Lecture 1, 1/22/20

Example: The set of all probability distributions form a manifold. This is called the statistical manifold.

- Each element in the manifold is a probability distribution; a function.
- This manifold is infinite dimensional.
- The set of Gaussian distributions in dimension k is a submanifold.
- Consider the set of all $N(\mu, \sigma)$ with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

Recall a probability distribution is a function satisfying

1. $p(x) \geq 0$
2. $\displaystyle\int p(x)dx = 1$

Say $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ are determined by $\mu_1, \mu_2, \sigma_1, \sigma_2$. It's a parameterization. The correct way to construct a metric is to view $(\mu, \sigma) \in P_+$ where P_+ is the Poincare upperplane. To find the distance between the points (μ_1, σ_1) and (μ_2, σ_2) in the $\mu\sigma$-plane, you must find the geodesic.

- "Connections on a manifold is just a derivative" just like one calls the derivative of $f : \mathbb{R}^n \to \mathbb{R}$ as the gradient

Lecture 2, 1/27/20

Abstract Manifolds and Examples.

Example. When we observe a sequence of points $\{(x^(i), y^(i))\}$ in the shape of a line, we know that there is some kind of correlation within the data. However, this is cheating, since we know what it looks like, but the computer does not see it. A way we handle this is by making the computer find the line of best fit.

The computer first locates a centroid within the data, and then rotates about that vertex. But how does it find the direction to point? One way to think about this is to imagine a normal vector from the line rotating around a circle S^1, as it tries to find the best normal vector.

The problem is well defined since S_1 is compact; hence, a desirable optimization exists. Also note that $\boldsymbol{n}$ and $-\boldsymbol{n}$ both determine the same line. Thus the set of lines passing through the centroid are in one to one correspondence with $S_1/\mathbb{Z}_2$. However, this is topologically a circle still; it's like we only focus on the upper half, and then identify the vertices of our semicircle together.

Example. Suppose we have a bunch of data points in the shape of a plane. Again, identifying the fact that we approximately have a plane is easy for us, although we'd like to teach a computer to understand the shape present.

We assign weights:

$$y = w_0x_0 + w_1x_1 + w_2x_2$$

and find w_0, w_1, w_2 so the plane fits the data but is not overfitting nor underfitting.

Again, in this situation, we look for a centroid of the data, say C. We then parse through all possible planes, rotating around C, seeking the optimal plane.

Note: Mathematically, we want to make the configuration space for the problem as simple as possible. A way to do this is to look at the normal plane assigned at C; hence we can just look at one point to control the plane. So our problem boils down to observing that points on S^2 correspond to all of our possible planes; with the exception that $\boldsymbol{n}$ and $-\boldsymbol{n}$ correspond to the same plane. Hence we are really looking at $S_2/\mathbb{Z}_2$.

Definition 0.1. A **real projective plane** $\mathbb{R}P^2$ is defined to be $S^2/\mathbb{Z}_2$. That is, a real projective plane is a set of all 2-planes in $\mathbb{R}^3$ passing through the origin.

Example. How do you count the set of lines? We presented a method before on how to do it, but there is another way. We can count the lines by measuring their intersections with a vertical line centered at $x = 1$. This then results in counting $\mathbb{R}^1 \cup \{\infty\} \cong S_1$.

Another way to do this is to count by angles; $0 \leq \theta \leq \pi$ which also corresponds to S_1.

In either case, we obtain the **projective real line**.

Definition 0.2. A **differentiable manifold** of dimension n is a set M and a family of injective mappings $x_\alpha : U_\alpha \subset \mathbb{R}^n \to M$ of open sets U_α of $\mathbb{R}^n$ into M such that

1. $\bigcup\limits_{\alpha} x_\alpha(U_\alpha) = M$

2. For any pair α, β with $x_\alpha(U_\alpha) \cap x_\beta(U_\beta) = W \neq \varnothing$ the set $x_\alpha^{-1}(W)$ and $x_\beta^{-1}(W)$ are open sets in $\mathbb{R}^n$ and the mapping $x_\beta^{-1} \circ x_\alpha$ is differentiable.

Lecture 3, 1/29/20

Definition 0.3. Let x be a random variable (which may be discrete, scalar or vector continuous). A statistical model $M = \{p(x, \zeta)\}$ with parameter ζ is a manifold when it satisfies the desired regularity conditions.

Information geometry studies the invariante geometrical structures of regular statistical models.

For the Gaussian random variable x, we have that

$$p(x : u, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

The set of Gaussian distributions is a two-dimensional manifold, where a point denotes a probability desity function and

$$\zeta = (\mu, \sigma).$$

Alternatively, we can let m_1, m_2 be the **first and second moments** of x, given by

$$m_1 = E[x] = \mu \qquad m_2 = E[x^2] = \mu^2 + \sigma^2.$$

Then $\zeta = (m_1, m_2)$ is a coordinate system. We could even have the coordiate system

$$\zeta = (\theta_1, \theta_2) \qquad \theta = \frac{\mu}{\sigma^2}, \theta_2 = -\frac{1}{2\sigma^2}.$$

called the **natural coordinates**.

This is similar to the idea where for a smooth function $f(x)$ we have

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots$$

and with a change of coordinates

$$f(x) - f(x_0) = f'(x_0)\overline{x} + \frac{f''(x_0)}{2!}\overline{x}^2 + \cdots$$

Now consider a discrete random variable with takes values on $X = \{0, 1, 2, \ldots, n\}$. Then if $p_i = \text{Prob}x = i > 0$, we can compute a probability vector

$$\boldsymbol{p} = (p_0, p_1, \ldots, p_n).$$

Since $\sum_{i=0}^{n} p_i = 1$, one of the vectors is linearly dependent, so we see that $\boldsymbol{p}$ is n-dimensional. Therefore, the set of probability distributions of a discrete form the **manifold of discrete**

distributions. In fact, these manifolds actually take the form of simplices S_n. The manifold becomes a n-dimensional simplex.

Consider the notation

$$\delta_i(x) = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{otherwise} \end{cases}.$$

Then we have that

$$p(x, \boldsymbol{\zeta}) = \sum_{i=1}^{n} \zeta_i \delta_i(x) + p_0(\boldsymbol{\zeta})\delta_0(x)$$

and remember that

$$p_0 = 1 - \sum_{i=1}^{n} p_i.$$

In information geometry, we use another coordinate system $\boldsymbol{\theta}$ where

$$\theta_i = \log\left(\frac{p_i}{p_0}\right) \quad i = 1, 2, \ldots, n.$$

That is, these are the best coordinates for S_n.

Let x be a variable taking value in $N = \{1, 2, \ldots, n\}$. Then we assigned a positive measure m_i to each element $i \in N$ so that

$$\zeta = (m_1, \ldots, m_n).$$

This defines a distribution of meaesures over N. The set of all positive measures forms an n-dimensional manifold. When we have $\sum n = 1$, this proabbility distributiion corresponds with the S_{n-1} simplex.

An example of this occurs when we describe the brigthness of an image by discretizing it into n^2 pixels. For each coordinate (i, j), we assign a brigthness to it, and this forms a positive measure.

Another examplee occurs with matrices. The set of all $n \times n$ matrices form an n^2-dimensional manifold. If A is symmetric and positive definite, then they form a $\frac{n(n+1)}{2}$-dimensional manifold, since we can use the elements in the upper diagonal of A as a coordinate system.

Another example includes the set of all neural networks which form a manifold. A neural network is specified by weights w_{ij} which connects neuron i to neuron j. The coordinate system for this manifold is the matrix $\boldsymbol{W} = (w_{ij})$.

Lecture 4, 2/3/19

Recall that

Definition 0.4. A **differentiable manifold** of dimension n is a set M and a family of injective mappings $x_\alpha : U_\alpha \subset \mathbb{R}^n \to M$ of open sets U_α of $\mathbb{R}^n$ into M such that

1. $\bigcup_{\alpha} x_\alpha(U_\alpha) = M$

2. For any pair α, β with $x_\alpha(U_\alpha) \cap x_\beta(U_\beta) = W \neq \varnothing$ the set $x_\alpha^{-1}(W)$ and $x_\beta^{-1}(W)$ are open sets in $\mathbb{R}^n$ and the mapping $x_\beta^{-1} \circ x_\alpha$ is differentiable.

3. The family $\{(U_\alpha, x_\alpha)\}$ is maximal relative to the coordinates (1) and (2).

- The pair (U_α, x_α) with $p \in x_\alpha(U_\alpha)$ is called a **parameterization** (or a system of coordiantes) of M at p.

- $x_\alpha(U_\alpha)$ is called a **coordiante neigborhood**.

- The family $\{(U_\alpha, x_\alpha)\}$ satisfying (1) and (2) is called a **differential** structure on M.

The reader may realize the above criteria is probably inconvenient to use to check if something is a manifold. The process isn't too bad for simple, obvious shapes but can be gradually more complicated. In real life, people use tricks to see if a set is a manfiold.

Example. Consider the set $X = \{(x, y, z) = x^2 + y^2 + z^2 = 1\}$. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ where

$$f(x, y, z) = x^2 + y^2 + z^2.$$

We can find a critical point of this by observing that

$$\nabla f = 0 \implies (2x, 2y, 2z) = (0, 0, 0) \implies (x, y, z) = (0, 0, 0).$$

Therefore $f(0, 0, 0) = 0$ is a critical value. So 1 is a regular value, which implies that $f^{-1}(1) = \{x^2 + y^2 + z^2 = 1\} = X$ is a manifold in $\mathbb{R}^3$.

Example. Consider the torus $T = S^1 \times S^1$. Both S^1 are manifolds, and the product of manifolds is a manifold, so that the torus is also a manifold.

We can summarize some tricks:

- View a given set as an image of a regular value of a differenitable map.

- Form a product of a given manifold to get another manifold.

- Form a **tangent bundle** TM of a manifold M to get another manifold. The tangent bundle can often be used to understand the dynamics of the manifold.

- By the method of "discontinuous action of a group."

- View a set as a homogeneous space. What does this mean? Let G be a lie group and suppose G acts on a manifold M such that, for all $x, y \in M$, there exists $g \in G$ such that
$$gx = y.$$
Find a subgroup H which fixes an element M. The set G/H id called a manifold called a **homogenous space**.

Example. Recall that $\mathbb{R}P^2 = S^2/\mathbb{Z}_2$ where we write $\mathbb{Z} = \{I, A\}$. Here, I is the identity and A is the antipodal action. For any $p \in S^2$, we write

$$I(p) = p \qquad A(p) = -p$$

where $-p$ can be thought of as a 180 degree rotation. We then say that $\mathbb{R}P^2$ is a manifold.

Example. Let $M = S^{n-1}$, and suppose $G = SO(n)$ acts on the manifold S^{n-1}. Let $v \in M$; then there exists a subgroup $H = SO(n-1)$ which fix v. For example, consider the subgroup $SO(2)$ where

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Then we have that $S^n = SO(n)/SO(n-1)$. This can also be imagined as a fiber bundle.

$$\begin{tikzcd} SO(3) \arrow[r] & SO(3) \arrow[d] \\ & S^2 \end{tikzcd}$$

Definition 0.5. The set of all lines passing through the origin of $\mathbb{R}^{n+1}$ is called the **real projective space** of dimension n, denoted at $\mathbb{R}P^n$.

Let $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$. Then $\mathbb{R}P^n$ can be identified as the quotient sapce $\mathbb{R}^{n+1} - \{0\}$ with the equivalence relation $(x_1, \dots, x_{n+1}) \cong (\lambda x_1, \dots, \lambda x_{n+1})$. The points of $\mathbb{R}P^n$ will be denoted by the equivalence class $[x_1, x_2, \dots, x_{n+1}]$. Notice that

$$[x_1, \dots, x_{n+1}] = [\frac{x_1}{x_i}, \dots, 1, \dots, \frac{x_{n+1}}{x_i}].$$

Define the subsets $V_1, V_2, \ldots, V_{n+1}$ of $\mathbb{R}P^n$ by

$$V_i = \{[x_1, x_2, \ldots, x_{n+1}] \mid x_i \neq 0\} \quad i = 1, 2, \ldots, n.$$

The claim to prove now is that $\mathbb{R}P^n$ can be covered by the sets $V_1, \ldots, V_n$. This is obtained by the maps

$$x_i(y_1, \ldots, y_n) = [y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_n]$$

where $y_j = \frac{x_j}{x_i}$.

One then has to check that $V_i \cap V_j \neq 0$ and that $x^{-1}(V_i \cap V_j)$ is open. One then must check that

$$\begin{aligned} x_j^{-1} \circ x_i(y_1, \ldots, y_n) &= x_j^{-1}([y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_n]) \\ = x_j^{-1}(\frac{y_1}{y_j}, \ldots, &\frac{y_{j-1}}{y_j}, 1, \frac{y_{j+1}}{y_j}, \ldots, \frac{y_{i-1}}{y_j}, \frac{1}{y_j}, \frac{y_{i+1}}{y_j}, \ldots, \frac{y_n}{y_j}) \\ = \frac{y_1}{y_j}, \ldots, &\frac{y_{j-1}}{y_j}, 1, \frac{y_{j+1}}{y_j}, \ldots, \frac{y_{i-1}}{y_j}, \frac{1}{y_j}, \frac{y_{i+1}}{y_j}, \ldots, \frac{y_n}{y_j}. \end{aligned}$$

Lecture 5, 2/5/19

Last time we discussed the real projective space $\mathbb{R}P^n$. Today, we'll discuss the space $\mathbb{C}P^n$, a manifold of real dimension $2n$, which is the set of complex lines in $\mathbb{C}^{n+1}$. This is known as the complex projective space, and is constructed similarly to how the real projective plane is constructed.

Definition 0.6. Let M_1 and M_2 be n and m-dimensional manifolds, respectively. Then a mapping

$$\phi : M_1 \to M_2$$

is differentiable at a point $p \in M_1$ if, given a parameterization

$$\overline{y} : V \subset \mathbb{R}^m \to M_2$$

containing $\phi(p)$, there exists a parameterization

$$\overline{x} : U \subset \mathbb{R}^n \to M_1$$

with $p \in U$ such that $\phi(\overline{x}(U)) \subset \overline{y}(V)$ and the mapping

$$\overline{y}^{-1} \circ \phi \circ \overline{x} : U \subset \mathbb{R}^n \to \mathbb{R}^m$$

is differentiable at $\overline{x}^{-1}(p)$. It suffices to show that all partial derivatives of $\overline{y}^{-1} \circ \phi \circ \overline{x}$ exist and are continuous.

Note that because of our definition of a manifold, the above definition does not depend of a choice of parameterization.

The map $\overline{y}^{-1} \circ \phi \circ \overline{x}$ is called the expression of ϕ in the parameterization $\overline{x}$ and $\overline{y}$

There aren't many cases for how n and m can behave.

$n = 1, m > 1$. Let M be a differentiablemanifold. A differentiable curve $\alpha : (a, b) \subset \mathbb{R} \to M$ is a differentiable curve in M

$n > 1, m = 1$. This is when we smash a manifold into a real line. This is a strategy in Morse Theory.

$n = 1, m = 1$. This case is pretty clear.

How do we define an abstract tangent vector? We define an abstract tangent vector as a differentiable operator. This is because we donn't have an ambient space for M to live.

Recall: In $\mathbb{R}^n$, let $\alpha : (-\epsilon, \epsilon) \to \mathbb{R}^n$ be a differenitable curve with $\alpha(0) = p$. Write $\alpha(t)$ as

$$\alpha(t) = (x_1(t), x_2(t), \ldots, x_n(t))$$

with $t \in (-\epsilon, \epsilon)$. Let $\boldsymbol{v} = \alpha'(0) = (x_1'(0), x_2'(0), \ldots, x_n'(0))$. Let f be a differentiable function defined in a neighborhood of p. We can restrict the curve *alpha* and express the directional derivative with respect to the vector $\boldsymbol{v} \in \mathbb{R}^n$ as

$$\begin{aligned} \frac{df \circ \alpha}{dt}\Big|_{t=0} &= \frac{d}{dt} f(x_1(t), x_2(t), \ldots, x_n(t))\Big|_{t=0} \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\Big|_{t=0} \frac{dx_i}{dt}\Big|_{t=0} \\ &= \left(\sum x_i'(0) \frac{\partial}{\partial x_i}\right) f. \end{aligned}$$

The characteristic property of this is that the directional derivative with respect to $\boldsymbol{v}$ is an operator on differenitable functions that depend uniquely on $\boldsymbol{v}$.

Note that what we're doing here is pairing vectors $\alpha'(0)$ with covectors $f : \mathbb{R}^n \to \mathbb{R}$; or in other words, simply differentiable functions which return a value given a desired direction.

Gu says to imagine the vectors $\boldsymbol{v}$ as moms, and the babies as all the curves α for which $\alpha'(0) = \boldsymbol{v}$.

Definition 0.7. Let M be a differentiable manifold. Let $\alpha : (-\epsilon, \epsilon) \to M$ be a differentiable curve in M with $\alpha(0) = p \in M$. The **tangent vector** to the curve at $t = 0$ is a function $\alpha'(0) : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is the set of differenitable functions at p which take in $\alpha'(0)$. That is,

$$\alpha'(0)(f) = \frac{df \circ \alpha(t)}{dt}\Big|_{t=0}$$

Now a tangent vector at p of M is the tangent vector at $t = 0$ of some curve $\alpha : (-\epsilon, \epsilon) \to M$ with $\alpha(0) = p$

Definition 0.8. We say that the **tangent space** at of M at p is the set T_pM, the set of all tangent vectors to M.

Claim: T_pM is a vector space. Moreover, if we choose a parameterization $\overline{x} : U \to M$ then T_pM has a basis $\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})\}$.

Now we define a **dual basis** of $\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})\}$ to be

$$\{dx_1, dx_2, \dots, dx_n\}.$$

Those are differenntial "general" 1-forms of "base." For example: $3dx_1 + 5x_2^2dx_2 + 20x_1dx_3$.

A key idea is to mimick the basis onn $\mathbb{R}^n$. Recall taht $\{v_1, v_2, \dots, v_n\}$ is orthonormal if $v_i \cdot v_j = \delta_{ij}$. Say $w = a_1v_2 + \cdots + a_nv_n$. Then to obtain a_i, we simply dot the expression by v_i. This is the purpose of the dual basis, since we want to mimick this kind of behavior. Hence we have that for a basis $\{e_1, e_2, \dots, e_n\}$, we define the dual basis $\{e_1^*, e_2^*, \dots, e_n^*\}$ where

$$e_i^*(e_j) = \delta_{ij}.$$

For example, in linear regression, we have

$$y = \theta_0x_0 + \theta_1x_1 + \cdots + \theta_nx_n = \begin{pmatrix} \theta_0 & \theta_1 & \cdots & \theta_n \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Hence we have that $dx_i\left(\frac{\partial}{\partial x_j}\right) = \delta_{ij}$

Lecture 6, 2/10/19

Recall that on $\mathbb{R}^n$, every symmetric, positive definite matrix gives rise to an inner product on $\mathbb{R}^n$. We define it as

$$\langle x, y \rangle_A = x^TAy.$$

Recall that an inner product is positive definite, symmetric, and bilinear. And note that

1. $\langle x, x \rangle = x^TAx \geq 0$

2. $\langle x, y \rangle = x^TAy = (x^Tay)^T = y^TA^T(x^T)^T = y^TAx = \langle y, x \rangle$.

3. $\langle x, ay + bz \rangle = a\langle x, y \rangle = b\langle x, z \rangle$.

Hence we really do have an inner product.

Conjugate Gradient Method on a Euclidean space is an algorithm for the numerical solution of a particular systems of linear equations, namely those who matrix is symmetric and positive definite. That is, we care about

$$Ax = b$$

where A is symmetric, positive definite. But Prof Gu: does this occur often, i.e, do we often run into these types of problems? Yes. Examples include the least squares or maximum likelihood extimation. And these types of problems arise in supervise learning. Assume a model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n.$$

The data to be fit is of the form $\{(x^{(i)}, y^(i))\}_{i=1}^{N}$. Then we desire the best θ_i's where

$$y^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)}.$$

Hence we have the matrix equation

$$\begin{pmatrix} \vdots \\ y^{(i)} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \\ 1 & x_1^{(i)} & x_2^{(i)} & \cdots & x_n^{(i)} \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

The conjugate gradient method is often applied to sparse systems that are too large to be handled by a direct implementation or other direct methods such as the Cholesky decomposition. It can also be used to solve unconstrained optimization problems such as energy minimization.

Thus we care about the equation $Ax = b$ where A is symmetric. Note however that if A is not even symmetric, we can multiply both sides by A^T to observe that

$$Ax = b \implies A^T A x = A^T b$$

and then apply our method since $A^T A$ is positive definite and symmetric.

Definition 0.9. Let u, v be vectors. Then we say u, v are conjugate if

$$u^T A v = 0.$$

Since A is symmetric and positive definite, the left hand side defines an inner product:

$$\langle u, v \rangle_A := \langle Au, v \rangle = \left\langle u, A^T v \right\rangle = \langle u, Av \rangle = u^T A v.$$

(from wikipedia).

Note that a conjugate pair of vectors are orthogonal in the sense of the inner product induced by a symmetric positive definite matrix. Now consider a set $P = \{p_1, p_2, \ldots, p_n\}$ of n mutually conjugate, vectors, with respect to some symmetric postiive definite matrix A. In this inner product, we can turn this set into an orthonormal one. Then this set forms an orthonormal basis, and the solution can be expressed in this basis.

Hence let $x_* = \sum_{i=1}^{n} \alpha_i p_i$. Note that this is extremely useful. For example, if we want the α_i coefficient, then we can take the inner product with respect to A to extract it. That is,

$$p_i^T A p_j = \delta_{ij}.$$

Also observe that we have

$$\begin{aligned} Ax_* &= \sum_{i=1}^{n} \alpha_i A p_i \\ p_k^T A x_* &= \sum_{i=1}^{n} p_k^T \alpha_i A p_i \\ p_k^T b &= \sum_{i=1}^{n} \alpha_i \delta_{ik} \implies \langle p_k, b \rangle = \alpha_k \langle p_k, p_k \rangle \end{aligned}$$

so that $\alpha_k = \dfrac{\langle p_k, b\rangle}{\langle p_k, l_k\rangle_A}$. Thus orthonormal bases give us a systematic way of finding the α_i coefficients. Note that we don't need to find all elements of P; we can still approximate the system. Thus this is an interative method. This is the same idea where in PCA, we throw away the small eigenvalues and their directions since we only care about the larger ones.

Note that since $Ax_* = b$, we see that x_* is the unique minimization of

$$f(x) = \frac{1}{2} x^T A x = -x^T b.$$

The existence of a minimal vector is guaranteed by the fact that

$$Df(x) = \frac{1}{2}(Ax + x^T A) - b = Ax - b.$$

so

$$D^2 f(x) = A.$$

Hence the critical point is the global minimum! Thus to begin the algorithm, we set $p_0 = b - Ax_0$, and calculate onwards by taking p_i to be orthogonal (in our inner product) to be orthogonal to p_{i-1}. Each time, we check the **residual**

$$r_k = b - Ax_k$$

and calculate onwards, stopping when we are fine with the error.

Definition 0.10. A **real Grassmanian manifold** $G_k\mathbb{R}^n$ is the set of k-planes through the origin $\mathbb{R}^n$. On the other hand, a **Stiefel manifold** $V_k\mathbb{R}^n$ consists of k-orthonormal frames in $\mathbb{R}^n$. Hence one can be viewed as a submanifold of the other.

Note: in general, $G_k\mathbb{R}^n \cong G_{n-k}\mathbb{R}^n$.

Lecture 7, 2/12/19: Riemannian manifold and metric.

Recall that a vector space of dimension n, denoted V^n, equipped with an inner product $\langle\rangle$, forms an **inner product space**. We call it a Euclidean space is $n < +\infty$. Also recall that for a regular surface, equipped with a first fundamental form, forms a Riemannian surface.

Suppose we have a manifold M. Let p be a point, and consider the tangent plane T_pM equipped with a metric. . Then if we vary our point within some neighborhood, we obtain another similar tangent plane equipped with a different metric. We want these metrics to continuously coincide with one another as we drag the points closer; hence we obtain some notion of a differentiable operator.

We can generalize this concept even further to define a differentiable manifold equipped with a Riemannian metric forms a **Riemannian manifold.**

Definition 0.11. A **Riemannian metric** on a differentiable n-manifold M is a correspondence which associates to each $p \in M$ an inner product $\langle\rangle_p$ on the tangent space T_pM which varies differentiably in the following sense.

If $x : U \subset \mathbb{R}^n \to M$ is a system of coordinates around p, with $x(x_1, x_2, \ldots, x_n) = q \in X(U) \subset M$ and $\dfrac{\partial}{\partial x_i}(q) = dx_q(0, \ldots, 1, \ldots, 0)$, then

$$\left\langle \frac{\partial}{\partial x_i}(q), \frac{\partial}{\partial x_j}(q) \right\rangle_q = g_{ij}(x_1, \ldots, x_n)$$

is a differentiable function on U. The function g_{ij} is called the **local representation of the Riemannian metric**.

Definition 0.12. Let M and N be Riemannian manifolds. A diffeomorphism $f : M \to N$ is called a **isometry** if

$$\langle u, v\rangle_p = \langle df_p(u), df_p(v)\rangle_{f(p)}$$

for all $p \in M$ and $u, v \in T_pM$. Under these conditions, we then say that M and N are isometric.

On the other hand, we can define local isometries.

Definition 0.13. Let M and N be Riemannian manifolds. A differentiable mapping $f : M \to N$ is a **local isometry** at $p \in M$ if there exists a neigborhood $U \subset M$ of p such that $f : U \to f(U)$ is a diffeomorphism and df_p preserves the inner product, i.e.,

$$\langle u, v\rangle_p = \langle df_p(u), df_p(v)\rangle_{f(p)}$$

for all $p \in M$ and $u, v \in T_pM$.

Lie Groups.

Definition 0.14. A **lie group** $(G, \cdot)$ is a group which is also a manifold with a differenitable structure such that the group operations, multiplication $\cdot$ and inverse $\quad^{-1}$, are differentiable. That is, the map

$$\begin{aligned} &G \times G \to G \\ &(x, y) \mapsto xy^{-1} \end{aligned}$$

is differentiable. That is, the group stucture and manifold structureare both compatible.

Examples of Lie groups include $S^1, S^3, SO(3), O(n), U(n), \dots$. We can make even more examples by multiplying the Lie groups together, since the product of lie groups forms a lie group. For example, the product $SO(3) \times SO(3) \times SO(2)$ is a Lie group, which models a robotic arm.

A Lie group is equipped with both left and right transformations. For example, the **left transformation** is a map where

$$\begin{aligned} L_x &: G \to G \\ y &\mapsto xy = L_x(y) \end{aligned}$$

and a **right transformation** is similarly a map where

$$\begin{aligned} R_x &: G \to G \\ y &\mapsto yx = R_x(y). \end{aligned}$$

With that said, we may define a Riemannian metric which is **left invariant** to be one for which

$$\langle u, v\rangle_y = \langle (dL_x)_y(u), (dL_x)_y(v)\rangle_{L_x(y)}$$

for all $x, y \in G$ and $u, v \in T_yG$. The definition is similar for a **right invariant** Riemannian metric. If a Riemannian metric is both left and right invariant, we say it is **bi-invariant**.

Theorem 0.1. If a Lie group G is compact, then there exists a bi-invariant metric on G.

Product Metric.

Let M_1, M_2 be Riemannian manifolds. Consider the manifold $M_1 \times M_2$. Let $\pi_1 : M_1 \times M_2 \to M_1$, and $\pi_2 : M_1 \times M_2 \to M_2$ be the natural projection maps. Introduce the Riemannian metric on $M_1 \times M_2$, where for $u, v \in T_{(p,q)}M_1 \times M_2$,

$$\langle u, v\rangle_{(p,q)} = \langle d\pi_1(u), d\pi_2(v)\rangle_p + \langle d\pi_2(u), d\pi_2(v)\rangle_q$$

for all $(p, q) \in M_1 \times M_2$.

For example, consider the n-torus

$$T^n = S^1 \times S^1 \times \cdots \times S^1$$

where the product repeats n times. We can put a Riemannian metric on each $S^1 \subset \mathbb{R}^2$, and extend this to a Riemannian metric on T. Then the torus T^n with this metric is called a **flat torus**.

Theorem 0.2. Every differentiable manifold has a Riemannian metric.

Now with a Riemannian metric on a differentiable manifold, we can define all kinds of measurements, including the length of a curve on a manifold:

$$\int_a^b = \sqrt{\left\langle \frac{d\alpha}{dt}, \frac{d\alpha}{dt} \right\rangle} dt.$$

We can also find the volume. For any point p, obseve that

$$x_i(p) = \frac{\partial}{\partial x_i}(p) = a_{i1}e_1 + \cdots + a_{in}e_n$$

and

$$x_k(p) = \frac{\partial}{\partial x_k}(p) = a_{k1}e_1 + \cdots + a_{kn}e_n.$$

We can then define

$$g_{ik} = \langle x_i(p), x_k(p) \rangle$$

which builds a matrix $G = [g_{ik}] = A^T A$. We then have that

$$\text{vol}(x_1(p), \ldots, x_n(p)) = \det(A) = \sqrt{\det(G)}$$

Lecture 8, 2/17/20: Grassman Manifolds

Recall that the Riemannian metric on a manifold M is basically the assignment of an inner product on each tangent space of M which changes smoothly. Also recall that a Lie group is a manifold with a group structure. Examples include

$$O(n) = \{A \in M_{n\times n}(\mathbb{R}) \mid A^T A = 1\}$$
$$SO(n) = \{A \in M_{n\times n}(\mathbb{R}) \mid \det A = 1, A^T A = I\}$$

Note that for $A(t), B(t) \in O(n)$, we have that

$$[A(t)B(t)]' = A'(t)B(t) + A(t)B'(t).$$

We can also define the tangent vector for a curve $A(t) \in O(n)$ where $A'(0) = I$. Observe that $A^T A = I$ implies that

$$\begin{aligned}[A^T A]' = T' \implies [A^T(t)]'A(t) + A^T(tA'(t) = 0 \implies [A'(t)]TA(t) + A^T(t)TA(t) \\ \implies [A'(0)]^T = -A'(0)\end{aligned}$$

after plugging in $t = 0$. This then implies that $A'(0)$ is a skew symmetric matrix. The nice thing about skew symmetric matrices is that the exponential matrix becomes an orthogonal matrix. Thus we see that the tangent space of $O(n)$ consist of skew symmetric $n \times n$ matrices.

We also sometimes deal with the concept of a Lie subgroup. For example, $SO(2)$ is a Lie subgroup of $SO(3)$. More generally, we also have that $SO(n)$ is a Lie subgroup of $O(n)$.

Definition 0.15. A **homogeneous space** is a differentiable manifold M of the form G/H, where G is a Lie group and H is a **isotropy subgroup**.

Examples include $O(n), SO(n), G_k\mathbb{R}^n$, and $V_k\mathbb{R}^n$.

Definition 0.16. Let G be a group. Then G acts (left) on a set X if there is a function $\phi : G \times X \to X$ where $(g, x) \mapsto \phi(g, x)$ which satisfying the following properties.

1. For all x, $\phi(e, x) = x$.
2. For all $g, h \in G$, and for each $x \in X$, we have that

$$\phi(gh, x) = \phi(g, \phi(h, x)).$$

Acting **right** on a set is defined similarly.

Finally, we define the isotropy subgroup.

Definition 0.17. Suppose G acts on a manifold M. Then the stabilizer of the group action, Gx, is the **isotropy subgroup**.

Recall that a **Grassmanian manifold** is the set of all k-planes which pass through the origin of $\mathbb{R}^n$.

For example, consider $G = SO(3)$ and $M = G_2\mathbb{R}^3$. Take an element in $G_2\mathbb{R}^3$; say xy-plane. The isotropy subgroup of $SO(3)$ is all the group elements of $SO(3)$ whihch fixes the xy-plane in $G_2\mathbb{R}^3$. However, this isotropy subgroup is then then

$$\left\{ \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \theta \in [0, 2\pi) \right\} \cong SO(2).$$

Therefore, we can write this as

$$G_2\mathbb{R}^3 \cong SO(3)/(SO(2) \times SO(1)).$$

where $SO(1)$ appears since we not only fix the xy-plane, but we also fix the z-axis.

As another example, consider $G_2\mathbb{R}^4$. Then we can imagine these 2-planes passing through the origin of $\mathbb{R}^4$ as a rotation which fixes one 2-plane, and its orthogonal complement, another 2-plane. Then we have that

$$G_2\mathbb{R}^4 \cong SO(4)/(SO(2) \times SO(2)).$$

Suppose H_x is the isotropy group of x and H_y is the isotropy group of y, with $x, y \in M$. Then if the group action is transitive, $H_x \cong H_y$. This is because $y = gx$ for some $g \in G$. Thus $H_y = gH_xg^{-1}$; so they are not only isomorphic, but they are also isotropic.

Now in general, we have that

$$G_k\mathbb{R}^n \cong SO(n)/(SO(k) \times SO(n-k)).$$

Grassmanian manifolds are particularly useful since we can look at the tangent planes of an n-manifold, and instead move their n-planes to the origin of $\mathbb{R}^n$. For example, we can embed the tangent 2-planes by passing them through the origin of $\mathbb{R}^3$. This results in obtaining a curve in $G_k\mathbb{R}^n$.

Also recall that a **Stiefel manifold** is the set of all k orthonormal frames in $\mathbb{R}^n$. That is, we can imagine them as matrices with orthonormal columns.

Lecture 9, 2/19/20

Today we'll consider how to take derivatives on a manifold. On a manifold, derivatives are called **connections**. Recall in calculus, we studied vector fields $f : \mathbb{R}^n \to \mathbb{R}^m$. First, we considered **affine connections**.

Recall that $f : \mathbb{R}^n \to \mathbb{R}$. If $n = 2$, we can turn this into a graph $(x, y, f(x, y))$. For some point $p = (p_1, p_2)$ in the xy-plane, we can examine the derivative on the manifold at $(p_1, p_2, f(p_1, p_2))$. Now consider a staight line in the xy-plane given by $\alpha(t) = p + tv$ which passes through p. Then this line forms a curve on our manifold. The curve is given by $\beta(t) = f(\alpha(t))$. Now observe that

$$\beta'(t) = f'(\alpha(t))\alpha'(t) = f'(p + tv)v.$$

which is the **directional derivative**. We can examine this further: observe that if $v = (v_1, v_2)$, then $f(\alpha(t)) = f((p_1, p_2) + t(v_1, v_2)) = (p_1 + tv_1, p_2 + tv_2)$. We can then say $x(t) = p_1 + tv_1, y(t) = p_2 + tv_2$. Then

$$\beta'(t) = f(x(t), y(t)) = \frac{\partial f}{\partial x}x'(t) + \frac{\partial f}{\partial y}y'(t) \implies \beta'(p) = \begin{pmatrix} \frac{\partial f}{\partial x}x'(t) \\ \frac{\partial f}{\partial y}y'(t) \end{pmatrix} \cdot \begin{pmatrix} x'(0) \\ y'(0) \end{pmatrix}.$$

We then have the directional derivative $\nabla f \cdot v$. More generally, we can consider a function $f : M^n \to \mathbb{R}$ and define the directional derivative as

$$D_v f = \langle \nabla f, v \rangle_g$$

where the inner product is the Riemannian metric. Note we can think of this more abstractly as a vector, v, eating a function $f : M^n \to \mathbb{R}$. Recall that

- $a(v_p + bw_p)[f] = av_p[f] + bw_p[f]$
- $v_p[af + bg] = av_p[f] + bv_p[g]$
- $v_p[fg] = v_p[f] \cdot g(p) + f(p) \cdot v_p[g]$

If we can take the derivative of a vector rr^n, we can take the derivatives of vector fields on our manifold.

Definition 0.18. Let W be a vector field in $\mathbb{R}^3$ and let v be tangent vector on a manifold M in $\mathbb{R}^3$. Then the **covariant derivative of** W with respect to v is the tangent vector

$$\nabla_v W = W(p + tv)'(0).$$

at the point p. Here, we view W as a function we are differentiating, and v as what we're taking the derivative with respect to (i.e., the direction). Moreover, $\nabla_v W$ measures the initial rate of change of $W(p)$ as p moves in the direction of v.

For example, consider the vector field $F : \mathbb{R}^n \to \mathbb{R}^m$ where

$$F(x_1, \ldots, x_n) = [f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)]$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are real valued functions. Then

$$\nabla_v F = (\nabla f_1 \cdot v, \nabla f_2 \cdot v, \ldots, \nabla f_m \cdot v).$$

With that said, we still have the following properties, for all vector fields Y, Z, numbers a, b, and tangent vectors v and w of p.

1. $\nabla_{av+bw} Y = a\nabla_v Y + b\nabla_W Y$
2. $\nabla_v(aY + bZ) = a\nabla_v Y + v\nabla_v A$
3. $\nabla_v(fY) = v[f]Y(p) + f(p)\nabla_v Y$ with f a differentiable function and $v[f]$ the directional derivative
4. $v[Y \cdot Z] = \nabla_v Y \cdot Z(p) + Y(p) \cdot \nabla_v Z$

In general, there's no reason to memorize these rules; all derivatives behave in the same way! They always satisfy linearity, distributivity, and the product rule.

To study the intrinsic geometry of a surface, one needs to generalize Gauss' idea regarding normal vectors mapped to the unit sphere.

$$dN_p : T_p(S) \to T_p(S).$$

The Gauss map basically shoves all of the tangent planes into $G_1\mathbb{R}^3$, where we normalize the normals. Specifically, we take a local parameterization $x(u, v)$ on a manifold for a point of interest. We then take the derivative x_u, x_v and state

$$N = \frac{x_u \times x_v}{||x_u \times x_v||}$$

Definition 0.19. Let W be a vector field, and consider a point p on our manifold M. Let y be a tangent vector at y. If U is a neighborhood at p, we consider a curve $\alpha : (-\epsilon, \epsilon) \to U$ where $\alpha(0) = p$ and $\alpha'(0) = y$. Let $w(t) : (-\epsilon, \epsilon) \to U$, $W(\alpha(t))$, be the restriction of W on the curve α. Then the **covariant derivative at** p **of the vector field** W relative to y is the normal projection of $w'(0)$ onto the plane $T_p(S)$.

This quantity is denoted as $\dfrac{Dw}{dt}(0)$.

Lecture 10, 2/24/20

Recall the concept of the covariant derivative. We imagine a vector field defined on our manifold; this means that at every point on our manifold, there is a vector. We then project these vectors down onto the tangent planes from which they originate to extract a type of derivative.

On the other hand, an affine connection on a differentiable manifold is a function $\nabla(X, Y)$ which takes vector fields X, Y and outputs $\nabla_X Y$.

What's interesting to consider when this vector field is normal to the tangent plane; hence the projection on the tangent plane becomes zero.

Definition 0.20. A vector field w along a paramterized curive $\alpha : I \to M$ on our manifold M such that $Dw(\alpha)/dt = 0$ for every $t \in I$ is called **parallel**.

Definition 0.21. Let M be a differentiable manifold, and suppose it has an affine connection and a Riemannian metric $\langle\rangle$. Then a connection is **compatible** with the metri when for any parallel vector fields P, P' on c we have that $\langle P, P' \rangle$ is constant.

The idea of a covariant derivative is analgous to the idea of a dot product. Based on its properties, we can come up with all kinds of different ways of defining metrics. With the covariant derivative, we have a similar case since we can come up with all kinds of different derivatives by interchanging X and Y.

Lecture 12, 3/2/20

Currently, the largest open problems in machine learning involve optimization. The current most effective optimization methods include Newton's Method and the Conjugate Gradient method. Machine learning also requires problems to be well-modeled, which is currently achieved with Stiefel and Grassmanian manifolds.

Recall that Newton's method simply seeks to find the minimum values of a single variable function $f(x)$ over $\mathbb{R}$. This is achieved by inputing an initial guess, which leads to a sequence $\{x_k\}$ that converges to a minimum input value.

Let Z be any $n \times p$ matrix. We want to decompose Z into tangential and normal components. That is, we seek a form

$$Z = \pi_n(Z) + \pi_p(Z).$$

Claim: $\pi_n(Z) = Y\text{sym}(Y^T Z)$ and $\pi_T(Z) = Y\text{skew}(Y^T Z) + (I - YY^T)Z$. First, recall that

$$A = \text{sym}(A) + \text{skew}(A) = \frac{A + A^T}{2} + \frac{A - A^T}{2}.$$

Also recall that

$$g_E(\Delta_1, \Delta_2) = \text{tr}\Delta_1^T \Delta.$$

The normal space at a point y consists of all matrices N which satisfy $\{N \mid \mathrm{tr}(\Delta^T N) = 0 \text{ for all } \Delta \in T_p M\}$. The set of $\{(Y, N) \mid Y \in M, N \in \text{ normal space at } Y\}$ is called the tangent bundle of M.

Note $\pi_N(Z) = Y\mathrm{sym}(Y^T Z)$ is perpendicular to $T_Y V_{P,N}$. Since $\mathrm{tr}(\Delta^T \Pi_N(Z)) = \mathrm{tr}(\Delta^T Y \frac{Y^T Z + Z^T Y}{2})$, note that $\Delta^T Y$ that this is a skew symmetric matrix. However, we already know that $\mathrm{sym}(Y^T Z)$ is symmetric. But the trace of a skew symmetric matrix multiplied a symmetric matrix is zero. Hence we see that

$$\mathrm{tr}(\Delta^T \pi_N(Z)) = \langle \Delta, \pi_N(Z) \rangle = 0.$$

Therefore, we see that

$$\{\pi_N(Z) \mid \text{ for all } Z \text{ is a } n \times p \text{ matrix}\} \subset \text{normal space at } Y$$

But since they have the same dimension, we must conclude that

$$\{\pi_N(Z) \mid \text{ for all } Z \text{ is a } n \times p \text{ matrix}\} = \text{ normal space } .$$

This is because $\mathrm{dim} T_Y V_{p,n} = np - \frac{p(p+1)}{2} = \frac{p(p+1)}{2} + p(n-p)$.

Note

$$\pi_N(Z) = Y\mathrm{skew}(Y^T) + (I - YY^T)Z$$

is perpendicular to $\pi_N(Z)$ since

$$\langle \pi_N(Z), \pi_T(Z) \rangle = 0.$$

which implies that

$$\begin{aligned} \mathrm{tr}([\pi_n(Z)]^T \pi_T(Z)) = \mathrm{tr}[\mathrm{sym}(Y^T Z)]^T Y^T [Y\mathrm{skew}(Y^T Z) + (I - YY^T)Z] \\ = \mathrm{tr}[\mathrm{sym}(Y^T Z)(0) + 0] \\ = 0. \end{aligned}$$

In a similar fashion, this implies that $\mathrm{dim}\pi_T(Z) = \mathrm{dim} T_Y V_{p,n}$.