihood $p_N\left(x^n; u\right)$. Remember that the point which minimizes the divergence is the orthogonal projection of $\hat{\eta}$ onto M along an m-geodesic. In other words, the estimating submanifold $A(u)$ of $\hat{u}_{mle}$ is m-autoparallel and orthogonal to M. Thus, **the maximum likelihood estimator is asymptotically efficient**.
More generally, consider any divergence D on $\mathcal{S}$ which induces the Fisher metric (or a constant multiple). Define the estimator $\hat{u}$:

$$D\left(\hat{\eta}||\eta(\hat{u})\right) = \min_{u\in M} D\left(\hat{\eta}||\eta(u)\right).$$

The estimating submanifold of $\hat{u}$ at u is:

$$A(u) = \left\{\eta : D\left(\eta||(\partial_a)_u\right) = 0, \forall a\right\},$$

because of the first-order necessary optimality condition. Thus, for $w = (u, v) \in A(u)$, we have:

$$\begin{aligned}g_{a\kappa} &= -D[\partial_\kappa||\partial_a], \\ &= -\frac{\partial}{\partial v^\kappa}\frac{\partial}{\partial u^a}D(v||u), \\ &= -\frac{\partial}{\partial v^\kappa}D(\eta(u,v)||(\partial_a)_u) = 0.\end{aligned}$$

Hence $\hat{u}$ is asymptotically efficient. This result was first shown by Eguchi.

6.3.4 Higher-order asymptotic theory of estimation

If we do not drop the higher order terms $1/\sqrt{N}$ and $1/N$, then we need to consider the curvature as well as the tangent space of $A(u)$. The goal is to compare various efficient estimators.

Using the higher order terms in Equation 22, we can rewrite Equaion 23 with higher order terms:

$$\widetilde{w}^\alpha = B^{\alpha i}\widetilde{x}_i - \frac{1}{2\sqrt{N}}C^\alpha_{\beta\gamma}\widetilde{w}^\beta\widetilde{w}^\gamma - \frac{1}{6N}D^\alpha_{\beta\gamma\delta}\widetilde{w}^\beta\widetilde{w}^\gamma\widetilde{w}^\delta + O\left(\frac{1}{N\sqrt{N}}\right),$$

where $C^\alpha_{\beta\gamma} = B^{\alpha i}C_{\beta\gamma i}$ and $D^\alpha_{\beta\gamma\delta} = B^{\alpha i}D_{\beta\gamma\delta i}$.
We know that:

$$\mathbb{E}[\widetilde{x}_i] = 0,$$
$$\mathbb{E}[\widetilde{w}^\beta\widetilde{w}^\alpha] = g^{\beta\alpha} + O\left(\frac{1}{N}\right).$$

Thus, defining $C^\alpha = C^\alpha_{\beta\gamma}g^{\beta\gamma}$, we have:

$$\mathbb{E}[\widetilde{w}^\alpha] = -\frac{1}{2\sqrt{N}}C^\alpha + O\left(\frac{1}{N}\right).$$

Hence, even if $\eta(u) \in A(u)$, the estimator $\hat{u}$ has a bias of order[26] $1/N$ whose coefficients are $C^a(u)$. The bias converges to 0 as $N \to \infty$.
To compensate for this bias, we define the **bias-corrected estimator**:

$$\hat{u}^{*a} = \hat{u}^a + \frac{1}{2N}C^a(\hat{u}).$$

Note that we use $C^a(\hat{u})$ instead of $C^a(u)$ which is unknown. Its bias is:

$$\mathbb{E}[\hat{u}^*] - u = O\left(\frac{1}{N^2}\right).$$

According to Wikipedia[27], we can write the bias correction explicitly:

$$\hat{b}^a = -\frac{1}{2N}C^\alpha(\hat{u}) = \frac{1}{N}g^{ab}g^{cd}\mathbb{E}_{\hat{u}}\left[\frac{1}{2}\frac{\partial^3 \log p_{\hat{u}}}{\partial u^b \partial u^c \partial u^d} + \frac{\partial \log p_{\hat{u}}}{\partial u^c}\frac{\partial^2 \log p_{\hat{u}}}{\partial u^b \partial u^d}\right].$$

Using complicated stuff (Hermite polynomials, Edgeworth expansion), it can be shown that:

Theorem 6.43. *The mean square error of a bias-corrected first-order efficient estimator is:*

$$\mathbb{E}\left[(\hat{u}^{*a} - u^a)(\hat{u}^{*b} - u^b)\right] = \frac{1}{N}g^{ab} + \frac{1}{2N^2}K^{ab} + O\left(\frac{1}{N^3}\right),$$

where K^{ab} is a function of:

- $\Gamma^{(m)}_M$ *the m-connection coefficients of M,*
- $H^{(e)}_M$ *the embedding e-curvature of M,*
- $H^{(m)}_A$ *the embedding m-curvature of $A(u)$.*

Remarks:

1. $\Gamma^{(m)}_M$ is not a tensor (it depends on the choice of coordinate system u for M). This reflects the fact that the square error is also not a tensor. However, once the coordinates u are chosen, this value is invariant on the choice of estimator.
   Using normal coordinates for the m-connection, it is always possible to chose u such that $\Gamma^{(m)}_M$ is 0 for a particular point. However, for $\Gamma^{(m)}_M$ to be identically 0, M must be m-flat (which is equivalent to e-flat by the duality).

[26] Remember that $\widetilde{w}$ is defined with a factor $\sqrt{N}$.
[27] `https://en.wikipedia.org/wiki/Maximum_likelihood_estimation`

2. $H_M^{(e)}$ is 0 if and only if M is e-autoparallel in $\mathcal{S}$, i.e., if and only if M is an exponential family (since $\mathcal{S}$ is an exponential family). Thus, the deviation of M from being an exponential family (measured by $H_M^{(e)}$) will increase the estimation error.

3. $H_A^{(m)}$ is the only term that depends on the choice of an estimator. If $\hat{u}$ is first-order efficient and if the m-curvature of $A(u)$ is 0, we say that $\hat{u}$ is **second-order efficient**.

Theorem 6.44. *The bias-corrected maximum likelihood estimator $\hat{u}^*_{mle}$ is second-order efficient.*

This is because for the maximum likelihood estimator, $A(u)$ is m-autoparallel and orthogonal to M.

7 Information Geometry and its Applications

Reading notes from [9].

- What does information geometry bring that wasn't known in "classical" statistical theory? The answer is: the e-connection (and m-connection). And that's all. Information geometry doesn't bring anything new, but just provide a geometric interpretation (which may or may not provide new insights). Everything else can be expressed outside information geometry including: canonical and expectation parameters of exponential families, Legendre-Fenchel duality, Bregman divergences etc. The e-connection and m-connection give rise to the Pythagorean relation for KL divergence. Information geometry consists in reformulating all statistical problems as an orthogonality condition. The Pythagorean relation provides the link between geometry and statistics. The algorithms are provided by e-projection and m-projection.

- Higher-order statistical inference. For any model, the MLE is unbiased and asymptotically first-order efficient. if the bias is corrected to become $O(1/N^2)$ instead of $O(1/N)$, then the bias-corrected MLE is asymptotically second-order efficient. However, the MLE is not asymptotically third-order efficient in general. I'm not sure if this has any relevance regarding machine learning. For finite data, only exponential families with expectation parameters have a first-oder efficient estimator, which is the MLE. It seems that there are no bounding of the expected error for finite data, either for MLE nor MAP (see a recent paper). MLE seems also to be linked to overfitting but I don't know why. MLE is equivalent to MAP with uniform prior (or with no pseudo-observations, no sure why this is the same).

- EM algorithm is used when a part of the data sample is not observed, i.e., the data sample is (x_i, z_i) and we only observe x_i, $i = 1, \dots, n$. In information geometry, it is a special case of the *em*-algorithm. We have a data manifold D which is the set of all possible observed points given x. Equivalently, it is the set of joint distributions $q(x, z) = q(z|x)q(x)$ such that $q(x)$ is the observed point for x. We also have a model manifold M. We alternate e-projection from M to D, then m-projection from D to M to find the pair of points $(q^*, p^*) \in D \times M$ that minimize $KL[q, p]$, $q \in D$, $p \in M$.

- Neyman-Scott problem. What if some of the parameters changed from one observation to the other. We want to estimate the parameter that don't change. In this case, MLE fails. It might be the correct formulation for supervised learning, where the input x_i is seen as a varying parameter. However, contrary to the Neyman-Scott problem, x_i is known so I am not sure.

- Total Bregman divergence is more robust. This is for k-means clustering. If you add an erroneous data, the centroid should not move too much. This is false using Bregman divergence, but true using total Bregman divergence.

- Chernoff information can be used to measure the "distance" between two distributions. It is in the setting of determining if a data sample x comes from a distribution p or a distribution q. Chernoff information is linked to an α-divergence for some α which depends on p and q. This is the first time I see a statistical interpretation of α-divergence for $\alpha \neq -1, 0, 1$.

- How to choose the kernel function for kernel methods (such as SVM)? Using non-Euclidean volume elements might help. (This interpretation is unrelated to information geometry).

- Bayesian duality of exponential families and **conjugate priors**, i.e., prior and posterior belongs to the same family, which leads to closed-form solutions for the posterior, which is nice. It is particularly nice for exponential family, where the *hyperparameters* (parameters of the prior) can be interpreted as adding β pseudo-observations using an exponential family with canonical parameters α. The Dirichlet distribution is the conjugate prior of categorical distribution, i.e., if the prior is Dirichlet and the likelihood is categorical, then the posterior is Dirichlet. Dirichlet distributions are the multivariate generalization of beta distributions and are an exponential family.
- Lots of stuff about Boltzmann machines and restricted Boltzmann machines. I skip these parts because RBM are no longer used for unclear reasons (slow and difficult to train it seems). I might come back to it later since the maths seem interesting.
- MAP seems to be equivalent to regularization for certain priors. Choosing the right prior might be the solution of adversarial vulnerability and generalization. I don't know anything about this but I think it is an already well-studied topic (or is it?). So, I'm not sure how I could bring anything new here ... I may be able to provide theoretical justifications for regularization, as opposed to ad-hoc empirically justified regularizers. Ridge regularization is probably well-understood in the linear case. No sure for the general case.

Questions:

- Statistical meaning of Hellinger distance / Fisher-Rao distance?
- Link between MLE and overfitting?
- How to choose the prior in Bayesian machine learning?
- Ridge regression in the linear case? Regularization and prior in the general case?
- Hopfield networks?

8 Graphical Models, Exponential Families, and Variational Inference

Reading notes from [10].

8.1 Basics of convex sets and functions

A **cone** K is a set such that, for any $x \in K$, the **ray** $\{\lambda x : \lambda > 0\}$ also belongs to K.
The **conical hull** of a collection of vectors $\{x_1, \dots, x_n\}$:

$$\left\{ y \in \mathbb{R}^d : y = \sum_{i=1}^{n} \lambda_i x_i, \lambda_i \geq 0 \right\},$$

is a *convex* cone. The set of symmetric positive semidefinite matrices is also a convex cone.
An **affine combination** is a sum $\sum_{i=1}^{k} \alpha_i x_i$ with $\sum_{i=1}^{k} \alpha_i = 1$.
A **convex combination** is a sum $\sum_{i=1}^{k} \alpha_i x_i$ with $\sum_{i=1}^{k} \alpha_i = 1$ and $\alpha_i \geq 0$.
The **affine hull** $\text{aff}(S)$ of a set S is the smallest set that contains all affine combinations.
The **convex hull** $\text{conv}(S)$ of a set S is the smallest set that contains all convex combinations.
Denote by $\mathcal{B}_\epsilon(z)$ the Euclidean ball of center z and radius $\epsilon > 0$.
The **interior** of a set $C \subseteq \mathbb{R}^d$ is:

$$\mathring{C} = \{z \in C : \exists \epsilon > 0 \text{ s.t. } \mathcal{B}_\epsilon(z) \subset C\}.$$

The **relative interior** of a set C is:

$$\text{ri}(C) = \{z \in C : \exists \epsilon > 0 \text{ s.t. } \mathcal{B}_\epsilon(z) \cap \text{aff}(C) \subset C\}.$$

For example, the interior of $[0,1] \subset \mathbb{R}^2$ is empty, while the relative interior of $[0,1]$ is $(0,1)$.
The relative interior of a convex set is always nonempty.
A convex set $C \subseteq \mathbb{R}^d$ is **full-dimensional** if its affine hull is equal to $\mathbb{R}^d$. Thus, for a full-dimensional convex set, the notion of interior and relative interior coincide.
A **polyhedron** P is a set that can represented as the intersection of a finite number of half-spaces:

$$P = \{x \in \mathbb{R}^d : \langle a_j, x \rangle \leq b_j, \forall j \in \mathcal{J}\}.$$

A bounded polyhedron is called a **polytope**.
A point $x \in P$ is an **extreme point** if there are no $y, z \in P$ and $\lambda \in (0,1)$ such that $x = \lambda y + (1-\lambda)z$.
A point $x \in P$ is a **vertex** if there exists $c \in \mathbb{R}^d$ such that, for all $y \in P, y \neq x$, we have $\langle c, x \rangle > \langle c, y \rangle$.
For a polyhedron, x is a vertex if and only if it is an extreme point.
Any nonempty polytope can be written as the convex hull of its extreme points (consequence of Minkowski-Weyl theorem). This convex hull representation is dual to the half-space representation. Conversely, the convex hull of any finite collection of vectors is a polytope.
It is convenient to allow convex functions to take the value $+\infty$, in particular for dual calculations. An **extended real-valued function** f on $\mathbb{R}^d$ takes values in the extended real line $\mathbb{R}_* = \mathbb{R} \cup \{+\infty\}$.
The **domain** of an extended convex function f is $\text{dom}(f) = \{x \in \mathbb{R}^d : f(x) < +\infty\}$.

8.2 Exponential families

Exponential families provide a link between **statistical inference** and **convex analysis**. Some statistical computations (e.g., marginalization, MLE) can be understood in terms of **mapping between mean parameters and canonical parameters**.

8.2.1 Motivation: Maximum Entropy

Let X be a scalar random variable. Let $X^1, \ldots, X^n$ be n iid observations. We compute the **empirical expectations** of certain functions $\phi_\alpha : \mathcal{X} \to \mathbb{R}$:

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^{n} \phi_\alpha(X^i),$$

for all α in some set $\mathcal{I}$. Based on $\hat{\mu}$ of dimension $|\mathcal{I}|$, we want to infer the probability distribution of X, represented as a density p absolutely continuous wrt some measure ν.
A distribution p is **consistent** with the data if:

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathcal{X}} \phi_\alpha(x) p(x) \nu(dx) = \hat{\mu}_\alpha, \tag{24}$$

for all $\alpha \in \mathcal{I}$. Expectations under p are matched to expectations under the empirical distribution. How can we choose a distribution p among all the distributions that satisfy Equation 24? Answer: by using the **Principle of Maximum Entropy**.
Define a functional of the density p, called the **Shannon entropy**:

$$H(p) = -\int_{\mathcal{X}} \log(p(x)) p(x) \nu(dx).$$

The Principle of Maximum Entropy is:

$$p^* = \arg\max_{p \in \mathcal{P}(\mathcal{X})} H(p) \text{ subject to } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I}.$$

One interpretation is to choose the distribution with **maximal uncertainty** as measured by $H(p)$ while remaining faithful to the data. If the problem is feasible and some technical conditions are satisfied, it can be shown that the optimal solution p^* takes the form:

$$p_\theta(x) \propto \exp\left\{\sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x)\right\}.$$

The parameters θ can be interpreted as the Lagrange multipliers associated with the constraints specified by $\hat{\mu}$. The optimal solution is obtained by ordinary calculus in the discrete case, and by calculus of variations in the continuous case. If there is only one ϕ_α which is equal to $\phi_\alpha(x) = x$, then the optimal solution is a Boltzmann distribution/Gibbs measure (i.e., for a fixed temperature, the distribution of velocities of a gas is a Boltzmann distribution).

8.2.2 Basics of exponential families

Let $X = (X_1, \ldots, X_m)$ be a random vector taking values in $\mathcal{X}^m$. Let $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ be a collection of functions $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$ known as **potential functions** or **sufficient statistics**. We write $d = |\mathcal{I}|$ such that $\phi : \mathcal{X}^m \to \mathbb{R}^d$. Let $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$ be a vector of **canonical** or **exponential** parameters. An **exponential family** associated with ϕ is the parameterized collection of density functions:

$$p_\theta(x_1, \ldots, x_m) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\},$$

taken wrt $d\nu$. The quantity A is known as the **log partition function** or **cumulant function** defined by:

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp\langle \theta, \phi(x) \rangle \nu(dx).$$

This function plays a prominent role. When $A(\theta)$ is finite, p_θ is properly normalized. Thus, define:

$$\Omega = \{\theta \in \mathbb{R}^d : A(\theta) < +\infty\}.$$

We will see that A is convex, implying that Ω is a convex set.
An exponential family is **regular** when Ω is an open set. In information geometry, this assumption ensures that we can differentiate with respect to θ, i.e., the family is a **smooth manifold**.
An exponential family is **minimal** if there is *no* nonzero vector $a \in \mathbb{R}^d$ such that the function:

$$x \mapsto \langle a, \phi(x) \rangle = \sum_{\alpha \in \mathcal{I}} a_\alpha \phi_\alpha(x)$$

is constant ν-almost everywhere. Equivalently, the $(d+1)$ functions $(1, \phi_\alpha)$ are linearly independent. Equivalently, no sufficient statistic ϕ_α can be expressed as a linear combination (plus a constant) of the others. Then, we say that θ is a **minimal representation** of p_θ, i.e., there is a unique θ associated with each distribution, i.e., the parameters θ are identifiable. In information geometry, this assumption ensures that θ is a **coordinate system** (i.e., $\theta \mapsto p_\theta$ is a diffeomorphism).
An exponential family is **overcomplete** or **nonminimal** if it is not minimal. Then, each distribution is associated with an affine subset of parameters θ. This notion is claimed to be useful in understanding the "sum-product algorithm".
Remark concerning observed variables or conditioning for the marginalization problem. We often condition on a subset of random variables that represent observed quantities, and we look for the marginal under the posterior distribution $p_\theta(y|x)$. For ease of notation, we will no longer make reference to observed variables and discuss marginalization only in reference to unconditional forms of exponential families. However, there is no loss of generality, since **the effects of observations can always be absorbed by modifying the canonical parameters θ and/or the sufficient statistics ϕ.** *Thus, it means that the parameters θ and/or the family itself depend on the observations. Then, what does it mean to estimate θ since they are partially determined by the observations?*

8.2.3 Mean parameterization and inference problems

Let p be a density wrt to an underlying base measure ν. We do not assume that p belongs to an exponential family. The **mean parameter** μ_α associated to a sufficient statistic $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$ is defined by:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)].$$

Define:

$$\mathcal{M} = \{\mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha, \forall \alpha \in \mathcal{I}\}.$$

The set $\mathcal{M}$ is always convex because $\lambda p + (1-\lambda)p'$ is a distribution (mixture distribution) for any $\lambda \in [0, 1]$.

9 Sensitivity to initial conditions of multidimensional linear stochastic differential equation

9.1 Multidimensional linear SDE

9.2 Geometry of Poincaré half-plane

9.3 Geometry of Siegel half-plane from an algebraic perspective

$\longrightarrow$ Is it true that the symplectic group acts transitively and by isometry on S_n? Is it then possible to describe all geodesics? Write 'geodesics Siegel half plane' in Google Scholar.

9.3.1 Introduction

Let V be a finite-dimensional real vector space.
The dual space to V is $V^T = \hom(V, \mathbb{R})$.
Let W be another real vector space, and let $A \in \hom(V, W)$. Given $v \in V$, the homomorphism A acts from the left on v, meaning that we write Av instead of $A(v)$.
The dual or transpose of A is $A^T \in \hom(W^T, V^T)$ defined by: for all $\alpha \in W^T$, $A^T\alpha = \alpha A$.
Consider the special case where $W = V^T$. Let $B \in \hom(V, V^T)$. The homomorphism B induces a bilinear form $(v, w) \mapsto (Bw)v$.
In this case, $B^T \in \hom((V^T)^T, V^T)$ is called the *adjoint* of B. There is a canonical identification of V with its double dual $(V^T)^T$ via the evaluation map $ev \in \hom(V, (V^T)^T), ev : v \mapsto ev_v$ defined by, for all $\alpha \in V^T$, $ev_v\alpha = \alpha v$ which consists of evaluating α at v. Hence, $B^T \in \hom(V, V^T)$ like B.
We have the following relation: for all $v, w \in V$, $(B^T w)v = (B^T ev_w)v = (ev_w B)v = ev_w(Bv) = (Bv)w$.
$B \in \hom(V, V^T)$ is called *symmetric* or *self-adjoint* if $B^T = B$ and *skew-symmetric* if $B^T = -B$.
$B \in \hom(V, V^T)$ is called positive definite (written $B > 0$) if, for all nonzero $v \in V$, $(Bv)v > 0$.
If $B > 0$, then B is invertible and B^T is also positive definite (because $(B^T v)v = (Bv)v$).
If $G \in \hom(V, V^T)$ is symmetric and positive definite, it is an *Euclidean structure*. G induces a bilinear form g which is an *inner product*.
Given an Euclidean structure G on V, a map $A \in \hom(V)$ is symmetric in the usual sense if GA is symmetric in our sense. The property of being symmetric depends on an inner product, although being diagonalizable does not. If $\Sigma \in \hom(V, V^T)$ is a skew-symmetric isomorphism, it is a *symplectic structure*. It induces a bilinear form σ which is alternating and non-degenerate.

9.3.2 Symplectic and orthogonal groups

Assume that $\dim V = 2n \geq 2$ and fix a symplectic structure Σ on V.
The symplectic group of (V, Σ) is:

$$\mathrm{Sp}(V) = \{A \in \hom(V) : A^T \Sigma A = \Sigma\}.$$

If λ is an eigenvalue of $A \in \mathrm{Sp}(V)$, then $1/\lambda$ is also an eigenvalue of A. (This is because $A^{-1} = \Sigma^{-1} A^T \Sigma$, and A^T has the same eigenvalues on V^T as A has on V, and the eigenvalues of A^{-1} are the inverses of the eigenvalues of A).
If $A \in \mathrm{Sp}(V)$, then $\det(A) = 1$.
Define:

$$\mathcal{C}(V) = \{J \in \hom(V) : J^2 = -I\},$$

where I is the identity.
If $J \in \mathcal{C}(V) \cap \mathrm{Sp}(V)$, then ΣJ is symmetric i.e., $(\Sigma J)^T = \Sigma J$. ΣJ induces a symmetric, non-degenerate bilinear form on V.
Given any G inducing a symmetric, non-degenerate bilinear form on V (but not necessarily positive definite), define the orthogonal group:

$$\mathrm{O}_G(V) = \{A \in \hom(V) : A^T G A = G\}.$$

The special orthogonal group $\mathrm{SO}_G(V)$ is the connected subgroup of $\mathrm{O}_G(V)$ containing the identity. If $G > 0$, $\mathrm{SO}_G(V)$ is simply the subgroup of $\mathrm{O}_G(V)$ with determinant 1.
The Lie algebra of $\mathrm{SO}_G(V)$ is $\mathfrak{so}_G(V) = \{A \in \hom(V) : (GA)^T = -GA\}$ which has dimension $2n^2 - n$. Hence, $\mathrm{O}_G(V)$ also has dimension $2n^2 - n$.
Let $J \in \mathcal{C}(V) \cap \mathrm{Sp}(V)$. Its tangent space is $T_J\mathrm{Sp}(V) = \{A \in \hom(V) : \Sigma JA = (\Sigma JA)^T\}$.
We have the following direct sum: $\hom(V) = T_J\mathrm{Sp}(V) \oplus \mathfrak{so}_G(V)$.

9.4 Geometry of Siegel half-plane from an algebraic perspective (simplified)

9.4.1 Basic definitions and results

Let F be either the real or the complex field.

The **symplectic group** is $\mathrm{Sp}_{2n}F = \{M \in GL_{2n}F : M^T JM = J\}$ with $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$.

Decompose $M \in M_{2n}F$ in four $n \times n$ blocks:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \text{ and } M^T = \begin{bmatrix} A^T & C^T \\ B^T & D^T \end{bmatrix}.$$

M is symplectic if and only if:

$$M^T JM = J,$$
$$\begin{bmatrix} A^TC - (A^TC)^T & A^TD - C^TB \\ -(A^TD - C^TB)^T & B^TD - (B^TD)^T \end{bmatrix} = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}.$$

Hence, M is symplectic if and only if A^TC and B^TD are symmetric and $A^TD - C^TB = I_n$.

Taking the determinant on both sides of the defining equation: $\det(M)^2 det(J) = det(J)$. The matrix J is invertible of inverse $-J$, thus we have $\det(M)^2 = 1$ and $\det(M) = \pm 1$.

Moreover, $M^T = JM^{-1}J^{-1}$, so M^{-1} is similar to M^T. But since M^T is similar to M, then we have that M^{-1} is similar to M.
If M is symplectic, then so is M^T:

$$MJM^T = MJ(JM^{-1}J^{-1}) = -MM^{-1}(-J) = J.$$

The inverse of M is $M^{-1} = J^{-1}M^TJ$ thus:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} -D^T & -B^T \\ C^T & A^T \end{bmatrix}.$$

Notice that if $n = 2$, the relation $A^TD - C^TB = I_n$ becomes $ad - cb = 1$ which is $\det(M) = 1$. Hence, $\mathrm{Sp}_2F = \mathrm{SL}_2F$.

The **Siegel upper half plane** is the set of all complex symmetric matrices with positive definite imaginary part.

$$\mathrm{SH}_n = \{X + iY : X, Y \in \mathrm{Sym}_n\mathbb{R}, Y > 0\}.$$

The action of the symplectic group on the Siegel upper half plane is:

$$\text{for } M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}_{2n}\mathbb{R} \text{ and } Z \in \mathrm{SH}_n, M(Z) = (AZ + B)(CZ + D)^{-1}.$$

It can be shown that $CZ + D$ is invertible and $M(Z) \in \mathrm{SH}_n$ so that the action is well defined[28].

These maps are called **generalized Möbius transformations**. The matrices M and $-M$ have the same action. This action generalize the action of $\mathrm{SL}_2\mathbb{R}$ over the Poincaré upper half space H, since $H = SH_1$ and $\mathrm{Sp}_2\mathbb{R} = \mathrm{SL}_2\mathbb{R}$. It is possible to get a closer connection between the two actions. Define an injective map:

$$H^n \to \mathrm{SH}_n$$
$$(z_1, \ldots, z_n) \mapsto \mathrm{diag}(z_1, \ldots, z_n)$$

[28] Why didn't we check that $M_1(M_2(Z)) = (M_1M_2)(Z)$?

There is a corresponding map for the groups:

$$\phi : (\mathrm{SL}_2\mathbb{R})^n \to \mathrm{Sp}_{2n}\mathbb{R}$$
$$(M_1, \ldots, M_n) \mapsto T(M_1 \oplus \ldots \oplus M_n)T^{-1} := M_1 \odot \ldots \odot M_n$$

where T is the permutation matrix that, when multiplied on the left, sends the odd lines $2j-1$ to j and the even lines $2j$ to $n+j$ (and similarly for the columns when transposed and multplied on the right[29]) for $j = 1, \ldots, n$. The symbol $\bigoplus$ is the matrix direct sum, which is the block diagonal matrix where each block is a term of the sum. If we write $M_j = \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix}$ then:

- the a_j element has indexes $(2j-1, 2j-1)$ in the direct sum, so it is sent to (j, j),
- the d_j element has indexes $(2j, 2j)$, so it is sent to $(n+j, n+j)$,
- the b_j element has indexes $(2j-1, 2j)$, so it is sent to $(j, n+j)$,
- the d_j element has indexes $(2j, 2j-1)$, so it is sent to $(n+j, j)$.

All of this to say that:

$$M_1 \odot \ldots \odot M_n = \begin{bmatrix} a_1 & & & b_1 & & \\ & \ddots & & & \ddots & \\ & & a_n & & & b_n \\ c_1 & & & d_1 & & \\ & \ddots & & & \ddots & \\ & & c_n & & & d_n \end{bmatrix}.$$

Then we have that the action of this matrix can be done componentwise:

$$(M_1 \odot \ldots \odot M_n)(\mathrm{diag}(z_1, \ldots, z_n)) = \mathrm{diag}(M_1(z_1), \ldots, M_n(z_n)).$$

In fact, ϕ is a group isomorphism onto its image.

The action of the symplectic group on the Siegel upper half plane is transitive. Every holomorphic bijective map from SH_n onto itself can be represented as a symplectic map.

9.4.2 Symplectic linear algebra

The matrix J can be used to define a skew-symmetric form in $M_{2n,1}\mathbb{R} \cong \mathbb{R}^{2n}$ as: $(u, v) := u^T Jv$. This form is called the **symplectic form**. Notice that a matrix is symplectic if and only if[30] $(Mu, Mv) = (u, v)$ for all $u, v \in \mathbb{R}^{2n}$.

A symplectic basis of $\mathbb{R}^{2n}$ is a basis such that the matrix of the symplectic form with respect to this basis is J. It is a basis $(e_1, \ldots, e_n, f_1, \ldots, f_n)$ such that $(e_j, f_k) = -(f_k, e_j) = \delta_{jk}$ and $(e_j, e_k) = (f_j, f_k) = 0$. The canonical basis of $\mathbb{R}^{2n}$ is a symplectic basis. Notice that a matrix is symplectic if and only if it is a change of basis matrix from one symplectic basis to another one.

A subspace of $\mathbb{R}^{2n}$ is called Lagrangean if it has dimension n and $(u, v) = 0$ for all u, v in the subspace. Given any symplectic basis $(e_1, \ldots, e_n, f_1, \ldots, f_n)$, we have that $\mathrm{Span}(e_1, \ldots, e_n)$ and $\mathrm{Span}(f_1, \ldots, f_n)$ are Lagrangean spaces.

A symplectic matrix M has determinant 1. This is different from real orthogonal matrix that can have determinant 1 or -1. Now let M be a symplectic matrix. Given that symplectic matrices are similar to their inverse, we can list the eigenvalues of M, with multiplicities, as follows: $(\lambda_1, \ldots \lambda_n, \lambda_1^{-1}, \ldots, \lambda_n^{-1})$. Moreover, we consider this list to be ordered as follows: $|\lambda_1| \geq \ldots \geq |\lambda_n| \geq 1 \geq |\lambda_n^{-1}| \geq \ldots \geq |\lambda_1^{-1}|$. The matrix M (and its induced transformation) is **hyperbolic** if M has no eigenvalue on the complex unit circle i.e., $|\lambda_j| \neq 1$ for all j.

The goal is now to find a normal form for hyperbolic symplectic matrices.

[29] Remember that the inverse of a permutation matrix is its transpose.

[30] This is the same criteria as for isometries but with a skew-symmetric form instead of a symmetric positive definite form.

9.5 Proof that multidimensional Gaussian laws are the Siegel half-plane

See Slepian-Bangs formula.

9.6 Calculus of variations

$$e(s) = \int_0^1 g(\Psi(s,t); \partial_s \Psi(s,t), \partial_s \Psi(s,t)) dt, \tag{25}$$

$\longrightarrow$ *Give an expression for $\Psi(s,t)$: for each s, the curve $t \mapsto \Psi(s,t)$ is a solution of the linear EDS with different initial conditions? Do we have an expression for g? What about a variation between a solution and a geodesic?*

9.7 Bounds

$\longrightarrow$ *Find bounds for $e(s)$ using Grönwall lemma, or using developments. Then, consider the extrema of $e(s)$.*

10 Uncertainty in Machine Learning

10.1 Robust Optimization

Confidence interval. Credibility interval. Credence interval.

10.2 Bayesian Neural Nets (BNN)

10.2.1 Variational Inference (VI)

In [11], Xue et al. propose a Bayesian Transformer Language Model (LM) applied to speech recognition. The model is composed of several Transformer decoders where only the first one is optimized with the VI method. A standard Transformer LM is used to set the mean of the prior distribution. The Bayesian Transformer LM can also be interpolated with the standard Transformer LM. The main results are a small increase in the generalization capability of the LM when using the Bayesian framework by comparison with the standard framework.

Extended Variational Density Propagation (exVDP) In [12], Dera et al. introduce an extended variational density propagation (exVDP) framework for propagating uncertainty in Convolutional Neural Networks (CNN). The weights of the filters of the convolutional layers are modeled as Tensor Normal Distribution. The weights of the fully-connected layers are also modeled as Gaussian vectors. As part of the VI framework, the ELBO is used as the objective function of the network. To estimate the expected log-likelihood (first term of the ELBO), the first and second moments of the weights distributions are propagated through the various layers (convolutional, activation, pooling, fully-connected). Once the posterior distribution is known, the expected log-likelihood is estimated with Monte Carlo sampling, and the full ELBO can be derived for the CNN. The exVDP is compared to other Bayesian networks as well as to classical CNN, on MNIST and CFAR-10 datasets. exVDP is shown to be able to resist Gaussian noise and adversarial attacks far better than other frameworks. Moreover, the propagated variance can be used to quantify the uncertainty of the network.

In [13], Dera et al. apply the exVDP framework to the classification of synthetic aperture radar (SAR) images. These are satellite images than can be used to classify the type of surface in each part of the image (water, crops, buildings etc.). Once again, the exVDP framework is able to resist Gaussian noise and adversarial attack. It is also possible to generate an uncertainty map allowing to visualize the uncertainty of the prediction of the network on each area of the image.

Unscented Variational Density Propagation (unVDP)

10.2.2 Laplace approximation

10.2.3 Stochastic Gradient Markov Chain Monte Carlo

10.2.4 MC-Dropout

10.2.5 Bayes By Backpropagation (BBP)

10.2.6 Preconditioned Stochastic Gradient Langevin Dynamics (p-SGLD)

10.3 Non-Bayesian Approaches

10.3.1 Threshold

10.3.2 Model Ensembling

DeepEnsemble

10.3.3 Prior Network (PN)

10.3.4 Stochastic Differential Equation Network (SDE-Net)

Uncertainty can be split into two components: aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty corresponds to the natural randomness of the task (due to class overlap, data noise etc.). It cannot be reduced with more data. Epistemic uncertainty, on the other hand, is caused by the ignorance of the model due to lack of data. It is high in regions where the data is sparse.

SDE-Net was introduced by Kong et al. [14]. It brings the following benefits:

- Separate aleatoric and epistemic uncertainty.
- No need to specify model prior distributions and infer posterior distributions (as with Bayesian methods).
- Applicable to classification and regression tasks.

DNNs can be viewed as state transformations of a dynamic system. The idea is to add a Brownian motion term to quantify epistemic uncertainty SDE-Net consists of:

- A **drift net** that parameterizes an ODE to fit the predictive function.
- A **diffusion net** that encourages high diffusion for data outside the training distribution.

SDE-Net is the discretization of the following stochastic differential equation:

$$dx_t = f(x_t, t)dt + g(x_t, t)dW_t,$$

where $g(x_t, t)$ is the variance of the Brownian motion and represents the epistemic uncertainty. The aleatoric uncertainty is captured by the drift net by representing model output as a probabilistic distribution (categorical for classification, Gaussian for regression).

The objective function is the following:

$$\begin{aligned}
&\min_{\theta_f} \mathbb{E}_{x_0 \sim P_{train}}[L(x_T)] + \min_{\theta_g} \mathbb{E}_{x_0 \sim P_{train}}[g(x_0; \theta_g)] + \max_{\theta_g} E_{\tilde{x}_0 \sim P_{OOD}}[g(\tilde{x}_0; \theta_g)], \\
&s.t.\ dx_t = f(x_t, t; \theta_f)dt + g(x_0; \theta_g)dW_t,
\end{aligned}$$

where $L(.)$ is the loss function dependent on the task, T is the terminal time of the stochastic process (the last "hidden layer"), P_{train} is the distribution for training data, P_{OOD} is the out-of-distribution data. In order to reduce the complexity: the parameters are shared by each layer, the variance g only depends on x_0 instead of x_t, and to avoid the explosion of solution, the variance g should be bounded by a hyper-parameter σ_{max}.

To quantify the uncertainty, multiple random realization $\{x_T\}_{m=1}^{M}$ are obtained. Then, the aleatoric uncertainty is given by the expected predictive entropy $\mathbb{E}_{p(x_T|x_0,\theta_{f,g})}[\mathcal{H}[p(y|x_T)]]$ for classification, and the

expected predictive variance $\mathbb{E}_{p(x_T|x_0,\theta_{f,g})}[\sigma(x_T)]$ for regression. The epistemic uncertainty is given by the variance of the final solution $\text{Var}(x_T)$. This is similar to ensembling methods but it requires the training of only one model.

For training, there is no closed form solution for x_T. Instead, the authors use the Euler-Maruyama scheme with fixed step size:

$$x_{k+1} = x_k + f(x_k, t; \theta_f)\Delta_t + g(x_0; \theta_g)\sqrt{\Delta t} Z_k, \tag{26}$$

where $Z_k \sim \mathcal{N}(0, 1)$ and $\Delta t = T/N$.

SDE-Net is tested on the following tasks: out-of-distribution detection, misclassification detection, adversarial sample detection, and active learning. It has been shown to have strong performance compared to other uncertainty quantification techniques

10.3.5 Particle optimization

10.3.6 Information Geometry

11 Miscellaneous

11.1 Generalization of the Mean Value Theorem

This section is motivated by a lack of explanations in the correction of exercise 49 question 1 of [15] p.152. We will only show the generalisation at order 2, but the derivation of higher orders is straightforward.

Let f be a C^2 real-valued function on an interval $[a, b]$. Using Taylor's Theorem with Integral Remainder, we get:

$$f(b) - f(a) - f'(a)(b-a) = \int_a^b (b-t)f''(t)dt.$$

Since f'' is continuous on the interval $[a, b]$, it is bounded and the bounds are reached. Moreover, $b - t \geq 0$ for every $t \in [a, b]$, hence we have that:

$$(\min_{[a,b]} f'') \int_a^b (b-t)dt \leq \int_a^b (b-t)f''(t)dt \leq (\max_{[a,b]} f'') \int_a^b (b-t)dt,$$

which can be rewritten as:

$$(\min_{[a,b]} f'')\frac{(b-a)^2}{2} \leq \int_a^b (b-t)f''(t)dt \leq (\max_{[a,b]} f'')\frac{(b-a)^2}{2}.$$

Let us consider the function ϕ defined on $[a, b]$ as $\phi : x \mapsto f''(x)\frac{(b-a)^2}{2}$. Since f'' is continuous on $[a, b]$, ϕ is also continuous on $[a, b]$. Then, using the Intermediate Value Theorem, we have that:

$$\phi([a, b]) = \left[(\min_{[a,b]} f'')\frac{(b-a)^2}{2}, (\max_{[a,b]} f'')\frac{(b-a)^2}{2}\right].$$

Since $\int_a^b (b-t)f''(t)dt \in [(\min_{[a,b]} f'')\frac{(b-a)^2}{2}, (\max_{[a,b]} f'')\frac{(b-a)^2}{2}]$, there exists a $c \in [a, b]$ such that:

$$\phi(c) = \int_a^b (b-t)f''(t)dt.$$

Replacing the integral with the expression of $\phi(c)$ in Taylor's formula, we finally get that there exists a real number $c \in [a, b]$ such that:

$$f(b) - f(a) - f'(a)(b-a) = f''(c)\frac{(b-a)^2}{2}.$$

We can show that in fact $c \in]a, b[$. To do this, we first consider the case where f'' is constant on $[a, b]$. Then, we can choose any value in $]a, b[$. If f'' is not constant, then at least one of the following inequality is true (at least one of them is strict): $(\min_{[a,b]} f'') \int_a^b (b-t)dt \leq \int_a^b (b-t)f''(t)dt \leq (\max_{[a,b]} f'') \int_a^b (b-t)dt$ then ...

11.2 l_p spaces are not Hilbert spaces unless $p = 2$

If $\|\cdot\|_p$ were induced by an inner product $\langle\cdot,\cdot\rangle_p$, it should verify the parallelogram identity:

$$\begin{aligned}\|x+y\|_p^2+\|x-y\|_p^2 &= \langle x+y, x+y\rangle_p + \langle x-y, x-y\rangle_p \\ &= 2\langle x,x\rangle_p + 2\langle y,y\rangle_p + 2\langle x,y\rangle_p - 2\langle x,y\rangle_p \\ &= 2\left(\|x\|_p^2+\|y\|_p^2\right).\end{aligned}$$

Consider the case $x = (1, 0, 0, 0, \dots)$ and $y = (0, 1, 0, 0, \dots)$. Then

$$\|x+y\|_p^2+\|x-y\|_p^2 = 2^{\frac{2}{p}+1},$$

and

$$2\left(\|x\|_p^2+\|y\|_p^2\right) = 4.$$

For the parallelogram inequality to be verified, we must have $2^{\frac{2}{p}+1} = 4 \iff p = 2$. The same counter-example works for $p = \infty$ since

$$\|x+y\|_\infty^2+\|x-y\|_\infty^2 = 2,$$

and

$$2\left(\|x\|_\infty^2+\|y\|_\infty^2\right) = 4.$$

This is very unfortunate because that means that there is no Riemannian metric on $\mathbb{R}^n$ corresponding to the l_p norm unless $p = 2$ (if I am not mistaken, maybe it is possible to have a Riemannian metric even if there is no global inner product). So we cannot use the tools of Riemannian geometry to study the interactions of the normed vector spaces $(\mathbb{R}^n, l_p)$ with a Riemannian manifold $(\mathbb{R}^n, g)$.

11.3 Immersed and embedded submanifolds

Let M be a n-dimensional smooth manifold.

Definition 11.1. A **regular submanifold** S is an open subset of M such that there exists a chart $(U, x^1, \dots, x^n)$ at any point $p \in S$, such that $S \cap U$ is characterized by $x^{k+1} = 0, \dots, x^n = 0$. Hence, $(S \cap U, x^1, \dots, x^k)$ is a coordinate chart for S. We say that S has dimension k and codimension $n - k$.

Remark 11.2.

- It can be shown that a regular submanifold is indeed a smooth manifold.
- The constant 0 in $x^{k+1} = 0, \dots, x^n = 0$ is arbitrary. If $S \cap U$ is characterized by $x^{k+1} = c^{k+1}, \dots, x^n = c^n$ for some constants $c^{k+1}, \dots, c^n$ then S is also a regular submanifold. This is because we could choose another coordinate system defined by $y^1 = x^1, \dots, y^k = x^k, y^{k+1} = x^{k+1} - c^{k+1}, \dots, y^n = x^n - c^n$ and (y^i) will satisfy the definition.

Let $i : N \to M$ be a smooth map between manifolds.

Definition 11.3. The image $i(N) \subset M$ is an

- **Immersed submanifold** of M if i is an injective immersion.
- **Embedded submanifold** of M if i is an homeomorphism onto its image and also an immersion. In this case, i is called an embedding.

The *injective* part is here to avoid self-intersection (in which case $i(N)$ wouldn't be a topological manifold, whatever the topology we endow it).

The *immersion* part is here to preserve the differential structure (i.e., the choice of a maximal differentiable atlas) of N.

Note that an embedded submanifold is necessarily immersed, because an homeomorphism is injective.

The difference between an immersed and an embedded submanifold is the topology we endow $i(N)$. If $i(N)$ is an immersed submanifold, then N can "keep" its own topology. More precisely, $i(N)$ might not even

be a topological manifold if we endow it with the induced topology from M. The only way for $i(N)$ to be a manifold is to endow it with the topology induced by i (i.e., U is open in $N \iff i(U)$ is open in $i(N)$). On the other hand, if $i(N)$ is an embedded submanifold, then $i^{-1} : i(N) \to N$ is continuous. Hence, N is "forced" to have the topology induced by M on $i(N)$.

Let us see some useful counterexamples.

- **A smooth injective map is not necessarily an immersion.** Consider $i^1 : \mathbb{R} \to \mathbb{R}^2$ defined by $i^1(t) = (t^2, t^3)$. It is injective because t^3 is injective. We have $i^1_*(t) = (2t, 3t^2)$. For i^1 to be an immersion, i^1_* must have rank 1 for every t. However, $i^1_*(0) = (0, 0)$ which has rank 0. Hence i^1 is not an immersion.

- **An immersion is not necessarily injective.** Consider $i^2 : \mathbb{R} \to \mathbb{R}^2$ defined by $i^2(t) = (t^2 - 1, t^3 - t)$. We have $i^2_*(t) = (2t, 3t^2 - 1)$ which has rank 1 for every t, so i^2 is an immersion. However, $i^2(-1) = i^2(1) = (0, 0)$ hence i^2 is not injective.

- **An injective immersion is not necessarily an homeomorphism onto its image.** This is the crucial difference between immersed and embedded submanifolds. As far as I know, all counterexamples rely on the idea of the map i getting close to itself at infinity without self-intersecting. Then, there exists a point p in $i(N)$ such that all of its neighborhoods in M contain points of $i(N)$ that are infinitely far from p in $i(N)$. Thus, we can find a neighborhood V of $i^{-1}(p) \in N$ such that, for any neighborhood U of p in M, $i^{-1}(U)$ is *not* included in V. In fact, any neighborhood V that "doesn't go to infinity" works. Hence i^{-1} is not continuous and i is not an homeomorphism.

Proposition 11.4. *If S is a regular submanifold of M, then the inclusion map $i : S \to M$ is an embedding. Hence S is an embedded submanifold.*

Conversely, if $i(N)$ is an embedded submanifold, then there exists a chart $(U, x^1, \dots, x^n)$ at any point $p \in i(N)$ such that $i(N) \cap U = \{q \in M : x^{k+1}(q) = 0, \dots, x^n(q) = 0\}$. Hence $i(N)$ is a regular submanifold.

11.4 Pushforward and Pullback

Let N and M be smooth manifolds of finite dimension n and m respectively.
Let $p \in N$ be a point of N and $q \in M$ a point of M. We denote by T_pN (resp. T_qM) the tangent space of N (resp. M) at p (resp. q). We denote by TN (resp. TM) the tangent bundle of N (resp. M).
Let $F : N \longrightarrow M$ be a smooth map of manifolds. We denote by $F_* : TN \longrightarrow TM$ the differential of F.

Add a figure representing the vector bundles and the map between them

11.4.1 Pushforward of Vectors and Vector Fields

Let $X_p \in T_pN$ be a tangent vector of N at p. Its *pushforward* is defined as $F_{*,p}(X_p) \in T_{F(p)}M$ ([16] p.159).

If X is a vector field of N, we would like to define a pushforward of X as a vector field Y on M such that $Y_{F(p)} = F_{*,p}(X_p)$. However, if F is not injective, then there exists $q \in M$ such that $q = F(p_1) = F(p_2)$ where $p_1 \neq p_2$ are distinct points of N. Then, it is possible that $F_{*,p_1}(X_{p_1}) \neq F_{*,p_2}(X_{p_2})$ such that Y_q is not well defined. If F is not surjective, then there exists $q \in M$ such that there is no $p \in N$ with $F(p) = q$. Once again, Y_q is not well defined. Hence, for $Y = F_*(X)$ to be well defined, a sufficient condition is that F be a diffeomorphism.

If N and M are Lie groups and F is a Lie group homomorphism, then the pushforward of left-invariant vector fields is well defined. This is due to the fact that there is an isomorphism between the space $L(N)$ of left-invariant vector fields on N and the tangent space at the identity T_eN (as well as between $L(M)$ and T_eM). We denote by $\tilde{A}$ the vector field generated by the tangent vector $A \in T_eN$. Then, the pushforward of $\tilde{A}$ is defined as $F_*(\tilde{A}) = (F_{*,e}A)^\sim$. Note that in this case, F is not necessarily a diffeomorphism ([16] p.185).

11.4.2 Pullback of Covectors, Differential Forms, and Functions

Let V and W be two vector spaces and $L : V \longrightarrow W$ a linear map. Let k be a non-negative integer and $A_k(V)$ be the space of k-covectors on V. The *pullback* map $L^* : A_k(W) \longrightarrow A_k(V)$ is defined for all $f \in A_k(W)$

and for all $(v_1, \ldots, v_k) \in V^k$ by:

$$(L^* f)(v_1, \ldots, v_k) = f(L(v_1), \ldots, L(v_k))$$

The pullback can be seen as a composition $L^* f : V \times \ldots \times V \xrightarrow{L \times \ldots \times L} W \times \ldots \times W \xrightarrow{f} \mathbb{R}$ where we originally had a covector f defined on $W \times \ldots \times W$, which is then *pulled back* by L to obtain a new covector $L^* f$ defined on $V \times \ldots \times V$.

The operator A_k such that $A_k : V \mapsto A_k(V)$ and $A_k : L \mapsto L^*$ is a contravariant functor from the category of vector spaces and linear maps to itself. If $k = 1$, then $A_1(V) = V^\vee$ is the dual space of V and L^* is the dual map of L. This general definition of pullback is used to define the pullback of differential forms and functions ([16] p.113).

Let $\omega : M \longrightarrow \bigwedge^k (T^* M)$ be a differential k-form (i.e., a k-covector field) where $\bigwedge^k (T^* M) \approx \bigcup_{p \in M} A_k(T_p M)$. The pullback of ω is defined pointwise as the pullback of k-covectors. For all $p \in M$ and for all $(v_1, \ldots, v_k) \in (T_p N)^k$ we write ([16] p.204):

$$(F^* \omega)_p (v_1, \ldots, v_k) = \omega_{F(p)} (F_{*,p}(v_1), \ldots, F_{*,p}(v_k))$$

The pullback can be seen as a composition $T_p N \times \ldots \times T_p N \xrightarrow{F_{*,p} \times \ldots \times F_{*,p}} T_{F(p)} M \times \ldots \times T_{F(p)} M \xrightarrow{\omega_{F(p)}} \mathbb{R}$.

If $k = 1$, we have from the previous paragraph that $F_{*,p}$ pushes forward vectors. Define the codifferential $(F_{*,p})^\vee : T^*_{F(p)} M \longrightarrow T^*_p N$ to be the dual map of $F_{*,p}$. The codifferential $(F_{*,p})^\vee$ pulls back a covector at $F(p)$ from M to N. Hence, we have $(F^*)_p = (F_{*,p})^\vee$. More precisely, if $\omega_{F(p)} \in T^*_{F(p)} M$ and $X_p \in T_p N$, then we have ([16] p.196):

$$(F^* \omega)_p (X_p) = F^* (\omega_{F(p)})(X_p) = ((F_{*,p})^\vee \omega_{F(p)})(X_p) = \omega_{F(p)} (F_{*,p} X_p)$$

Similarly, one can write the pullback as a composition $T_p N \xrightarrow{F_{*,p}} T_{F(p)} M \xrightarrow{\omega_{F(p)}} \mathbb{R}$.

If $k = 0$, $A_0(T_p M) = \mathbb{R}$ such that the 0-forms on M are the C^∞ functions on M. Let $h : M \longrightarrow \mathbb{R}$ be such a function. Applying the definition of the pullback of k-forms, we could write $(F^* h)_p (\varnothing) = h_{F(p)}(\varnothing) = (h \circ F)(p)$. Hence, the pullback of h by F is defined by:

$$F^* h = h \circ F$$

Here, the pullback is simply the composition $N \xrightarrow{F} M \xrightarrow{h} \mathbb{R}$.

Unlike vector fields which in general cannot be pushed forward under a smooth map, every k-form (including covector fields) can be pulled back by a smooth map. This difference in behavior of vector fields and forms under a map can be traced back to a basic asymmetry in the concept of a function: *every point in the domain maps to only one image point in the range, but a point in the range can have several preimage points in the domain.*

Remark 1: Functions are many-to-one relations. What about the other possible type of relations? One-to-many relations correspond to the preimage of a point by a function, so the concept is already encapsulated in the concept of function. Many-to-many relations can rewritten as functions between collections of sets (or is it?). Hence, functions are truly general.

Remark 2: What about the pushforward of k-forms or the pullback of vector fields? At a point q, a k-form on M is simply a k-covector of $A_k(T_q M)$. A k-covector takes as *inputs* vectors of $T_q M$. So, all we can do is to pull back these vectors by seeing them as outputs of F_*. Hence, the pushforward of a k-form simply does not make any sense. From a duality perspective, we observe that a vector field *outputs* a vector of $T_p N$ at a given point p. It is then natural to push forward this vector using F_*. But recall that if F is not bijective, $Y = F_*(X)$ may not be well defined. Once again, the pullback of a vector field does not make sense (or does it?).

11.5 The Uncertainty Principle

11.5.1 Gabor's Uncertainty Principle

In this paragraph, we provide an intuitive explanation of Gabor's Uncertainty Principle before presenting the Paley-Wiener theorem.

11.6 The curvature of a curve in the Euclidean space $\mathbb{R}^3$

Let $r : I \to U$ be a smooth curve from an interval $I \subset \mathbb{R}$ to an open set U of a Riemannian manifold M.

The geodesic curvature defined by:

$$\kappa(t) = \frac{\sqrt{|\dot{r}(t)|^2 |D_t \dot{r}(t)|^2 - \langle D_t \dot{r}(t), \dot{r}(t) \rangle^2}}{|\dot{r}(t)|^3}, \tag{27}$$

reduces to:

$$\kappa(t) = \frac{|\dot{r}(t) \times \ddot{r}(t)|}{|\dot{r}(t)|^3}, \tag{28}$$

when $M = \mathbb{R}^3$ with the Euclidean metric i.e., the inner product matrix is the identity at every point:

$$\langle u, v \rangle = \sum_{i,j=1}^{n} \delta_{ij} u^i u^j = u^T v,$$
$$|u|^2 = \langle u, u \rangle.$$

So, we assume that $M = \mathbb{R}^3$. Let's use the canonical basis (e_1, e_2, e_3) of $\mathbb{R}^3$ as a global coordinate system. We identify the tangent space at each point with $\mathbb{R}^3$ itself. In coordinates, the covariant derivative of $\dot{r}$ along r is defined by:

$$D_t \dot{r}(t) = \sum_{i,j,k=1}^{n} (\ddot{r}^k(t) + \dot{r}^i(t) \dot{r}^j(t) \Gamma_{ij}^k(r(t))) e_k, \tag{29}$$

where Γ_{ij}^k are the Christoffel symbols of the second kind:

$$\Gamma_{ij}^k = \frac{1}{2} g^{kl} (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}), \tag{30}$$

with g_{ij} the metric and g^{kl} its inverse. Since the metric is $g_{ij} = \delta_{ij}$, the Christoffel symbols Γ_{ij}^k are all identically zero. So the covariant derivative reduces to the vector of second derivatives :

$$D_t \dot{r}(t) = \sum_{k=1}^{n} \ddot{r}^k(t) e_k = \ddot{r}(t). \tag{31}$$

Now, coming back at equation 27, we have:

$$\begin{aligned}
\kappa(t) &= \frac{\sqrt{|\dot{r}(t)|^2 |D_t \dot{r}(t)|^2 - \langle D_t \dot{r}(t), \dot{r}(t) \rangle^2}}{|\dot{r}(t)|^3}, \\
&= \frac{\sqrt{|\dot{r}(t)|^2 |\ddot{r}(t)|^2 - \langle \ddot{r}(t), \dot{r}(t) \rangle^2}}{|\dot{r}(t)|^3} \\
&= \frac{|\dot{r}(t) \times \ddot{r}(t)|}{|\dot{r}(t)|^3} \text{ (using Lagrange's identity } |u \times v|^2 = |u|^2 |v|^2 - \langle u, v \rangle^2),
\end{aligned}$$

which is precisely equation 28.

11.7 Differential Geometry in Deep Learning

As far as I know, there are three very different settings for applying differential geometry to deep learning:

- For unsupervised learning and dimensionality reduction. In this setting, there are applications of algebraic topology, category theory, and information geometry.
- To exploit specific structures in the data (images, time series, graphs etc.), for example by designing architectures that rely on symmetries. In this setting, there are applications of geometry and its links with algebra (see "Algebraic Differential Geometry"), and information geometry.
- To study robustness and learning in supervised learning (and eventually reinforcement learning) from a geometric-statistical perspective. This is specific to information geometry.

11.8 Finally the Truth about Fisher Information?

In this subsection, I want to answer the following question: *why is the Fisher Information the right metric for parameterized families of probability distributions?* Here are some leads:

- In this tweet[31], Gabriel Peyré says: "The Fisher metric is the unique Riemannian structure on parametric families of densities which is invariant/covariant under reparametrization of the samples/parameter spaces. For 1D Gaussians, it corresponds to the Poincaré hyperbolic half plane."
- In a comment of the same tweet, he says: "Presenting Fisher through the prism of KL (or any divergence) is a bit misleading. One should really think about a probability density as being the square of a vector on the unit sphere. Then the natural distance is Hellinger[32]. Fisher is just its restriction to a parametric sub-manifold. Then all invariances are not magic anymore."
- In this blog post[33], Tom Leinster explains that the infinitesimal metric of a large family of entropies (which are deformations of the relative entropy, including the Rényi entropies and q-logarithmic entropies) is the Fisher metric (up to a constant factor). The surprising result is that the infinitesimal metric is always the Fisher metric, whatever the entropy you choose in this family.
- Some other properties of the FIM are explained in the wikipedia page[34]. The most well known property is *Chentsov's theorem* which states that the FIM is the unique Riemannian metric on a statistical manifold that is invariant under sufficient statistics (up to rescaling).

But what is the meaning of all this??

11.9 Is Machine Learning a Pseudoscience?

In this video[35], Lê Nguyen Hoang argues that Machine Learning is like neoclassical economics: it relies on assumptions that are not verified in the real world.

In particular, machine learning assumes that the data (x, y) are i.i.d. samples of a "true" but unknown distribution, and thus it makes sense to define a generalization error that learning models should minimize.

Another assumption is the "double descent". It relies on empirical evidences to say that when a model has a large number of parameters compared to the number of training points, it somehow escapes the bias-variance trade-off and achieves low error.

11.9.1 Programme de recherche en IA et critiques de Lê.

Le positionnement de mes recherches en IA par rapport à celles de Lê Nguyen Hoang (et al.). Lê s'est focalisé sur les attaques de type *poisoning*. Il s'agit d'une vision centrée sur les données (plutôt que sur les modèles) et qui donne une grande importance aux *sources* des données. Lê s'interesse donc aux IA qui sont entraînées avec des données issues de multiples sources, typiquement collectées auprès d'utilisateurs (par exemple les utilisateurs de plateformes en ligne). Lê distingue deux types de problèmes:

- **Fairness**. L'apprentissage va favoriser les sources les plus actives. Cela pose un problème *démocratique* puisque le principe "une personne, une voix" n'est pas respecté.
- **Security**. Des acteurs malveillants (typiquement des organismes de propagande issus de "régimes autoritaires") peuvent injecter des données (poisoning) afin de modifier l'apprentissage. On parle alors *d'attaques Byzantines*. Plus généralement, cela pose la question du *système de vote*: comment s'assurer

[31] `https://twitter.com/gabrielpeyre/status/1033957406714286081?lang=fr`

[32] `https://fr.wikipedia.org/wiki/Distance_de_Hellinger`

[33] `https://golem.ph.utexas.edu/category/2018/05/the_fisher_metric_will_not_be.html`. See also `https://golem.ph.utexas.edu/cgi-bin/MT-3.0/mt-search.cgi?IncludeBlogs=3&Template=category&search=fisher`.

[34] `https://en.wikipedia.org/wiki/Fisher_information_metric`

[35] `https://youtu.be/IVqXKP91L4E`

qu'un acteur rationnel aura intérêt à communiquer ses véritables préférences (c'est le problème[36] du *vote utile*).

Pour résoudre le premier problème, Lê préconise un modèle "user-centric". Au lieu de voir les données comme un "pot commun", il faut pouvoir distinguer les sources, afin de s'assurer que chaque source aura un impact limité sur l'apprentissage. Pour résoudre le second problème, Lê préconise une "normalisation". Par exemple, utiliser la médiane géométrique plutôt que la moyenne. De manière équivalente, cela consiste à dire que chaque point tire avec une force unitaire, plutôt qu'une force proportionnelle à sa distance au centroïde (qui inciterait à exagérer sa position afin d'avoir une force plus grande). Néanmoins, la médiane peut subir le "vote utile" dans certains cas particulier. Aussi, la plateforme Tournesol utilise une autre "normalisation" basée sur la norme l_∞ (je ne sais pas ce que ça veut dire). Lê n'a pas encore expliqué cette normalisation (elle est décrite dans le papier Robust Sparse Voting que je n'ai pas lu).
Point intéressant. Lê critique plusieurs hypothèses classiques du machine learning (sans pour autant donner d'alternatives):

- L'hypothèse iid, qui peut conduire à des résultats dangereux (selon Lê). Il n'explique pas pourquoi et je n'ai pas assez de connaissances pour deviner moi-même. Il faudrait que je comprenne mieux les processus stochastiques. Mais intuitivement, on comprend que si la distribution change (distribution shift) ou que les données sont dépendantes, alors l'apprentissage peut être faussé.

- L'hypothèse d'une distribution sous-jacente dont serait issue les données. Je comprends qu'une telle distribution n'existe pas vraiment dans la réalité. Néanmoins, je ne comprends pas comment on peut démontrer quoi que ce soit si on retire cette hypothèse. Que mettre à la place ? D'où viennent les données ?

- L'hypothèse que le but du machine learning est de généraliser à des données non-observées. Autant j'aurais pu prévoir les deux critiques précédentes, autant celle-là me laisse sur le cul. Lê n'explique rien, mais je suppose qu'il parle de la minimisation du risque espéré $\mathbb{E}_{p(x,y)}[\mathcal{L}(y, f(x)]$. Selon lui, cela favorise le "statu-quo". Il enchaîne en disant qu'il n'est pas un "idéologue". Pourtant, s'opposer au statu-quo est une position idéologique (que je partage). Mais oublions la question politique pour le moment. Par quoi remplacer le risque espéré ? Surprenamment, je m'étais interrogé il y a quelques semaines sur la pertinence du risque espéré pour la robustesse aux "attaques adversariales". En effet, on peut écrire $p(x, y) = p(y|x)q(x)$. Le risque espéré donne un poids négligeable aux erreurs commises dans les régions où $q(x) \approx 0$. Je conjecturais que les attaques adversariales sont situées dans de telles régions, et que l'adversarial training augmente le $q(x)$ de certaines régions, *mais pas de toutes, loin s'en faut*. Dans ce sens, le risque espéré favorise le statu-quo en donnant presque tout son poids aux régions les plus probables. Impossible de savoir si c'est ce que Lê voulait dire. Puisque Lê rejette l'utilisation de la distribution $p(x, y)$, je ne sais même pas si le risque espéré fait du sens pour lui. Puisqu'il est difficile de combler les régions $q(x) \approx 0$ (par définition quasi-absentes des données d'entraînement), je pensais que le seul moyen de "généraliser correctement" (c'est-à-dire sans utiliser le risque espéré) était d'utiliser un prior "correct" qui reste à déterminer.

Pour Lê, le risque existentiel des IA est déjà là: il s'agit des systèmes de recommandation actuels qui ne sont pas "robustement bénéfiques". Les questions d'alignement et de long-termisme semblent le désintéresser (et il a bien raison).
Pour ma part, je m'intéresse plutôt aux modèles, et surtout à créer des modèles "réellement intelligents". Je suppose donc que les données sont "parfaites" (à voir ce que cela veut dire). En particulier, il n'y a qu'une seule source qui est supposée fiable. On veut construire un modèle qui "comprend vraiment la tâche" et ne sera donc pas sensible à des "perturbations non pertinentes". Même si je ne m'y suis jamais intéressé, cela pourrait être lié à l'apprentissage avec peu de données. Comment généraliser correctement avec peu de données? Qu'est-ce que "généraliser correctement"? Mes recherches sont donc plutôt orthogonales à celles

[36] Le second problème des systèmes de vote est la *sensibilité aux alternatives non pertinentes*: l'ajout d'une nouvelle alternative modifie les préférences de la population (telles qu'agrégées par le système de vote) des alternatives déjà présentes (i.e., l'ordre des alternatives déjà présentes est changé). Exemple: dans une élection Mélenchon-Le Pen, la population préfère Mélenchon à Le Pen. Si on ajoute Macron, les gens préfèrent maintenant Le Pen à Mélenchon car certains votes Mélenchon se sont reportés sur Macron. C'est un problème propre au scrutin uninominal.

de Lê. Lê m'invite cependant à reconsidérer ce que sont des "données", d'où elles viennent, qu'est-ce qu'une "tâche"?

11.10 The Likelihood principle.

Il s'agit d'un principe fondamental des statistiques, mais qui ne fait pas du tout l'unanimité.
Pour des données x fixées, la fonction de vraisemblance est $\theta \mapsto p(x;\theta)$.
Le principe de vraisemblance affirme que *toute l'information sur un paramètre à inférer est contenue dans la fonction de vraisemblance.* En particulier, le paramètre inféré ne peut pas dépendre d'évènements qui ne se sont pas produits.

Voici un exemple classique. Imaginons qu'Alice demande à Bob de tester si une pièce de monnaie est équilibrée ou pas. Alice demande à Bob de faire 12 lancers. Bob obtient 3 piles. En particulier, le 12e lancer est un pile. A ce moment, Bob découvre qu'il avait mal compris les instructions d'Alice. Alice voulait en fait que Bob continue ces lancers jusqu'à obtenir trois piles. Bob est content parce que ça correspond parfaitement à ce qu'il a fait. Mais Charlie lui fait alors remarquer que ça change le résultat de son expérience.

En effet, la pièce suit une loi de Bernoulli de paramètre inconnue p (probabilité d'obtenir pile). Le nombre de piles suit une loi binomiale de paramètres $n = 10$ et p inconnue. On souhaite tester l'hypothèse nulle $H_0 : p = 0.5$ versus $H_1 : p < 0.5$. On va calculer la p-value. On suppose que le test est significatif (i.e., on rejette H_0) si la p-value est inférieure à 0.05.

- Dans le premier cas de figure (12 lancers avec 3 piles), la probabilité d'obtenir ce résultat ou pire sous H_0, c'est-à-dire trois piles ou moins en supposant la pièce équilibrée, est:

$$\mathbb{P}_0 = \sum_{i=0}^{3} \binom{n}{i} p^i (1-p)^{n-i} = \left(\frac{1}{2}\right)^{12} \sum_{i=0}^{3} \binom{12}{i} = 0.073.$$

  Dans ce cas, le test n'est pas significatif et H_0 n'est pas rejetée.

- Dans le second cas de figure, on s'arrête au 3e pile. On a donc 11 lancers avec 2 piles puis un lancer qui est un pile. Ici, on va calculer la probabilité d'avoir 2 piles ou moins parmi 11 lancers suivi par un pile (en supposant que la pièce est équilibrée):

$$\mathbb{P}_1 = \left(\sum_{i=0}^{2} \binom{n-1}{i} p^i (1-p)^{n-1-i} \right) p = \left(\frac{1}{2}\right)^{12} \sum_{i=0}^{2} \binom{11}{i} = 0.0164.$$

  Dans ce cas, le test est significatif et H_0 est rejetée.

Pourtant, la vraisemblance est $p \mapsto p^3(1-p)^9$ dans les deux cas, donc si le principe de vraisemblance était respecté, on devrait obtenir le même résultat. Certaines personnes considèrent que cet exemple prouve que le principe de vraisemblance est faux. D'autres considèrent qu'il prouve que ce genre de test (et en particulier l'usage de la p-value) est erroné.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley InterScience, 2006.

[2] J. M. Lee, *Introduction to Smooth Manifolds.* Springer, second edition ed., 2013.

[3] P. Molino, *Riemannian Foliations.* Birkhäuser, 1988.

[4] S. Lavau, "A short guide through integration theorems of generalized distributions," *Differential Geometry and its Applications*, vol. 61, pp. 42–58, Dec. 2018.

[5] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry.* American Mathematical Society, 2000.

[6] J. M. Lee, *Riemannian Manifolds: An Introduction to Curvature.* No. 176 in Graduate Texts in Mathematics, Springer, 1997.

[7] L. W. Tu, *Differential Geometry*, vol. 275 of *Graduate Texts in Mathematics*. Springer International Publishing, 2017.

[8] S.-i. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*. New York, NY: Springer New York, 1985.

[9] S.-i. Amari, *Information Geometry and Its Applications*, vol. 194 of *Applied Mathematical Sciences*. Springer Japan, 2016.

[10] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, 2007.

[11] B. Xue, J. Yu, J. Xu, S. Liu, S. Hu, Z. Ye, M. Geng, X. Liu, and H. Meng, "Bayesian transformer language models for speech recognition," in *ICASSP*, pp. 3–8, 2021.

[12] D. Dera, G. Rasool, and N. C. Bouaynaya, "Extended variational inference for propagating uncertainty in convolutional neural networks," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[13] D. Dera, G. Rasool, N. C. Bouaynaya, A. Eichen, S. Shanko, J. Cammerata, and S. Arnold, "Bayes-sar net: Robust sar image classification with uncertainty estimation using bayesian convolutional neural network," *2020 IEEE International Radar Conference*, pp. 362–367, 2020.

[14] L. Kong, J. Sun, and C. Zhang, "SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates," *arXiv:2008.10546 [cs, stat]*, Aug. 2020.

[15] F. Rouvière, *Petit Guide de Calcul Différentiel à l'usage de la licence et de l'agrégation, quatrième édition*. Cassini, 2014.

[16] L. W. Tu, *An Introduction to Manifolds, Second Edition*. Springer, 2010.