Compendium of Mathematics & Physics

Nicolas Dewolf

July 27, 2022

Contents

Chapter 1

Data Analysis

The main reference for the sections on optimization problems is [1]. For the geometry of clustering methods, see [2]. The main references for the section on *conformal prediction* are [3,4]. Although a part of this chapter is a continuation of the previous one, the focus here lies more on the computational aspect of the analysis of large data sets. For this reason the chapter starts with some sections on applied linear algebra (for a refresher see Chapter **??**).

1.1 Data sampling

1.1.1 Inverse CDF sampling

Although one of the most straightforward sampling algorithms, this approach makes the strong assumption that the cumulative distribution function **??** is invertible.

Method 1.1.1. Sample a point λ uniformly from the unit interval $[0, 1]$. This number gives the cumulative distribution of the point to be sampled. (The CDF $F_X(X)$ is itself uniformly distributed.) The new point x' is simply given by $F_X^{-1}(\lambda)$.

In the case where F_X is discrete λ might not lie in the image of F_X and the inverse might not admit an algorithmically useful expression, so one should use a different approach. Given a point $x \in \mathbb{R}$ and its associated cumulative probability $F_X(x)$ one can sample a new point x' as follows. One increases (or decreases) x until the unique point x' is found such that $F_X(x' - 1) < \lambda \leq F_X(x')$.

1.1.2 Uniform rejection sampling

This method again uses the fact that the value of the cumulative distribution function is itself uniformly distributed on the unit interval $[0, 1]$. The CDF does not have to be invertible for this method, but the probability density should be compactly supported.

Method 1.1.2. Consider an interval $[a, b]$ such that f_X vanishes outside this interval and let q_0 be an upper bound for f_X. Now, sample a point x' uniformly on $[a, b]$ and sample a point q uniformly on $[0, q_0]$. If $f_X(x') \geq q$, then x' is a good sample. If not, repeat this procedure.

The proof of that this algorithm works is quite easy and mainly depends on the fact that

$$\Pr(Q \leq f_X(X)) = \frac{1}{q_0(b-a)}.$$

This value is often called the **acceptance probability**. From this expression it is clear that if one chooses q_0 too large, the acceptance probability becomes very small and the algorithm will take a long time to produce a sample.

1.1.3 Monte Carlo sampling

The general idea of (Markov chain) Monte Carlo methods is to construct a sequence of points such that (starting from a given index) all points represent good samples and such that the sequence forms a Markov chain.

The first Monte Carlo algorithm uses an acceptance threshold:

Method 1.1.3 (Metropolis-Hastings). Assume that a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is given. The first element is given by $x_0 = \mu$. Subsequent points are constructed as follows:

- Sample a point x' from the normal distribution.
- Calculate the **acceptance ratio**

$$\lambda := \frac{f_X(x')}{f_X(x_{i-1})}.$$

- If $\lambda \geq 1$, $x_i = x'$.
- If not, sample a point q uniformly on $[0, 1]$. If $\lambda \geq q$, $x_i = x'$, else $x_i = x_{i-1}$.[1]

To obtain an efficient algorithm, it is helpful to choose $\mu = \mathrm{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$. This ensures that the points are sampled in a region that resembles the form of f_X.

In fact, this method can be drastically generalized. First of all it is possible to replace the normal distribution by any symmetric transition probability:

$$g(x' \mid x) = g(x \mid x'). \tag{1.1}$$

In case the transition probability is not symmetric, the acceptance ratio needs to be modified:

$$\lambda := \frac{f_X(x')g(x_{i-1} \mid x')}{f_X(x_{i-1})g(x' \mid x_{i-1})}. \tag{1.2}$$

Remark 1.1.4. It is clear from the definition of the acceptance ratio that one does not need f_X to be normalized. This avoids costly calculations of the normalization factor.

1.2 Optimization

1.2.1 Linear equations

Method 1.2.1 (Normal equation). Given the equation

$$Ax = b$$

as in Section **??**, one can try to numerically solve for x by minimizing the ℓ^2-norm $\|Ax - b\|^2$:

$$\hat{x} := \arg\min_x (Ax - b)^T (Ax - b). \tag{1.3}$$

This leads to the so-called normal equation[2]

$$A^T Ax = A^T b. \tag{1.4}$$

This can be formally be solved by $x = (A^T A)^{-1} A^T b$, where $(A^T A)^{-1} A^T$ is the pseudoinverse of A.

[1] This step could be merged with the previous one since $\lambda \geq 1$ always implies $\lambda \geq q$.

[2] The name stems from the fact that the equation $A^T Ax = A^T b$ implies that the residual is orthogonal (normal) to the range of A.

Remark 1.2.2. It is easy to see that the above linear problem is obtained when trying to extremize the quadratic form associated to a symmetric matrix.

Method 1.2.3 (Tikhonov regularization). Consider a linear (regression) problem

$$Ax = b.$$

The most straightforward way to solve for x is the least squares method introduced in Chapter **??**, where the solution is (formally) given by the normal equation: $x = (A^T A)^{-1} A^T b$. However, sometimes it might happen that A is nearly singular (it is said to be **ill-conditioned**). In this case a regularization term can be added to the minimization problem:

$$\|Ax - b\|^2 + \|\Gamma x\|^2, \tag{1.5}$$

where Γ is called the **Tikhonov matrix**. In the case that $\Gamma = \lambda \mathbb{1}$, one speaks of ℓ^2**-regularization**. This regularization technique benefits solutions with smaller norms.

Remark 1.2.4. The ℓ^2-regularization can be generalized by replacing the 2-norm by any p-norm $\|\cdot < |_p$. For $p = 1$ and $p = 2$ the names **lasso** and **ridge** regression are often used. For general $p \geq 0$ one sometimes speaks of **bridge** regression.

The minimization procedures for $p \leq 1$ have the important property that they not only shrink the coefficients, but even perform feature selection, i.e. some coefficients become identically zero. However, it can be shown that the optimization problem for $p < 1$ is nonconvex and, hence, is harder to solve. In general it is found that lasso regression gives the best results.

A benefit of ℓ^2-regularization is that it can be derived from a Bayesian approach. By choosing a Gaussian prior $\mathcal{N}(0, \lambda^{-1})$, Bayesian inference immediately gives the ℓ^2-regularized cost function as the posterior distribution. Accordingly the ℓ^2-regularized linear regressor is equivalent to the maximum a posteriori estimator with Gaussian priors. One can obtain ℓ^p-regularization in a similar way by replacing the Gaussian priors with generalized normal distributions (such as the Laplace distribution for $p = 1$).

Definition 1.2.5 (Multicollinearity). Consider a finite set of random variables $\{X_i\}_{1 \leq i \leq n}$. These random variables are said to be perfectly (multi)collinear if there exists an affine relation between them, i.e. there exist variables $\{\lambda_i\}_{0 \leq i \leq n}$ such that

$$\lambda_0 + \lambda_1 X_1 + \cdots + \lambda_n X_n = 0. \tag{1.6}$$

The same concept can be applied to data samples. The data is said to be (multi)collinear if the above equation holds for all entries of the data set. However, in this case one also define ''near multicollinearity'' if the variables X_i are related as above up to some error term ε. If the variance of ε is small, the matrix $X^T X$ might have an ill-conditioned inverse which might render the algorithms unstable.

Definition 1.2.6 (Variance inflation factor). The VIF is an estimate for how much the variance of a coefficient is inflated by multicollinearity. The VIF of a coefficient β_i is defined as follows:

$$\mathrm{VIF}_i := \frac{1}{1 - R_i^2}, \tag{1.7}$$

where R_i^2 is the R^2-value obtained after regressing the predictor $\hat{X}_i$ on all other predictors. The rule of thumb is that VIF ≥ 10 implies that a significant amount of multicollinearity is present in the model.

1.2.2 Gradient descent

The gradient descent algorithm is first introduced in the case of quadratic forms:

Method 1.2.7 (Steepest descent). Consider the quadratic form

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

Assume that A is symmetric and positive-definite such that $Ax = b$ gives the minimum of f. Like most recursive algorithms, gradient descent starts from an arbitrary guess x_0. It then takes a step in the direction of steepest descent (or largest gradient), i.e. in the direction opposite to $\nabla f(x_0) = Ax_0 - b =: -r_0$:

$$x_{i+1} := x_i + \alpha r_i. \tag{1.8}$$

The quantities r_i are called the **residuals**. This procedure is repeated until convergence, i.e. until the residual vanishes up to a fixed numerical tolerance.

A naive gradient descent method would require to fine-tune the step size α. However, a more efficient method is given by the **line search algorithm**, where the value of α is optimized in every step as to minimize f along the line defined by r_i. A standard calculus argument leads to the following form of the step size:

$$\alpha_i = \frac{r_i^T r_i}{r_i^T A r_i}. \tag{1.9}$$

This choice forces the descent direction to be orthogonal to the previous one since

$$\frac{d}{d\alpha} f(x_i) = -\nabla f(x_i) \cdot \nabla f(x_{i-1}).$$

As a consequence this minimization scheme often results in a chaotic zigzag trajectory through the configuration space. The higher the **condition number** $\kappa = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$, the worse the zigzag motion will be. A very narrow valley (or some higher-dimensional analogue) will make the trajectory bounce back and forth between the walls, instead of moving towards the minimum.

1.2.3 Conjugate gradient

As noted in the previous section, a common problem with gradient descent is that the direction of steepest descent is often not the same as the direction pointing to the optimal solution and, hence, convergence might only occur after a long time.

A simple solution can be obtained by considering multiple orthogonal directions and taking a suitable step once in every direction. This way one obtains an algorithm that converges in n steps, where n is the dimension of the coefficient matrix A. By requiring that the error at step $i+1$ is orthogonal to the direction d_i, it is assured that no direction is used twice. However, the main problem with this idea is that the exact error e_i is not known and, hence, one cannot calculate the required steps.

By modifying the orthogonality condition one can avoid this problem. This is the idea behind conjugate direction methods:

Definition 1.2.8 (Conjugate vectors). Consider a symmetric positive-definite matrix A. Any such matrix induces an inner product as follows:

$$\langle v|w\rangle_A := v^T A w. \tag{1.10}$$

Two vectors v, w are said to be (A-)conjugate if they are orthogonal with respect to $\langle\cdot|\cdot\rangle_A$. The general approach to obtain a basis of A-conjugate vectors is a modified version of the Gram-Schmidt procedure **??** where the ordinary Euclidean inner product is replaced by (1.10). This modification is called the **Arnoldi method**.

By taking the input vectors of the Arnoldi method to be the residuals r_i, one obtains the **conjugate gradient** (CG) algorithm. It is interesting to note that the residuals themselves satisfy a recursion relation:

$$r_{i+1} = r_i - \alpha_i A d_i, \tag{1.11}$$

where the step size α_i is defined similar to the step size for ordinary steepest descent:

$$\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}. \tag{1.12}$$

Since the directions are constructed using the residuals, they span the same subspace. By denoting the subspace spanned by the first i directions by $\mathcal{D}_i$, the relation $r_{i+1} \in \mathcal{D}_i + A d_i$ leads to the following expression because of the above recursion relation:

$$\mathcal{D}_i = \text{span}\big\{r_0, Ar_0, \ldots, A^{i-1} r_0\big\}. \tag{1.13}$$

Because of their prominence in the literature on numeric optimization techniques, these subspaces have earned their own name:

Definition 1.2.9 (Krylov subspace). A vector space $\mathcal{K}$ of the form

$$\mathcal{K} := \text{span}\{v, Av, \ldots, A^n v\} \tag{1.14}$$

for some matrix A, vector v and natural number $n \in \mathbb{N}$. Given such an A and v, one often denotes the associated Krylov subspace of dimension n by $\mathcal{K}_n(A, v)$.

The fact that the spaces $\mathcal{D}_i$ are Krylov spaces also has an import implication for the numerical complexity of the CG algorithm. The residual r_{i+1} can be shown to be orthogonal to the space $\mathcal{D}_{i+1}$ (this is generally called the **Galerkin condition**). But since $A\mathcal{D}_i \subset \mathcal{D}_{i+1}$, this also implies that r_{i+1} is A-conjugate to $\mathcal{D}_i$. It follows that the only relevant contribution in the Arnoldi method is given by the last direction d_i. This reduces the complexity (both time-wise and memory-wise) per iteration from $O(n^2)$ to $O(n)$.

The steps in the CG algorithm are summarized below:

Method 1.2.10 (Conjugate gradient). Let x_0 be the initial guess with the associated residual $r_0 := b - Ax_0$ acting as the first direction vector d_0. The following scheme gives an iterative n-step (n being the dimension of the coefficient matrix A) algorithm to obtain the solution to $Ax = b$:

$$\alpha_i := \frac{r_i^T r_i}{d_i^T A d_i} \tag{1.15}$$

$$x_{i+1} := x_i + \alpha_i d_i \tag{1.16}$$

$$r_{i+1} := r_i - \alpha_i A d_i \tag{1.17}$$

$$d_{i+1} := r_{i+1} + \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} d_i. \tag{1.18}$$

Remark 1.2.11. In exact arithmetic the above optimization scheme would result in an exact solution after n iterations (in fact the number of iterations is bounded by the number of distinct eigenvalues of A). However, in real life one is not working in exact arithmetic and one has to take into account the occurrence of floating-point errors. These not only ruin the accuracy of the residual recursion relation (1.17), but more importantly[3] it might result in the search directions not being A-conjugate.

Now, what about general coefficient matrices A, for example those resulting in under- or overdetermined systems? For nonsymmetric or nondefinite square matrices one can still solve the normal equation (1.4) using the same methods, since A^TA is both symmetric and positive-definite. For underdetermined systems an exact solution does not always exact, but the numerical methods will always be able to find a solution that minimizes the ℓ^2-error. For overdetermined systems A^TA will be nonsingular and the numerical methods can find an exact solution. However, the condition number of A^TA is the square of that of A and, hence, the algorithms will convergence much slower.

A different approach exists where the CG algorithm is not applied to the matrix A^TA, but the individual matrices are used A, A^T directly. This way not one Krylov space is generated, but two dual ''copies'' are constructed:

$$\mathcal{D}_i := \operatorname{span}\{r_0, Ar_0, \dots, A^{i-1}r_0\}$$
$$\widetilde{\mathcal{D}}_i := \operatorname{span}\{\widetilde{r}_0, A^T\widetilde{r}_0, \dots, (A^T)^{i-1}\widetilde{r}_0\},$$

where $\widetilde{r}_0$ does not have to be related to r_0. In this case there are two Galerkin conditions $r_i \perp \mathcal{D}_i$ and $\widetilde{r}_i \perp \widetilde{\mathcal{D}}_i$ (only the first one is relevant). The residuals form biorthogonal bases of the Krylov subspaces:

$$\langle r_i | r_j \rangle = \|r_i\|^2 \delta_{ij}. \tag{1.19}$$

As a consequence the search directions also form biconjugate bases:

$$\langle d_i | d_j \rangle_A = \|d_i\|_A^2 \delta_{ij}. \tag{1.20}$$

1.2.4 Nonlinear conjugate gradients

Of course, many real-world applications are determined by nonlinear equations and, hence, it would be pleasant if one could salvage some of the above ideas even when linear algebra is not the natural language. The main requirement would be that one can calculate the gradient of the function to be minimized.

On the level of the implementation, the structure of the algorithm remains more or less the same. What does change is the form of the Arnoldi method, in particular, the prefactor in Equation (1.18). For linear CG there are multiple equivalent formulas, but for nonlinear CG these do not lead to the same algorithm. The two most common choices are given below.

Method 1.2.12 (Nonlinear CG). Since there is no linear equation related to the minimization problem, the residuals are always defined as $r_i := -\nabla f(x_i)$. The algorithm consists of the following iterations:

$$\alpha_i := \arg\min_{\alpha} f(x_i + \alpha d_i) \tag{1.21}$$
$$x_{i+1} := x_i + \alpha_i d_i \tag{1.22}$$
$$r_{i+1} := -\nabla f(x_i) \tag{1.23}$$
$$d_{i+1} := r_{i+1} + \beta_{i+1} d_i, \tag{1.24}$$

[3] The residual problem can be solved by computing the residual ''exactly'', i.e. by the formula $r_i = b - Ax_i$, every k iterations.

where β_{i+1} is computed by one of the following formulas:

- **Fletcher-Reeves formula**:

$$\beta_{i+1} := \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}. \tag{1.25}$$

- **Polak-Ribière formula**:

$$\beta_{i+1} := \max\left\{\frac{r_{i+1}^T (r_{i+1} - r_i)}{r_i^T r_i}, 0\right\}. \tag{1.26}$$

Some general remarks have to be made concerning the nonlinear CG algorithm:

Remark 1.2.13. As was already mentioned for the linear version, floating-point errors might lead to a loss of conjugacy. For the nonlinear extension this becomes worse. The more f deviates from a quadratic function, the quicker conjugacy is lost (for quadratic formulas the Hessian is exactly the matrix A, but for higher-degree functions the Hessian varies from point to point). Another problem, one that did not occur for quadratic functions, is that nonlinear functions might have multiple local minima. The CG method does not care about local vs. global and, hence, it will not necessarily converge to the global minimum. A last remark concerns the fact that there is no theoretical guarantee that the method will converge in n steps. Since the Gram-Schmidt procedure can only construct n conjugate vectors, the simplest solution is to perform a restart of the algorithm every n iterations.[4]

For linear CG a simple formula for finding the optimal value of α_i was obtained. However, for nonlinear CG one cannot solve Equation (1.21) as easily. The main idea, i.e. that f' should be orthogonal to the previous search direction remains, is still valid. Here, only the **Newton-Raphson approach** is considered:[5]

$$\alpha_i = \frac{\nabla f(x_i)^T d_i}{d_i^T \mathrm{Hess} f(x_i) d_i}. \tag{1.27}$$

To obtain the optimal α-value, one should iteratively apply the Newton-Raphson method in every CG iteration. If the action of the Hessian f'' on d_i cannot be simplified, i.e. if the full Hessian has to be computed in every iteration, this can lead to considerable computational overhead. The general rule of thumb is to perform only a few Newton-Raphson iterations and obtain a less accurate but more efficient algorithm. To make sure that the search descent direction is indeed a direction of descent (and not one of ascent), one can check that $r^T d \geq 0$ and restart the procedure if it is negative.

1.2.5 Krylov methods

Generally one starts from an iterative fixed-point based technique to solve the linear equation $Ax = b$ as before, i.e. one iterates $x_{i+1} = b + (\mathbb{1} - A)x_i$. Using the residuals $r_i = b - Ax_i$ this can be rewritten as

$$x_i = x_0 + \sum_{k=0}^{i-1} r_k = x_0 + \sum_{k=0}^{i-1} (\mathbb{1} - A)^k r_0. \tag{1.28}$$

It is clear that this results in $x_i - x_0 \in \mathcal{K}_i(A, r_0)$. The main idea is then to find optimal degree-k polynomials P_k such that $x_i - x_0 = \sum_{k=0}^{i-1} P_k(A) r_0$.

[4] The max operation in Equation (1.26) is already a form of restarting, due to the fact that the Polak-Ribière version of nonlinear CG sometimes results in cyclic behaviour.

[5] Another common method is the *secant method*.

Method 1.2.14 (Jacobi method). Consider a linear problem $Ax = b$ where A has spectral radius less than 1. First, decompose A as the sum of a diagonal matrix D and and a matrix E with zero diagonal elements. If one assumes that D is invertible, the following recursive scheme is obtained:

$$x_{i+1} := D^{-1}(b - Ex_i). \tag{1.29}$$

A sufficient condition for convergence is strict diagonal dominance, i.e. $|D_{ii}| > \sum_{j \neq i} |E_{ij}|$.

?? COMPLETE (e.g. Lanczos)??

1.3 Constrained optimization

1.3.1 Lagrange multipliers

A common generalization of the above optimization problems is the addition of constraints involving equalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad g_i(x) = 0 \qquad \forall 1 \leq i \leq n. \tag{1.30}$$

The general approach to solving such constrained problems is by extending the optimization loss:

Method 1.3.1 (Lagrange multipliers). Given a constrained optimization problem of the form (1.30), one can construct the enhanced loss function

$$\mathcal{L}(x, \lambda_1, \ldots, \lambda_n) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x). \tag{1.31}$$

The solution to the original problem is obtained by extremizing this loss with respect to x and the Lagrange multipliers λ_i (as usual this might fail globally for nonconvex problems):

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial \mathcal{L}}{\partial \lambda_i} = 0 \quad \forall 1 \leq i \leq n. \end{cases} \tag{1.32}$$

The situation becomes even more interesting when one also allows constraints involving inequalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad \begin{cases} g_i(x) = 0 & \forall 1 \leq i \leq m \\ h_j(x) \leq 0 & \forall 1 \leq j \leq n. \end{cases} \tag{1.33}$$

Problems of this form are called **primal optimization problems**. By defining an enhanced loss using Lagrange multipliers as before

$$\mathcal{L}(x, \alpha, \beta) := f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{n} \beta_i h_i(x), \tag{1.34}$$

it is not hard to see that

$$\max_{\alpha, \beta; \beta_j \geq 0} \mathcal{L}(x, \alpha, \beta) = \begin{cases} \infty & \text{if a constraint is violated} \\ f(x) & \text{if all constraints are satisfied.} \end{cases} \tag{1.35}$$

Definition 1.3.2 (Primal optimization problem). Denote the maximum of $\mathcal{L}(x, \alpha, \beta)$ by $\theta_P(x)$.

$$p^* := \min_x \theta_P(x) = \min_x \max_{\alpha,\beta;\beta_i \geq 0} \mathcal{L}(x, \alpha, \beta). \tag{1.36}$$

By interchanging the max and min operators in the primal formulation, another problem is obtained:

Definition 1.3.3 (Dual optimization problem).

$$d^* := \max_{\alpha,\beta;\beta_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha,\beta;\beta_i \geq 0} \min_x \mathcal{L}(x, \alpha, \beta). \tag{1.37}$$

From basic calculus it is known that $\max \min \leq \min \max$ and, hence, that $d^* \leq p^*$. The difference $p^* - d^*$ is called the **duality gap** and, if $d^* = p^*$, one says that **strong duality** holds. The real question then becomes: ''*When does strong duality hold?*''.

Definition 1.3.4 (Slater conditions). Consider a convex optimization problem, i.e. a problem of the form (1.33) where f is convex, the g_i are convex and the h_j are affine. This problem is said to satisfy the Slater condition(s) if there exists an x that is strictly **feasible**, i.e. $h_j(x) < 0$ for all $1 \leq j \leq n$.

Property 1.3.5 (Strong duality). If a convex problem satisfies the Slater conditions, strong duality holds. The solutions x and (α, β) that attain this duality are called primal optima and dual optima respectively.

The following property gives a set of sufficient conditions:

Property 1.3.6 (Karush-Kuhn-Tucker conditions). If there exist x, α and β such that strong duality holds, the following conditions are satisfied:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \end{cases} \quad \forall 1 \leq i \leq m \qquad \text{and} \qquad \begin{cases} \beta_j h_j(x) = 0 \\ h_j(x) \leq 0 \\ \beta_j \geq 0. \end{cases} \quad \forall 1 \leq j \leq n \tag{1.38}$$

Conversely, if there exists values x, α and β that satisfy the KKT conditions, they give strongly dual solutions for the primal and dual problems.

Remark 1.3.7 (Complementary slackness). The third equation in the KKT conditions has an important implication. It says that if there is an index j such that the constraint h_j is not **active**, i.e. $h_j(x) < 0$, the associated Lagrange multiplier is 0 and, conversely, if there is an index j such that the Lagrange multiplier $\beta_j > 0$, the constraint h_j is active.

Remark 1.3.8. It is not hard to see that the KKT conditions reduce to the conditions for Lagrange multipliers when all h_j are identically 0. For this reason the quantities α and β are called the **KKT multipliers**.

1.3.2 Riemannian gradient descent

In many situations the full parameter space of an optimization problem is constrained in such a way that the resulting admissible subset admits the structure of a smooth manifold and, in particular that of a Riemannian manifold. When trying to extend gradient descent algorithms to this setting, one has to take into account that most manifolds are not linear spaces and, hence, that linear updates will often lead outside the manifold.

The first point that we have to treat is the occurence of the gradient in these algorithms. In ordinary Euclidean space, one simply takes the gradient to be the vector of partial derivatives. However, on general smooth manifolds this object is actually given by the de Rham differential, which is a covariant vector. However, a short proof shows that the Riemannian gradient from Remark **??** actually gives the direction of steepest ascent. So even on Riemannian manifolds the gradient is the correct direction to work with. However, as mentioned above, the form of the update will be a problem in general.

?? COMPLETE ??

1.4 Approximation theory

1.4.1 Bayes optimality

Definition 1.4.1 (Bayes risk). The minimal risk over all models:

$$R^* := \inf_{f:\mathcal{X}\to\mathcal{Y}} R(f). \tag{1.39}$$

Definition 1.4.2 (Bayes classifier). Given a joint probability distribution P over the instance space $\mathcal{X} \times \mathcal{Y}$ and a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the pointwise Bayes predictor is defined as follows:

$$f^* : x \mapsto \arg\min_{y\in\mathcal{Y}} \int_{\mathcal{Y}} l(y, y')dP(y' \mid x). \tag{1.40}$$

Property 1.4.3 (Bayes optimality). The risk of the pointwise Bayes predictor is minimal, i.e. $R(f^*) = R^*$. In practice, however, one cannot achieve Bayes optimality (through the pointwise Bayes predictor), since this would require the knowledge of the distribution.

Definition 1.4.4 (Approximation error). Given a data set $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$, the empirical risk R_{emp} is defined as follows:

$$R_{\text{emp}} : \hat{y} \mapsto \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} l(\hat{y}(x), y). \tag{1.41}$$

Consider a hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (e.g. selected by a choice of model architecture). The minimizer of the true risk in $\mathcal{H}$ is denoted by h^*, while the empirical risk minimizer is denoted by $\hat{h}$. The **approximation error/uncertainty** is defined as the difference $R(\hat{h}) - R(h^*)$, while the **model uncertainty** is defined as the difference $R(h^*) - R^*$. The total error made by $\hat{h}$ (with respect to the Bayes optimum) is then simply the sum of these two.

1.4.2 PAC theory and empirical risk minimization

Property 1.4.5. Note that by the (strong) law of large numbers **??**, the empirical risk converges to the true risk almost surely whenever the data poitns are sampled i.i.d. However, in practice, the true question becomes how fast this convergence happens.

Definition 1.4.6 (Probably approximately correct). A model $\hat{f}$ is said to be ε-accurate at the confidence level $1 - \delta$ (with respect to the hypothesis space $\mathcal{H}$), or simply (ε, δ)-PAC, if it satisfies

$$\Pr\left(R(\hat{f}) - \inf_{f\in\mathcal{H}} R(f) > \varepsilon\right) < \delta. \tag{1.42}$$

Corollary 1.4.7. If one can prove a PAC bound, one gets the following bound for free:

$$R(\hat{f}) \leq \inf_{f \in \mathcal{H}} R(f) + 2\varepsilon \tag{1.43}$$

with probability $1 - \delta$, i.e. with high probability the true risk for the minimizer is only slightly higher than the empirical risk and, hence, the minimizer is a reasonbaly good estimate.

Property 1.4.8 (Bounded losses). Consider a hypothesis space $\mathcal{H}$ with a bounded loss function. The probability that for a given number of data points $n \in \mathbb{N}_0$ their exists at least one model $f \in \mathcal{H}$ for which the empirical risk deviates significantly from the true risk is bounded as follows:

$$\begin{aligned}
\Pr(\exists f \in \mathcal{H} : |R_{\text{emp}}(f) - R(f)| \geq \varepsilon) &= \Pr(\sup_{f \in \mathcal{H}} |R_{\text{emp}}(f) - R(f)| \geq \varepsilon) \\
&\leq \sum_{f \in \mathcal{H}} \Pr(|R_{\text{emp}}(f) - R(f)| \geq \varepsilon) \\
&\leq \sum_{f \in \mathcal{H}} 2e^{-2n\varepsilon^2} \\
&= 2|\mathcal{H}|e^{-2n\varepsilon^2},
\end{aligned} \tag{1.44}$$

where the second inequality comes from Hoeffding's inequality **??** (the empirical risk is an average). In combination with a similar expression for the one-sided tails, using the one-sided Hoeffding inequality, the PAC bound becomes

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{2n}}, \tag{1.45}$$

with probability at least $1 - \delta$. Note that this bound is distribution-free, i.e. it does not depend on the data-generating distribution.

This bound on the empirical risk also implies a bound on the expected risk of the minimizer $\hat{f}$:

$$\mathrm{E}\Big[R(\hat{f})\Big] \leq \inf_{f \in \mathcal{H}} R(f) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{2n}} + \delta, \tag{1.46}$$

where $\delta > 0$ is now arbitrary.

The issue with the above two bounds is that they only apply to the case where the hypothesis space $\mathcal{H}$ is finite. To obtain useful bounds for countable hypothesis spaces, a new tool is required:

Definition 1.4.9 (Complexity regularizer). A function $c : X \to \mathbb{R}^+$ such that

$$\sum_{x \in X} e^{-c(x)} \leq 1. \tag{1.47}$$

A simple example would be the log-probabilities $c(x) := \ln(P(x))$ when a probability distribution P on X is given.

Using a complexity regularizer on $\mathcal{H}$, the following risk bound can be obtained with probability $1 - \delta$:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{c(f) + \ln(1/\delta)}{2n}}. \tag{1.48}$$

1.5 Classification problems

1.5.1 Clustering

Probably the most well-known and simplest algorithm for clustering in the unsupervised setting is the k-means algorithm:

Method 1.5.1 (k-means algorithm). Assume that an unlabelled dataset $\mathcal{D} \subset \mathbb{R}^n$ is given. For every integer $k \in \mathbb{N}$, usually satisfying $k \ll |\mathcal{D}|$, and any choice of k distinct **centroids** $\{c_i \in \mathbb{R}^n\}_{i \leq k}$, the k-means algorithm is defined through the following iterative scheme:

1. To every point $d \in \mathcal{D}$ assign a cluster C_i based on the following criterion:
$$i = \arg\min_{j \leq k} \|d - c_j\|^2. \tag{1.49}$$

2. Update the centroids c_i to represent the center of mass of the associated cluster C_i:
$$c_i \longleftarrow \frac{1}{|C_i|} \sum_{d \in C_i} d. \tag{1.50}$$

This algorithm optimizes the following global cost function with respect to the centroids c_i:

$$\mathcal{L}_{k\text{-means}}(c_1, \ldots, c_k) = \sum_{i=1}^{k} \sum_{d \in C_i} \|d - c_i\|^2. \tag{1.51}$$

Given the above idea, one could ask for a more general algorithm where clustering is performed with respect to a divergence function **??**. In the case of Bregman divergences **??** it can be shown that all one needs to do is replace the Euclidean distance by the divergence D_f:

Property 1.5.2 (Centroid position). Let D_f be a Bregman divergence. The minimizer

$$\arg\min_{\kappa} \sum_{i=1}^{k} D_f(x_i \| \kappa) \tag{1.52}$$

is given by the arithmetic average

$$\kappa = \frac{1}{k} \sum_{i=1}^{k} x_i. \tag{1.53}$$

If instead of a cluster $C = \{x_i \in \mathbb{R}^n\}_{i \leq k}$, one is given a probability distribution p, one simply has to replace the arithmetic average by the expectation value with respect to p. It can be furthermore be shown that for any Bregman divergence the k-means algorithm always converges in a finite number of steps (however, the clustering is not necessarily optimal).

The cluster boundaries $H(c_1, c_2) = \{x \in \mathbb{R}^n \mid D_f(x\|c_1) = D_f(x\|c_2)\}$ admit a simple geometric construction:

Property 1.5.3 (Cluster boundaries). Let D_f be a Bregman divergence and consider the k-means problem associated to D_f for $k = 2$ (higher-dimensional problems can be treated similarly). The boundary $H(c_1, c_2)$ is exactly the geodesic hypersurface orthogonal to the dual geodesic connecting c_1 and c_2. This partitioning of the data manifold is a generalization of *Voronoi diagrams* to (Bregman) divergences.[6]

[6] See [5] for more information. This is also introduced in [6], but there the author has confusingly interchanged the affine and dual coordinates.

1.5.2 Nearest neighbour search

?? COMPLETE ??

1.6 Garden

?? ADD (e.g. trees, forests) ??

1.7 Support-vector machines

1.7.1 Kernel methods

This section will introduce the mathematics of kernel methods. This mainly involves the language of Hilbert spaces (see Chapter **??** for a refresher).

Definition 1.7.1 (Kernel[7]). A function $k : X \times X \to \mathbb{C}$ that is (conjugate) symmetric and for which the Gram-matrix $K_{ij} := K(x_i, x_j)$ is positive-definite for all $n \in \mathbb{N}$ and $\{x_i \in X\}_{i \leq n}$.

Definition 1.7.2 (Reproducing kernel Hilbert space). A Hilbert space $\mathcal{H} \subset \text{Map}(X, \mathbb{C})$ of functions over a set X for which all evaluation functionals $\delta_x : f \mapsto f(x)$ are bounded (or continuous by Property **??**). Reproducing kernel Hilbert spaces are often abbreviated as **RKHS**s.

Using the Riesz representation theorem **??** one can express every evaluation functional δ_x on $\mathcal{H}$ as a function $K_x \in \mathcal{H}$. This allows for the introduction of a kernel on X:

Definition 1.7.3 (Reproducing kernel). Let $\mathcal{H}$ be an RKHS on a set X. The (reproducing) kernel k on X is defined as follows:

$$k(x, y) := \delta_x(K_y) \overset{\text{Riesz}}{=} \langle K_x | K_y \rangle_{\mathcal{H}}. \tag{1.54}$$

Because k is given by an inner product, it is not hard to see that the reproducing kernel is a kernel 1.7.1.

Starting from a kernel one can also characterize an RKHS as follows:

Alternative Definition 1.7.4 (RKHS). A Hilbert space $\mathcal{H} \subset \text{Map}(X, \mathbb{C})$ of functions over a set X such that there exists a kernel k on X with the following properties:

1. **Reproducing property**: For all $x \in X, f \in \mathcal{H}$ the evaluation functional δ_x satisfies $\delta_x(f) = \langle k(\cdot, x) | f \rangle_{\mathcal{H}}$.
2. **Density**: The span of $\{k(\cdot, x) \mid x \in X\}$ is dense in $\mathcal{H}$.

The density property is often replaced by the property that $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$.

Property 1.7.5 (Convergence). In an RKHS, convergence in norm implies pointwise convergence.

Theorem 1.7.6 (Moore-Aronszajn). *There exists a bijection between RKHSs and kernels.*

Proof. One direction of the theorem is, as mentioned before, rather simple to see. The other direction is constructive:

Given a kernel k, one defines the function $K_x := k(\cdot, x)$ for all $x \in X$. The RKHS is then constructed as the Hilbert completion of $\text{span}\{K_x \mid x \in X\}$, where the inner product is

[7] Also called a **Mercer kernel**. See Mercer's theorem below for more information.

defined as follows

$$\left\langle \sum_{x\in X} a_x K_x \middle| \sum_{y\in X} b_y K_y \right\rangle := \sum_{x,y\in X} \overline{a_x} b_y k(x,y). \tag{1.55}$$

Formula 1.7.7. Let $\mathcal{H}$ be an RKHS with kernel k. If $\{e_i\}_{i\leq \dim(\mathcal{H})}$ is an orthonormal basis for $\mathcal{H}$, then

$$k(x,y) = \sum_{i=1}^{\dim(\mathcal{H})} e_i(x)\overline{e_i(y)}. \tag{1.56}$$

Remark 1.7.8. Note that one can use different conventions in the above definitions, e.g. the definition $k(x,y) = \langle K_y | K_x \rangle_{\mathcal{H}}$ is also valid.

Theorem 1.7.9 (Mercer). *Let X be a finite measure space and consider a (conjugate) symmetric function $k \in L^2(X\times X, \mathbb{C})$. If k satisfies the* ***Mercer condition***

$$\iint_{X\times X} k(x,y)\overline{f(x)}f(y)dxdy \geq 0 \tag{1.57}$$

for all $f \in L^2(X,\mathbb{C})$, the Hilbert-Schmidt operator

$$T_k : L^2(X,\mathbb{C}) \to L^2(X,\mathbb{C}) : f \mapsto \int_X k(\cdot,x)f(x)dx \tag{1.58}$$

admits a countable orthonormal basis $\{e_i\}_{i\in\mathbb{N}}$ with nonnegative eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ such that

$$k(x,y) = \sum_{i=1}^{\infty} \lambda_i e_i(x)\overline{e_i(y)}. \tag{1.59}$$

Theorem 1.7.10 (Bochner). *A continuous function satisfies the Mercer condition if and only if it is a kernel.*

Alternative Definition 1.7.11 (Kernel). Consider a set X. A function $k : X\times X \to \mathbb{C}$ is called a (Mercer) kernel on X if there exists a Hilbert space $\mathcal{H}$ together with a function $\phi : X \to \mathcal{H}$ such that

$$k(x,y) = \langle \phi(x) | \phi(y) \rangle_{\mathcal{H}}. \tag{1.60}$$

When using Mercer's theorem, the feature maps are given by

$$\phi_i : x \mapsto \sqrt{\lambda_i} e_i(x). \tag{1.61}$$

Remark 1.7.12. The kernel expressions in the Mercer and Moore-Aronszajn theorems are related by the fact that the RKHSs induced by kernels satisfying the assumptions of the Mercer theorem are of the form

$$\mathcal{H} = \left\{ f \in L^2(X,\mathbb{C}) \middle| \sum_{i=1}^{\infty} \frac{\langle f, e_i\rangle_{L^2}^2}{\lambda_i} < +\infty \right\}. \tag{1.62}$$

Remark 1.7.13 (Vector-valued functions). Much of this section can be generalized to the setting of vector-valued functions $f : X \to \mathbb{C}^d$. In this case the kernels $k : X\times X \to \mathbb{C}$ are generalized to a matrix-valued functions $k : X\times X \to \mathbb{C}^{d\times d}$.

1.7.2 Decision boundaries

Consider a linear model for a classification problem $y = w^T x + b$. The object x_i is said to belong to the positive (resp. negative) class if $y > 0$ (resp. $y < 0$). This is implemented by the sign activation function

$$\mathrm{sgn}(y) = \begin{cases} 1 & y > 0 \\ -1 & y < 0 \end{cases} \tag{1.63}$$

to the linear model. The **decision boundary** $y = 0$, where the decision becomes ambiguous, forms a hyperplane in the feature space. However, it should be clear that in generic situations there are multiple hyperplanes that can separate the two classes for a finite number of data points. The problem then becomes to obtain the hyperplane with the maximal separation, i.e. the hyperplane for which the distance to the nearest data point is maximal.

The unit vector $\frac{w}{\|w\|}$ defines the normal to the hyperplane and, therefore, one can obtain the distance $d(x)$ from a data point x to the decision boundary by projecting onto this unit vector. The point $x - d(x)\frac{w}{\|w\|}$ is an element of the decision boundary and, hence, satisfies the hyperplane equation. Rewriting this gives an expression for the distance

$$d(x) = \frac{w^T x + b}{\|w\|}. \tag{1.64}$$

To account for the direction of the arrow, this number should be multiplied by the class $\mathrm{sgn}(y) = \pm 1$. This result is called the **geometric margin** $\gamma(x) := \mathrm{sgn}(y) d(x)$. The numerator in the geometric margin is called the **functional margin**. The geometric margin is preferable since it is invariant under simultaneous scale transformations of the parameters w, b.

The optimization objective now becomes

$$\max_w \frac{\gamma}{\|w\|} \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq \gamma \|w\| \qquad \forall 1 \leq i \leq n, \tag{1.65}$$

where $\gamma = \min_{i \in \{1,\ldots,n\}} \gamma(x_i)$ for x_i ranging over the training set. The problem is formulated in terms of the functional margin $\gamma\|w\|$ to avoid the nonconvex constraint $\|w\| = 1$. This allows the application of the Slater conditions for strong duality. Since the geometric margin is invariant under scale transformations, one can without loss of generality work with the assumption $\gamma\|w\| = 1$. The optimization problem is then equivalent to the following minimization problem:

$$\min_w \|w\|^2 \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq 1 \qquad \forall 1 \leq i \leq n. \tag{1.66}$$

The KKT conditions for this problem give the following results:

$$w = \sum_{i=1}^{n} \beta_i y_i x_i \tag{1.67}$$

and

$$\sum_{i=1}^{n} \beta_i y_i = 0, \tag{1.68}$$

where the quantities β_i are the KKT multipliers for the affine constraints $1 - y_i(w^T x_i + b) \leq 0$. Using these relations the quantity y can be expressed for a new data point as follows:

$$y \equiv w^T x + b = \sum_{i=1}^{n} \beta_i y_i \langle x_i | x \rangle + b. \tag{1.69}$$

Two observations can be made at this point. First of all, complementary slackness 1.3.7 implies that the only relevant vectors x_i in this calculation are the ones that satisfy $\gamma(x_i) = 0$. These are called the **support vectors** and they give their name to a class of models called **support-vector machines** (SVMs). These are the models that are trained using the above optimization problem. Furthermore, y can be written in terms of an inner product. It is exactly this last observation that allows for the generalization of the above model to nonlinear decision boundaries. The previous section showed that inner products are equivalent to (Mercer) kernels. Hence, by choosing a nonlinear kernel function, one can implicitly work with nonlinear feature maps. This is often called the **kernel trick**. As an example, polynomial kernels represent feature maps from x to monomials in the coefficients of x.

However, as often happens with data analysis algorithms, this procedure is sensitive to outliers. This is especially the case for kernels that are based on feature maps to infinite-dimensional spaces (e.g. the *RBF kernel*). To solve this problem one can introduce a regularization term in the cost function. The simplest such term for support-vector machines is a simple ℓ^1-penalty:

$$\min_w \|w\|^2 + C\sum_{i=1}^{n} \xi \qquad \text{such that} \qquad \begin{cases} \xi_i \geq 0 & \forall 1 \leq i \leq n \\ y_i(w^T x_i + b) \geq 1 - \xi_i & \forall 1 \leq i \leq n. \end{cases} \tag{1.70}$$

The resulting KKT conditions are as follows:

$$0 \leq \beta_i \leq C \tag{1.71}$$

and

$$\beta_i = 0 \implies y_i(w^T x_i + b) \geq 1 \tag{1.72}$$

$$\beta_i = C \implies y_i(w^T x_i + b) \leq 1 \tag{1.73}$$

$$\beta_i \in]0, C[\implies y_i(w^T x_i + b) = 1. \tag{1.74}$$

?? COMPLETE (e.g. geometry)??

1.8 Vapnik-Chervonenkis theory

1.8.1 VC dimension

Definition 1.8.1 (Shatter coefficient). Let Ω denote the universe of discours and consider a set $C \subset P(\Omega)$. C **shatters** a set $A \subset \Omega$ if for every subset $a \subseteq A$ there exists a subset $c \in C$ such that

$$a = c \cap A. \tag{1.75}$$

The shatter(ing) coefficients of C are defined as follows:

$$S_n(C) := \max_{x_1,\ldots,x_n \in \Omega} \big|\big\{\{x_1, \ldots, x_n\} \cap c \,\big|\, c \in C\big\}\big|. \tag{1.76}$$

It should be clear that every shatter coefficient S_n is bounded above by 2^n. If $S_n(C) = 2^n$, then C shatters some set of cardinality n.

Given a collection of binary functions $\mathcal{F} \subseteq \{0,1\}^X$, the shatter(ing) coefficients (or **growth functions**) of $\mathcal{F}$ are given by:

$$S_n(\mathcal{F}) := \max_{x_1,\ldots,x_n \in X} \big|\big\{\{f(x_1), \ldots, f(x_n)\} \,\big|\, f \in \mathcal{F}\big\}\big|. \tag{1.77}$$

The shatter coefficients give a notion of the effective size of $\mathcal{F}$, since they say how many different results the functions can produce given a data set of n points.

Definition 1.8.2 (Vapnik-Chervonenkis dimension). The Vapnik-Chervonenkis dimension of a collection of functions $\mathcal{F}$ is defined as follows:

$$\mathrm{VC}(\mathcal{F}) := \max\{k \in \mathbb{N} \mid S_k(\mathcal{F}) = 2^k\}. \tag{1.78}$$

Note that if a collection shatters n points, it also necessarily shatters a subset of these points, therefore, one also has

$$S_n(\mathcal{F}) = 2^n \tag{1.79}$$

for all $n < \mathrm{VC}(\mathcal{F})$.

A collection is called a **Vapnik-Chervonenkis class** if its VC dimension is finite.

Property 1.8.3 (Sauer's lemma[8]). The VC dimension bounds shatter coefficients in the following way:

$$S_n(\mathcal{F}) \leq \sum_{k=0}^{\mathrm{VC}(\mathcal{F})} \binom{n}{k} \tag{1.80}$$

for all $n \in \mathbb{N}$. For $n \leq \mathrm{VC}(\mathcal{F})$, the right-hand side is just the full binomial expansion for 2^n and, accordingly, the shatter coefficients grow exponentially. For $n \geq d$, the binomial series is truncated and polynomial behaviour is obtained:

$$\forall n \geq \mathrm{VC}(\mathcal{F}) : S_n(\mathcal{F}) \leq \left(\frac{e\,n}{\mathrm{VC}(\mathcal{F})}\right)^{\mathrm{VC}(\mathcal{F})}, \tag{1.81}$$

where e is the Euler number.

The generalization bounds of Section 1.4.2 on empirical risk minimization can be extended to uncountable hypothesis spaces as follows:

Property 1.8.4 (Generalization bound). The expected risk of the empirical risk minimizer $\hat{f}$ is bounded as follows:

$$R(\hat{f}) \leq \inf_{f \in \mathcal{H}} R(f) + 4\sqrt{2\frac{\mathrm{VC}(\mathcal{H})\ln(e\,n/\mathrm{VC}(\mathcal{H})) + \ln(2/\delta)}{n}} \tag{1.82}$$

with probability at least $1 - \delta$.

?? CHECK ALL THESE BOUNDS ??

1.8.2 Rademacher complexity

Remark 1.8.5 (Real-valued functions). The VC dimension of a collection of arbitrary real-valued functions can be defined as the VC dimension of the corresponding collection of indicator functions (Heaviside functions):

$$\mathrm{VC}(\mathcal{F}) := \mathrm{VC}\big(\{\theta(f(x) - \lambda) \,\big|\, f \in \mathcal{F}, \lambda \in \mathbb{R}\}\big). \tag{1.83}$$

This definition is equivalent to the following one based on subgraphs:

$$\mathrm{VC}(\mathcal{F}) = \mathrm{VC}\big(\{C_f := \{(x, \lambda) \in X \times \mathbb{R} \mid \lambda < f(x)\} \,\big|\, f \in \mathcal{F}\}\big). \tag{1.84}$$

[8] Sometimes called the **Sauer-Shelah lemma**.

Example 1.8.6 (Linear spaces). Every vector space V of real-valued functions has VC dimension at most $\dim(V) + 1$.

Example 1.8.7 (Translations). The set of translations of a real-valued function has VC dimension 1.

Although this remark says that one can in theory extend ordinary VC theory to arbitrary (real-valued) functions, this does not mean that the VC bounds obtained before above make sense in this setting. To obtain useful bounds, it is important to introduce a new notion of effective size:

Definition 1.8.8 (Rademacher complexity). Consider a collection of functions $\mathcal{F} := \{f : X \to \mathbb{R}\}$. The Rademacher complexity is defined as follows:

$$\mathfrak{R}_n(\mathcal{F}) := \mathrm{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right], \tag{1.85}$$

where the σ_i are Rademacher variables **??** and the expectation is taken over both the Rademacher variables and the sample $(X_1, \ldots, X_n)$.

Property 1.8.9 (Risk bound). Consider a collection of bounded functions $\mathcal{F} \subseteq [a, b]^{\mathcal{X}}$.

$$\mathrm{E}[R(f)] \leq R_{\mathrm{emp}}(f) + 2\mathfrak{R}_n(\mathcal{F}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \tag{1.86}$$

with probability at least $1 - \delta$.

Property 1.8.10 (VC dimension). The shatter coefficient bounds the Rademacher complexity in the following way:

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\ln(S_n(\mathcal{F}))}{n}}. \tag{1.87}$$

1.8.3 Relation to Glivenko-Cantelli classes

Property 1.8.11. Recall Property **??**. The empirical L^1-norm only depends on the values of the functions $f \in \mathcal{F}$ at the given data points and, therefore, the covering number of $\mathcal{F}$ is bounded above by the covering number of $\{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$. The latter is itself bounded above by the shatter coefficient $S_n(\mathcal{F})$. If $\mathcal{F}$ has finite VC dimension, Sauer's lemma 1.8.3 implies that this coefficient grows polynomial in n, so

$$\frac{1}{n}\ln N_C(\varepsilon, \mathcal{F}_M, \|\cdot\|_1) \xrightarrow{d} 0 \tag{1.88}$$

and, thus, $\mathcal{F}$ is Glivenko-Cantelli.

Theorem 1.8.12. *A class of sets is Vapnik-Chervonenkis if and only if it is Glivenko-Cantelli* **??**.

1.9 Time series analysis

Definition 1.9.1 (Time series). A $\mathbb{N}$- or $\mathbb{Z}$-indexed stochastic process. Since $\mathbb{N}$ and $\mathbb{Z}$ are isomorphic in a very simple way, the two conventions for time series will be used interchangeably.

1.9.1 Stationarity

Definition 1.9.2 (Strict stationarity). A time series $(X_n)_{n\in\mathbb{N}}$ is (strictly) stationary if for any two integers $r, s \in \mathbb{N}$, the joint distribution satisfies the following condition:

$$P(X_{t_1}, \ldots, X_{t_r}) = P(X_{t_1+s}, \ldots, X_{t_r+s}). \tag{1.89}$$

Definition 1.9.3 (Weak stationarity). A time series $(X_n)_{n\in\mathbb{N}}$ is said to be weakly (or **covariance**) stationary if it satisfies the following conditions:

1. **Mean-stationary**: $\mathrm{E}[X_n] = \mathrm{E}[X_0]$ for all $n \in \mathbb{N}$.
2. **Finite covariance**: $\mathrm{cov}(X_i, X_j) < \infty$ for all $i, j \in \mathbb{N}$.
3. **Covariance-stationary**: $\mathrm{cov}(X_i, X_{i+j}) = \mathrm{cov}(X_0, X_j)$ for all $i, j \in \mathbb{N}$.

The following definition is a reformulation of Birkhoff ergodicity **??**:

Definition 1.9.4 (Ergodicity). A time series $\{X_t\}_{t\in\mathbb{Z}}$ is ergodic if for every measurable function f the following equation holds for all $t \in \mathbb{Z}$:

$$\lim_{T\to\infty} \frac{1}{2T+1} \sum_{k=-T}^{T} f(X_k) = \mathrm{E}[f(X_t)]. \tag{1.90}$$

Intuitively this means that state space averages can be evaluated as time averages.

1.9.2 Correlation

Definition 1.9.5 (Autocorrelation function). Consider a time series $(X_n)_{n\in\mathbb{N}}$. The autocovariance (resp. autocorrelation) function of this time series is defined as the covariance (resp. autocorrelation) function of the random variables $(X_n)_{n\in\mathbb{N}}$.

Definition 1.9.6 (Spectral density). Consider a (weakly) stationary time series $(X_n)_{n\in\mathbb{N}}$. If the associated autocovariance is in ℓ^1, one can define the spectral density as the discrete Fourier transform of the autocovariance function:

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{i\omega k}, \tag{1.91}$$

where $\gamma(k)$ is the autocovariance function at lag k.

Under the assumption that the spectral density exists, the time series is said to have **short memory** if $f(0)$ is finite. Otherwise the series is said to have **long memory**.

Definition 1.9.7 (Lag operator[9]). The lag operator sends a variable in a time series to the preceding value:

$$BX_t = X_{t-1}. \tag{1.92}$$

An important concept, especially in the context of autoregressive models, is that of a **lag polynomial** (the notation for these is not completely fixed in the literature, but the θ-notation is a common choice):

$$\theta(B) = 1 + \sum_{i=1}^{k} \theta_i B^i \tag{1.93}$$

$$\varphi(B) = 1 - \sum_{i=1}^{k} \varphi_i B^i. \tag{1.94}$$

[9] Also called the **backshift operator**.

Notation 1.9.8 (Difference operator). The difference operator Δ is defined as follows:

$$\Delta = 1 - B. \tag{1.95}$$

In a similar way one can define the **seasonal** difference operator:

$$\Delta_s = 1 - B^s. \tag{1.96}$$

Method 1.9.9 (Ljung-Box test). A test to see if a given set of autocorrelations of a time series is different from zero. Consider a time series of n elements and let $\{\rho_i\}_{1\leq i\leq k}$ be the first k lagged autocorrelation functions. The test statistic is defined as

$$Q = n(n+2)\sum_{i=1}^{k}\frac{\rho_k}{n-k}. \tag{1.97}$$

If the null hypothesis ''there is no correlation'' is true, the Q-statistic will asymptotically follow a χ^2-distribution with k degrees of freedom.

Method 1.9.10 (Augmented Dickey-Fuller test). Consider a time series $(X_t)_{t\in T}$. The (augmented) Dickey-Fuller test checks if the time series is (trend) stationary. For this test one considers the following regression model (similar to the ARIMA-*models* discussed in the next section):

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^{p-1}\theta_i \Delta X_{t-i} + \varepsilon_t. \tag{1.98}$$

The test statistic is

$$\mathrm{DF} = \frac{\gamma}{\mathrm{SE}(\gamma)}, \tag{1.99}$$

where SE denotes the standard error. The null hypothesis states that $\gamma = 0$, i.e. there is a *unit root* $(1-B)$ present in the model. Comparing the test statistic to tabulated critical values will give an indication whether to reject the hypothesis or not (the more negative the statistic, the more significant the result).

1.9.3 Autoregressive models

Definition 1.9.11 (AR(p)-model). Consider a time series $(X_t)_{t\in T}$. The autogressive model of order p is defined as the multiple linear regression model of X_t with respect to the first p lagged values $X_{t-1}, \ldots, X_{t-p}$ of the time series:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + \varepsilon_t. \tag{1.100}$$

Definition 1.9.12 (Partial autocorrelation function). The p^{th} autocorrelation function is defined as the p^{th} coefficient in the AR(p)-model.

Remark 1.9.13. The optimal order p of an autoregressive model is the one for which all higher partial autocorrelation functions (almost) vanish.

Definition 1.9.14 (MA(p)-model). Consider a time series $(X_t)_{t\in T}$ where every X_t contains a white noise contribution $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$. The moving average model of order p is defined as the multiple linear regression model of X_t with respect to the first p lagged values $\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}$ of the error term:

$$X_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \cdots + \beta_p \varepsilon_{t-p} + \varepsilon_t. \tag{1.101}$$

Since the error terms are assumed to have mean zero, one can see that the intercept term β_0 gives the mean of the time series.

Remark 1.9.15. The optimal order p of an autoregressive model is the one for which all higher autocorrelation functions (almost) vanish.

Definition 1.9.16 (Invertibility). An MA(q)-model is said to be invertible if all roots of its associated lag polynomial $\theta(B)$ lie outside the unit circle. This condition implies that the polynomial is invertible, i.e. $1/\theta(B)$ can be written as a convergent series in the operator B. This in turn implies[10] that one can write the MA(q)-model as an AR(p)-model, where possibly $p = \infty$. The analogous property for AR(p)-models leads to a definition of **stationarity**.

In practice it is not always possible to describe a data set using either an autoregressive or a moving average model. However, these two types of models can be combined:

Definition 1.9.17 (ARMA(p, q)**-model).**

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t \tag{1.102}$$

As above, one can find the optimal values for p and q by analyzing the autocorrelation and partial autocorrelation functions.

Using the lag polynomials one can rewrite the ARMA(p, q)-model as follows:

$$\varphi(B)X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{1.103}$$

By considering the special case where the polynomial $\mathcal{B}_\alpha^-$ has a unit root $1 - B$ with multiplicity d, one can obtain a generalization of the model:

$$\varphi(B)(1 - B)^d X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{1.104}$$

The interpretation of this additional factor $(1 - B)^d$ is related to the stationarity of the time series. The operator $1 - B$ is a finite difference operator:

$$\begin{aligned}(1 - B)X_t &= X_t - X_{t-1}\\ (1 - B)^2 X_t &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})\\ &\dots\end{aligned}$$

By successive applications, one can obtain a stationary time series from a nonstationary time series. This combination of differencing, autoregression and moving averages is called the **ARIMA**-model[11].

Remark 1.9.18. Including so-called *exogenous* variables, i.e. external predictors, leads to an **ARIMAX**-model.

Remark 1.9.19 (Fitting AR- and MA-models). As is clear from the definition of an AR(p)-model, the parameters θ_i can easily be found using standard techniques for multivariate linear regression such as ordinary least squares. However, in contrast to AR-models where the predictors are known, the estimation of coefficients in MA-models is harder since the error terms ε_t are by definition unknown.

To estimate the coefficients in a MA-model, people have introduced multiple techniques (see for example [7]). One of the most famous ones is the method by *Durbin*:

[10] Sometimes this is used as a definition of invertibility.

[11] The 'I' stands for "integrated".

Method 1.9.20 (Durbin). By restricting to invertible MA(q)-models (or by approximating a noninvertible model by an invertible one), one can first fit an AR(p)-model with $p > q$ to obtain estimates for the errors ε_t and then, in a second step, use a least squares-method to solve for the coefficients in the MA-model.

As a last modification one can introduce seasonal components. Simple trends such as a linear growth are easily removed from the time series by detrending or differencing. However, a periodic pattern is harder to remove and, in general, ARIMA-models are not suited to accompany this type of features. Luckily one can easily modify the ARIMA-model to incorporate seasonal variations. The multiplicative SARIMA-model is obtained by inserting operators similar to the ones of the ordinary ARIMA-model, where the lag operator B is replaced by the seasonal lag operator B^s (where s is the period of the seasonal variation):

Definition 1.9.21 (ARIMA$(p,q,d)(P,Q,D)_s$**-model).**

$$\Phi(B^s)\varphi(B)\Delta_s^D\Delta^d X_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{1.105}$$

1.9.4 Causality

Definition 1.9.22 (Granger causality). Consider two time series $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$. The time series X_n is said to Granger-cause Y_n if past values of X_n help to predict future values of Y_n. More formally this can be stated as follows:

$$P[Y_{t+k} \in A \mid \Omega(t)] \neq P[Y_{t+k} \in A \mid \Omega\backslash X(t)] \tag{1.106}$$

for some k, where $\Omega(t)$ and $\Omega\backslash X(t)$ denote the available information at time t with and without removing the variable X from the universe.

This formulation of causality was introduced by *Granger* under the following two assumptions:

- The cause always happens prior to the effect.
- The cause carries unique information about the effect.

Remark 1.9.23. A slightly different but for computational purposes often more useful[12] notion of Granger-causality is as follows. A time series $(X_n)_{n\in\mathbb{N}}$ is said to **Granger-cause** a time series $(Y_n)_{n\in\mathbb{N}}$ if the variance of predictions of Y_n becomes smaller when the information contained in X_n is taken into account.

Remark 1.9.24. Assume that two uncorrelated models giving predictions of a time series are given. One way to check if they have the same accuracy is the *Diebold-Mariano test*. However, when testing for Granger-causality one should pay attention. This test is not valid for nested models and, hence, is not applicable to two models that only differ by a set of extra predictors (in this case an external time series).

1.10 Uncertainty modelling

1.10.1 Prediction regions

One of the simplest ways to express uncertainty about predictions or parameter estimates is to give a set of possible values instead of a single value. However, to be meaningful, these sets should satisfy some conditions.

[12] In fact this was the original definition by *Granger*.

Definition 1.10.1 (Validity). Consider a measurable function $\Gamma : X \to P(Y)$ and let P be the joint distribution on the instance space $Z \equiv X \times Y$. Γ is said to be valid (at **significance level** $\alpha \in [0, 1]$ or **confidence level** $1 - \alpha$) if it satisfies

$$P(y \in \Gamma^\alpha(x)) \geq 1 - \alpha. \tag{1.107}$$

One sometimes also distinguishes between exact validity and **conservative** validity, where the former is the subcase of the latter for which the inequality becomes an equality.

In fact, one can define two notions of validity: pointwise and asymptotic. Equation (1.107) characterizes pointwise validity in the sense that the probability of having an error is given by a Bernoulli process with parameter α. Asymptotic validity is a frequentist notion in the following sense:

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma)}{n} \leq \alpha\,, \tag{1.108}$$

where $\mathrm{Err}_n(\Gamma)$ is the number of errors made by Γ after n trials. It should be clear that pointwise validity (both exact and conservative) implies asymptotic validity.

Remark 1.10.2 (Confidence regions). The definition of valid confidence predictors above is similar to the definition of confidence regions (Section **??**). However, in contrast to confidence regions for population parameters, the size of confidence regions for predictive distributions does not go towards zero in the infinite data limit. This follows from the fact that in general all observations are subject to noise and, hence, even with perfect knowledge about the data generating distribution, an exact prediction is impossible.[13]

1.10.2 Conformal prediction

A very general framework for the construction of valid prediction intervals in a model-independent manner is given by the conformal prediction framework by *Vovk et al.* The main ingredients for the construction are randomization and ''conformity measures''.

The first step will be studying the behaviour under randomization of the existing data (be it measurements or past predictions). To ensure that the procedure satisfies the required confidence (or probability) bounds, one has to make some assumptions. One of the main benefits of this framework is that one can relax the condition of the data being i.i.d. to it being exchangeable:

Definition 1.10.3 (Exchangeability). Consider an ordered data sample $\{z_i\}_{1\leq i\leq N}$. The joint distribution $P(z_1, \ldots, z_N)$ is said to be exchangeable if it is invariant under any permutation of the data points. A distribution Q is said to be exchangeable if Q^n is exchangeable for all $n \in \mathbb{N}$.

This definition can be restated in a purely combinatorial way. First, define the notion of a **bag** obtained from a (possibly ordered) data sample $\{z_i\}_{1\leq i\leq N}$ as the (unordered) multiset $\mathcal{B}$ containing these elements. The joint distribution P is then said to be exchangeable if the probability of finding any sequence of data points is equal to the probability of drawing this same sequence from the bag of these elements. Since this probability is purely combinatorial and, hence, completely independent of the ordering, it should be clear that this coincides with the first definition. The set of bags in a space X is sometimes denoted by X^∞.

Definition 1.10.4 (Nonconformity measure). Consider a bag $\mathcal{B}$ together with a new element z^* in a probability space (Z, Σ, P). A noncomformity measure $A : Z^\infty \times Z \to \mathbb{R}$ is a measurable function that gives a number indicating how different z^* is from the content of $\mathcal{B}$.

[13] Note that when observations are not sampled according to a distribution, but are perfectly predictable, this is of course not true.

Remark. One could restate all statements in this section in terms of ‘‘conformity measures’’ and, hence, look at similarities instead of dissimilarities. It will become clear that the procedure is invariant under monotone transformations and, hence, everything can be multiplied by -1.

Example 1.10.5 (Point predictors). A general class of nonconformity measures is obtained from point predictors for a metric space (Y, d) . Given a point predictor $\rho : X \to Y$ trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_\rho(\mathcal{B}, (x, y)) := d\big(\rho(x), y\big). \tag{1.109}$$

Example 1.10.6 (Interval predictors). For every interval predictor $(l, u) : X \to \mathbb{R}^2$, i.e. a pair of functions $u, l : X \to \mathbb{R}$ such that $\forall x \in X : l(x) \leq u(x)$, trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_{u,l}(\mathcal{B}, (x, y)) := \max\big(l(x) - y, y - u(x)\big). \tag{1.110}$$

It should be noted that although common, nonconformity measures and, by extension, conformal prediction are also applicable to nonmetric spaces:

Example 1.10.7 (Nested region predictors). Let T be a totally ordered set **??**. For every model $(C_t)_{t\in T} : X \to P(Y)$ that predicts a sequence of nested regions, i.e.

$$s \leq t \implies C_s(x) \subseteq C_t(x)\,, \tag{1.111}$$

trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_T(\mathcal{B}, (x, y)) := \inf\{t \in T \mid y \in C_t(x)\}. \tag{1.112}$$

Construction 1.10.8 (Conformal predictor). Consider a data sample given as a bag $\mathcal{B} \in Z^\infty$ together with a nonconformity measure A and let α denote the confidence level of the prediction region to be constructed. For any new element $z^* \in Z$ the algorithm proceeds as follows:

1. Denote the nonconformity score $A(\mathcal{B}, z^*)$ by μ_{z^*}.
2. For every element z in $\mathcal{B}$, define μ_z by replacing z by z^* in the bag and calculating the nonconformity score as in the previous step.
3. Calculate the conformal p-value as the fraction of elements $z \in \mathcal{B} \cup \{z^*\}$ for which $\mu_z \geq \mu_{z^*}$:
$$p^* := \frac{|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z \geq \mu_{z^*}\}|}{|\mathcal{B}| + 1}. \tag{1.113}$$
4. Include an element z^* in the prediction region C^α if and only if $p^* > \alpha$.

It should be noted that, in general, the construction of these regions can be quite time-consuming because if the nonconformity measure A depends on model that has to be trained on $\mathcal{B}$, this training has to be reperformed for every element $z \in \mathcal{B}$ in step 2. For low-dimensional regions it can often be achieved by solving inequalities derived from the specific form of the nonconformity measure.

Property 1.10.9 (Optimality). A conformal predictor satisfies the following conditions:

- Regions are (conservatively) valid, i.e.
$$P(p^* \leq \alpha) \leq \alpha\,, \tag{1.114}$$
and thus also
$$P(y^* \in \Gamma^\alpha(x^*, \mathcal{B})) \geq 1 - \alpha\,, \tag{1.115}$$
where the probability is taken over both the bag $\mathcal{B}$ and the point (x^*, y^*).

- Regions are nested, i.e. $\alpha \leq \beta \implies \forall x \in X : \Gamma^{\alpha}(x) \subseteq \Gamma^{\beta}(x)$.

Given any region predictor satisfying these three properties below, there exists a conformal predictor that is more efficient, i.e. produces smaller prediction regions.

Property 1.10.10 (Smooth conformal predictors). One can modify the above construction in such a way that the resulting conformal predictors are not only conservatively valid but are also exactly valid. To this end one replaces the conformal p-value (1.113) by

$$p^*(\theta) := \frac{|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z > \mu_{z^*}\}| + \theta\,|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z = \mu_{z^*}\}|}{|\mathcal{B}| + 1}\,, \tag{1.116}$$

where θ is independently and uniformly sampled from the unit interval $[0, 1]$. Exact validity is then obtained by also marginalizing over the random variable θ.

Now, one could wonder if the assumption of exchangeability is a realistic assumption. Obviously if one applies the procedure to independent observations, everything is fine since i.i.d. sequences are clearly exchangeable. However, some important classes of data sequences are clearly not exchangeable, e.g. time series. This kind of data often contains intrinsic correlation and, hence, the exchangeability assumption is almost always violated. However, a solution exists. One can restate the construction above using an explicit randomization as is done in [8]. There, one replaces the nonconformity measure by a function that acts on ordinary sequences instead of unordered bags. The fraction p^* can then be expressed as follows:

$$p^* = \frac{1}{|S_{N+1}|} \sum_{\sigma \in S_{N+1}} \mathbb{1}(A(\sigma \cdot \vec{\boldsymbol{z}}) \geq A(\vec{\boldsymbol{z}}))\,, \tag{1.117}$$

where $\vec{\boldsymbol{z}} \equiv (z_1, \ldots, z_N, z^*)$. Using these explicit permutations, one can generalize the construction of conformal predictors to arbitrary randomization schemes, i.e. to subgroups of S_{N+1}. However, in general this will ruin the validity of the procedure.

?? FINISH (is this even relevant for this compendium) ??

At last a computationally efficient modification of the original CP algorithm is introduced. For most applications, especially those in machine learning and big data, the computational inefficiency of conformal predictors would make them hard to use. To overcome this issue *Papadopoulos et al.* introduced the following modification:

Construction 1.10.11 (Inductive CP). Consider a data set $\mathcal{D} \subset Z$ and a nonconformity measure A based on an underlying predictor. First, split $\mathcal{D}$ into a training set $\mathcal{T}$ and a calibration set $\mathcal{C}$. Using $\mathcal{T}$, train the underlying predictor of A. Then, for every point $z \equiv (x, y) \in \mathcal{C}$, construct the nonconformity score $\mu_z := A(z)$. As before, for every new element $z^* \in Z$ the conformal p-value is defined as the fraction of elements in $\mathcal{C}$ for which the nonconformity measure is larger than the one for z^*:

$$p^* := \frac{|\{z \in \mathcal{C} \cap \{z^*\} \mid \mu_z \geq \mu_{z^*}\}|}{|\mathcal{C}| + 1}. \tag{1.118}$$

As in the original CP algorithm, a new observation z^* is included in the prediction region if and only if $p^* > \alpha$:

$$\Gamma^{\alpha}(x) := \{y \in Y \mid p(x, y) > \alpha\}. \tag{1.119}$$

It should be clear that the underlying predictor only needs to be trained once using this scheme.

Remark 1.10.12 (Terminology). The name ‘‘inductive CP’’ stems from the fact that the general behaviour is deduced from a small subset of all observations. For this reason one sometimes calls the original algorithm a ‘‘transductive’’ method.

Property 1.10.13 (Validity). Although the above ICP algorithm is already computationally much more efficient than its transductive counterpart, one can go even further. However, in this case one needs to pay attention in order not to ruin the validity. To use Equation (1.118) one does not have to retrain the model every time, but one still needs to reevaluate the nonconformity score for possible $y^* \in Y$. It would be even better if one could extract the boundaries of the prediction region straight from the calibration data.

If the data is exchangeable (and the resulting nonconformitys score are too) it is not hard to see that the rank of any new nonconformity score among the calibration scores $A(\mathcal{C}) := \{\mu_z \in \mathbb{R} \mid z \in \mathcal{C}\}$ is uniformly sampled from $\{1, \ldots, |\mathcal{C}| + 1\}$. So, given a quantile level β, the probability of finding a new nonconformity score smaller than or equal to the β-quantile of $A(\mathcal{C})$ is

$$P\Big(\mu_{z^*} \leq q_\beta(A(\mathcal{C}))\Big) = \frac{\lceil \beta|\mathcal{C}| \rceil}{|\mathcal{C}| + 1}. \tag{1.120}$$

If one constructs a prediction region by including all points $(x^*, y^*) \in Z$ such that the associated nonconformity score is smaller than the $(1-\alpha)$-quantile, one immediately obtains

$$P(y^* \in \Gamma^\alpha(x^*, \mathcal{C})) = \frac{\lceil (1-\alpha)|\mathcal{C}| \rceil}{|\mathcal{C}| + 1}. \tag{1.121}$$

However, this is where the TCP and ICP algorithms differ. When using Equation (1.113) for the transductive algorithm, the above probabilities would contain a factor $|\mathcal{C}| + 1$ in both the numerator and the denominator, so the coverage condition is satisfied. However, for this quantile-based reformulation of the inductive algorithm, a minimal modification also leads to validity:

$$\Gamma^\alpha(x) \longrightarrow \Gamma^\alpha(x) := \left\{y \in Y \mid A(x, y) \leq q_{(1+1/|\mathcal{C}|)(1-\alpha)}(A(\mathcal{C}))\right\}, \tag{1.122}$$

i.e. the empirical quantiles are replaced by ''inflated'' quantiles. When using Equations (1.118) and (1.119) to determine the ICP region, one is essentially including all points (x^*, y^*) such that the nonconformity score is smaller than or equal to the $(1-\alpha)$-quantile of the enhanced calibration curve $A(\mathcal{C}) \cup \{A(x^*, y^*)\}$. It is not hard to show that this quantile is equivalent to the inflated quantile of the ordinary calibration curve.

The smoothing of Property 1.10.10 can also be implied in this case. If one does not use a smoothened conformal predictor but assumes that all calibration scores are distinct, the following property is obtained:

$$P(y^* \in \Gamma^\alpha(x, \mathcal{C})) \geq 1 - \alpha + \frac{1}{|\mathcal{C}|}. \tag{1.123}$$

So, in the limit of large calibration sets, the exact validity is recovered.

Remark 1.10.14. One does have to pay attention when interpreting the above statement. The validity property holds in probability with respect to both the new instance (x^*, y^*) and the calibration set $\mathcal{C}$. However, this does not not mean that for a fixed calibration set $\mathcal{C}$ the error fraction

$$\frac{|\{1 \leq i \leq k : y_i \in \Gamma^\alpha(x_i, \mathcal{C})\}|}{k} \tag{1.124}$$

is bounded by α for $k \longrightarrow \infty$. Because the events are not independent, the error can be much larger than α.

The above offline ICP algorithm can be generalized to an online algorithm:

Construction 1.10.15 (Online ICP). Consider an increasing sequence of positive integers $(m_n)_{n\in\mathbb{N}_0}$ of ''update thresholds''. The prediction region $\Gamma^\alpha(\mathcal{S})$ for the data sample $\mathcal{S} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is defined as follows:

- If $n \leq m_1$, use a fixed conformal predictor to construct $\Gamma^\alpha(\mathcal{S})$.
- If $m_k < n \leq m_{k+1}$, construct $\Gamma^\alpha(\mathcal{S})$ as follows:

$$\Gamma^\alpha(\mathcal{S}) := \left\{ y \in Y \;\middle|\; \frac{|\{m_k < j \leq n \mid \mu_{z_j} \geq \mu_{z_n}\}|}{n - m_k} > \varepsilon \right\}, \tag{1.125}$$

  where

$$\mu_{z_j} := A(\mathcal{B}(z_1, \ldots, z_{m_k}), (x_j, y_j)),$$
$$\mu_{z_n} := A(\mathcal{B}(z_1, \ldots, z_{m_k}), (x_n, y)).$$

It is clear that for $k \ll |\mathcal{C}|$ the offline ICP algorithm approximates the online version. The major benefit of the online algorithm is that one does not need to use the inflated quantiles to obtain valid prediction regions, because the calibration set grows every time new data is observed and, hence, the finite-size fluctuations get supressed.

?? COMPLETE ??

1.10.3 Classifier calibration

A specific instance of regression problems are classification tasks, where the one tries to model a function $f : X \to Y$ with Y finite (or discrete). For the case of probabilistic classifiers, the notion of validity 1.10.1 admits the following formulation:

Definition 1.10.16 (Calibration). A probabilistic (multiclass) classifier $\hat{P}(\cdot \mid \cdot) : Y \times X \to [0, 1]$ is said to be (well-)calibrated if

$$\Pr\Big(y \;\Big|\; \hat{P}(y \mid x) = p\Big) = p \tag{1.126}$$

for all $p \in [0, 1]$. Here, the confidence level $1 - \alpha$ is the output of the classifier $\hat{P}$, i.e. instead of asking for a region satisfying a given confidence level, the model yields the confidence level required to include a given class. A possible way to visually investigate the calibration of a probabilistic classifier is to draw **calibration plots** or **reliability diagrams** for all classes, where for every class label y and for a suitable partition $\mathcal{P} := \{0 = p_1, p_2, \ldots, p_n = 1\}$ of the interval $[0, 1]$ the i^{th} point is the proportion of instances for which y was predicted with probability between p_i and p_{i+1}. For a calibrated model, these points should lie on the diagonal.

Consider now the case of binary classification: $X \to \{0, 1\}$. Here one can easily apply the conformal prediction framework from the previous section. To this end, define the following nonconformity measure:

$$A_{\text{binary}}(\mathcal{B}, (x, y)) := 1 - \hat{P}(y \mid x; \mathcal{B}), \tag{1.127}$$

where the classifier is possibly estimated on $\mathcal{B}$. At significance α, the model predicts all classes such that the conformal p-value is greater than α. As before, one can adopt an inductive framework and use the $(1 - \alpha)$-quantile of the calibration scores as a threshold.

In fact, for classification tasks, where the codomain is finite (or at least discrete), it makes sense to slightly change the approach. Instead of fixing the significance level α, one could (or should) consider some specific values:

- **Confidence**: $\sup\{1-\alpha \mid |\Gamma^\alpha(x)| \leq 1\}$, and
- **Credibility**: $\inf\{\alpha \mid |\Gamma^\alpha(x)| = 0\}$.

The former is the highest probability with which the model can predict a single class. If one wants a higher probability, more than one class has to be considered. The latter gives a measure of how ''credible'' the predictions are. The smaller this number, i.e. the greater the confidence with which the model predicts a vacuously false result (every sample necessarily belongs to a class), the less reliable the model is for a given sample.

Various othe approaches exist to calibrate existing probabilistic classifiers. In practice it has been observed that many models show sigmoidal distortions in their calibration plots. A possible approach is then to modify the final sigmoidal layer in these models (or add an extra sigmoidal layer if the model did not use one). This gives rise to **Platt scaling** and **temperature scaling**. For the former one takes the output $\hat{P}$ and fits a sigmoidal layer of the form

$$\hat{P}_{\text{Platt}}(y \mid x; A, B) = \frac{1}{1+\exp(-A\hat{P}(y \mid x) - B)}. \tag{1.128}$$

For the latter one modifies the existing logits as follows for some parameter $T > 0$:

$$\hat{P}_{\text{temp}}(y \mid x; T) = \frac{1}{1+\exp(-\hat{z}(y \mid x)/T)}, \tag{1.129}$$

where $\hat{z}$ represents the logits of the estimator $\hat{P}$.

Remark 1.10.17 (Accuracy). Because temperature scaling does not change the maximum of the softmax function, the eventual predictions do not change.

1.10.4 Normalizing flows

One of the main problems for obtaining valid uncertainty estimates is that the underlying distribution of a given data set is often unknown. Except for the conformal prediction framework, all other methods make some assumptions about the underlying data generating process (even CP makes the exchangeability assumption). One way around this problem is by first transforming the data such that it is sampled according to a well-known distribution.

1.10.5 Conditionality

The above sections considered confidence predictors that were valid in a global sense, i.e. averaged over the whole instance space $X \times Y$. However, as in ordinary probability theory, in many settings it makes more sense to consider conditional statements:

$$P(y \in \Gamma^\alpha(x) \mid \kappa(x,y)) \geq 1-\alpha\,, \tag{1.130}$$

where the function $\kappa : X \times Y \to K$ represents the conditioning statement (the set K can be any set). In the conformal prediction literature it is often called a **taxonomy function**.

In practice, the label set K is often finite and discrete, i.e. one considers a finite subdivision of $X \times Y$. The simplest solution to obtain conditional validity in this case is to apply any known algorithm to every conditional class individually. In the case of conformal prediction this gives rise to the notion of **Mondrian conformal predictors**.

In a perfect world, the ultimate goal should be to have exact objectwise validity:

$$P(y \in \Gamma^\alpha(x) \mid x) \geq 1-\alpha. \tag{1.131}$$

However, in general, one can show that this cannot be attained:

Property 1.10.18 (No-go theorem). Let X be a separable metric space equipped with its canonical Borel σ-algebra and consider a confidence predictor Γ^α at conditional significance level $\alpha \in [0,1]$. For every probability distribution P on $X \times Y$ and P_X-almost all non-atoms $x \in X$ one has

$$\Pr(\lambda(\Gamma^\alpha(x)) = +\infty) \geq 1 - \alpha \tag{1.132}$$

and

$$\Pr(\mathrm{co}(\Gamma^\alpha(x)) = \mathbb{R}) \geq 1 - 2\alpha\,, \tag{1.133}$$

where λ and co denote the Lebesgue measure and convex hull, respectively.

?? FINISH ??

1.10.6 Distribution-shift

An important problem in the field of data science, in particular that of machine learning, is the change of the data generating process. Consider a classic train-test routine. If the model was trained on a data set sampled from the distribution P_0, but applied to a data set sampled from the distribution P_1, there is no reason the expect that the model will still give reasonable results. Furthermore, even if the crude point predictions are still somewhat sensible, this does not mean that their theoretical properties, such as the validity of confidence regions, is preserved.

A first step to resolve this problem is the detection of a possible distribution shift. Conformal prediction, as introduced in the previous section, can be used to detect such shifts in an online fashion. The main idea of the algorithm is to construct a martingale **??** from the p-values produced by a conformal predictor. For every (reasonable) martingale $(X_n)_{n\in\mathbb{N}}$ the Doob-Ville inequality **??** says that the growth is bounded. However, if the sequence $(X_n)_{n\in\mathbb{N}}$ is constructed in such a way that the martingale property is lost once the data distribution changes, the Doob-Ville inequality can be violated and the crossing of a given threshold can be regarded as evidence for this distribution shift [4].

The general expression of the martingale will be of the form

$$X_i \equiv \prod_{k=0}^{i} f_k(p_k)\,, \tag{1.134}$$

where p_k is the p-value of the k^{th} data point as produced by some conformal predictor. The functions f_k are called the **betting functions**. For $(X_n)_{n\in\mathbb{N}}$ to be a martingale with respect to the natural filtration of the p_k's, the betting functions should satisfy

$$\int_0^1 f_k(p)\,dp = 1. \tag{1.135}$$

Method 1.10.19 (Power martingale). For every constant $\varepsilon \in [0,1]$ one defines the power martingale as

$$X_i^\varepsilon := \prod_{k=0}^{i} \varepsilon p_k^{\varepsilon-1}. \tag{1.136}$$

One can also construct a **simple mixture** martingale by integrating over ε:

$$X_i := \int_0^1 X_i^\varepsilon\,d\varepsilon. \tag{1.137}$$

Because $\varepsilon - 1 \leq 0$, the above martingales will start to become very large if the conformal predictor produces small p-values, i.e. when unlikely values are observed.

However, not all distribution shifts will give rise to such behaviour. For example, it is possible that, although the p-values are not distributed uniformly anymore (since the data is not exchangeable), they are concentrated in the upper half of the unit interval and, hence, do not let the ''martingale'' grow strongly. For this reason it is convenient to construct betting functions that take into account the distribution of the p-values.

Method 1.10.20 (Plug-in martingale). Let $\hat{\rho}_i(p)$ denote an estimate of the probability density constructed using the p-values $\{p_1, \dots, p_i\}$. The plug-in martingale is defined by the betting functions

$$f_i := \hat{\rho}_{i-1}. \tag{1.138}$$

Now, if the empirical distribution functions of the p-values converge weakly **??** to an absolutely continuous distribution and $\log(\hat{\rho}_i(p)) \longrightarrow \log(\rho(p))$ uniformly, where ρ is the limit density of the empirical distributions, it can be shown that plug-in martingale grows quicker than the martingale associated to any other (continuous) betting function.

A different approach is to use p-values that are constructed directly from disitributional data. For inspiration, consider the following property:

Property 1.10.21. Consider a data sample $(x_n)_{n\in\mathbb{N}}$ and let f, g be two possible probability densities describing this sample. If f is the true density, the likelihood process

$$X_i := \prod_{k=0}^{i} \frac{g(x_i)}{f(x_i)} \tag{1.139}$$

is a martingale.

So likelihood ratios already give rise to test martingales. However, this martingale cannot be used to check for distribution shifts, since it is only a martingale when the true density is used.

Example 1.10.22 (Likelihood nonconformity). For every two probability densities f, g one can define a nonconformity measure as follows:

$$A_{\mathrm{NP}}(\mathcal{B}, z) := \log f(z) - \log g(z)\,, \tag{1.140}$$

i.e. the log-likelihood ratio.[14]

By using this nonconformity measure in combination with the above plug-in approach, one can obtain a likelihood-based changepoint detection algorithm. If the initial distribution is not known, it can be estimated based on the bag $\mathcal{B}$.

[14]The subscript refers to the Neyman-Pearson lemma **??** for which this function gives the (logarithm) of the test statistic.

List of Symbols

The following symbols are used throughout the summary:

Abbreviations

AIC	Akaike information criterion
ARMA	autoregressive moving-average model
BCH	Baker-Campbell-Hausdorff
CCR	canonical commutation relation
CDF	cumulative distribution function
CFT	conformal field theory
CIS	completely integrable system
CP	completely positive
CPTP	completely positive, trace-preserving
CR	Cauchy-Riemann
DGA	differential graded algebra
DGCA	differential graded-commutative algebra
EPR	Einstein-Podolsky-Rosen
ETCS	Elementary Theory of the Category of Sets
FWHM	full width at half maximum
GA	geometric algebra
GHZ	Greenberger-Horne-Zeilinger
GNS	Gel'fand-Naimark-Segal
HoTT	Homotopy Type Theory
KKT	Karush-Kuhn-Tucker
LIVF	left-invariant vector field
MPO	matrix product operator
MPS	matrix product state
MTC	modular tensor category
NDR	neighbourhood deformation retract
OPE	operator product expansion
OZI	Okubo-Zweig-Iizuka
PAC	probably approximately correct
PL manifold	piecewise-linear manifold
PVM	projection-valued measure

RKHS	reproducing kernel Hilbert space
SVM	support-vector machine
TQFT	topological quantum field theory
VIF	variance inflation factor
ZFC	Zermelo-Frenkel set theory with the axiom of choice
TVS	topological vector space

Operations

Ad_g	adjoint representation of a Lie group G
ad_X	adjoint representation of a Lie algebra $\mathfrak{g}$
arg	argument of a complex number
$\Box$	d'Alembert operator
$\deg(f)$	degree of the polynomial f
e	identity element of a group
$\Gamma(E)$	set of global sections of a fibre bundle E
Im	imaginary part of a complex number
$\mathrm{Ind}_f(z)$	index of a point $z \in \mathbb{C}$ with respect to a function f
$\hookrightarrow$	injective function
$\cong$	is isomorphic to
Par_t^γ	parallel transport map with respect to the curve γ
Re	real part of a complex number
Res	residue of a complex function
$\twoheadrightarrow$	surjective function
$\{\cdot,\cdot\}$	Poisson bracket
∂X	boundary of a topological space X
$\overline{X}$	closure of a topological space X
$X^\circ, \overset{\circ}{X}$	interior of a topological space X
$\sphericalangle(\cdot,\cdot)$	angle between two vectors
$X \times Y$	cartesian product of the sets X and Y
$X + Y$	sum of the vector spaces X and Y
$X \oplus Y$	direct sum of the vector spaces X and Y
$V \otimes W$	tensor product of the vector spaces V and W
$\mathbb{1}_X$	identity morphism on the object X
$\approx$	is approximately equal to
$\hookrightarrow$	is included in
$\cong$	is isomorphic to
$\mapsto$	mapsto

Collections

Ab	category of Abelian groups
$\mathrm{Aut}(X)$	automorphism group of an object X
$\mathcal{B}_0(V,W)$	space of compact bounded operators between the Banach spaces V and W

$\mathcal{B}(V, W)$	space of bounded linear maps from the space V to the space W
CartSp	the category of Euclidean spaces and "suitable" homomorphisms (e.g. linear maps, smooth maps, ...)
$C_\bullet$	chain complex
$\mathbf{Ch}(\mathbf{A})$	category of chain complexes with objects in the additive category $\mathbf{A}$
$\mathbf{C}^\infty$	category of smooth spaces
$C_p^\infty(M)$	ring of smooth functions $f : M \to \mathbb{R}$ on a neighbourhood of $p \in M$
$C^\omega(V)$	the set of all analytic functions defined on the set V
$\mathrm{Conf}(M)$	conformal group of (pseudo-)Riemannian manifold M
$C(X, Y)$	set of continuous functions between two topological spaces X and Y
$\mathbf{C}^\infty\mathbf{Ring}, \mathbf{C}^\infty\mathbf{Alg}$	category of smooth algebras
Diff	category of smooth manifolds
DiffSp	category of diffeological spaces and smooth maps
D^n	standard n-disk
$\mathrm{dom}(f)$	domain of a function f
$\mathrm{End}(X)$	endomorphism monoid of a an object X
$\mathcal{E}\mathrm{nd}$	endomorphism operad
$\mathbf{FormalCartSp}_{\mathbf{diff}}$	category of infinitesimally thickened Euclidean spaces
$\mathrm{GL}(V)$	general linear group, the group of automorphisms of a vector space V
$\mathrm{GL}(n, K)$	general linear group: the group of all invertible $n \times n$-matrices over the field K
Grp	category of groups and group homomorphisms
Grpd	category of groupoids
$\mathrm{Hol}_p(\omega)$	holonomy group at the point p with respect to the principal connection ω
$\mathrm{Hom}_{\mathbf{C}}(V, W)$	set of homomorphisms from an object V to an object W in a category $\mathbf{C}$
hTop	homotopy category
$\mathrm{im}(f)$	image of a function f
$K^0(X)$	K-theory over a (compact Hausdorff) space X
Kan	category of Kan complexes
$\mathcal{K}_n(A, v)$	Krylov subspace of dimension n generated by the matrix A and the vector v
L^1	space of integrable functions
Law	category of Lawvere theories
Lie	category of Lie groups
$\mathfrak{Lie}$	category of Lie algebras
$\mathfrak{X}^L$	space of left-invariant vector fields on a Lie group
LX	free loop space on X
$\mathbf{Man}^p$	category of C^p-manifolds
Meas	category of measure spaces and measure-preserving functions
$N\mathbf{C}$	the simplicial nerve of a small category $\mathbf{C}$
$\mathbf{Open}(X)$	category of open subsets of a topological space X
$\mathrm{O}(n, K)$	group of $n \times n$ orthogonal matrices over a field K
$P(S), 2^S$	power set of S

$\mathrm{Pin}(V)$	pin group of the Clifford algebra $C\ell(V, Q)$
$\mathbf{Psh}(\mathbf{C}), \widehat{\mathbf{C}}$	category of presheaves on a (small) category $\mathbf{C}$
$\mathbf{Sh}(X)$	category of sheaves on a topological space X
$\mathbf{Sh}(\mathbf{C}, J)$	category of J-sheaves on a site $(\mathbf{C}, J)$
$\mathbf{\Delta}$	simplex category
$\mathrm{SL}_n(K)$	special linear group: group of all invertible n-dimensional matrices with unit determinant over the field K
S^n	standard n-sphere
$S^n(V)$	space of symmetric rank n tensors over a vector space V
$W^{m,p}(U)$	the Sobolov space in L^p of order m
$\mathbf{Span}(\mathbf{C})$	span category over $\mathbf{C}$
$\mathrm{Spec}(R)$	spectrum of a commutative ring R
$\mathrm{supp}(f)$	support of a function f
$\mathrm{Syl}_p(G)$	set of Sylow p-subgroups of a finite group G
S_n	symmetric group of degree n
$\mathrm{Sym}(X)$	symmetric group on the set X
$\mathrm{Sp}(n, K)$	group of matrices preserving a canonical symplectic form over the field K
$\mathrm{Sp}(n)$	compact symplectic group
$\mathrm{TL}_n(\delta)$	Temperley-Lieb algebra with $n-1$ generators and parameter δ.
T^n	standard n-torus (the n-fold Cartesian product of S^1)
Top	category of topological spaces
Topos	the 2-category of (elementary) topoi and geometric morphisms
$U(\mathfrak{g})$	universal enveloping algebra of a Lie algebra $\mathfrak{g}$
$\mathrm{U}(n, K)$	group of $n \times n$ unitary matrices over a field K
$\mathbf{Vect}(X)$	category of vector bundles over a manifold X
$\mathbf{Vect}_K$	category of vector spaces and linear maps over a field K
Y^X	set of functions from a set X to a set Y
$\emptyset$	empty set
$\pi_n(X, x_0)$	n^{th} homotopy space over X with basepoint x_0
$[a, b]$	closed interval
$]a, b[$	open interval
$\Lambda^n(V)$	space of antisymmetric rank n tensors over a vector space V
ΩX	(based) loop space on X
$\Omega^k(M)$	$C^\infty(M)$-module of differential k-forms on the manifold M
$\rho(A)$	resolvent set of a bounded linear operator A
$\mathfrak{X}(M)$	$C^\infty(M)$-module of vector fields on the manifold M
Units	
C	coulomb
T	tesla

Bibliography

[1] Jonathan R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, 1994.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705--1749, 2005.

[3] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371-–421, 2008.

[4] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[5] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281--307, 2010.

[6] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Publishing Company, Incorporated, 2016.

[7] Niclas Sandgren and Petre Stoica. On moving average parameter estimation. Technical Report 2006-022, Department of Information Technology, Uppsala University, 2006.

[8] Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732--749. PMLR, 2018.

[9] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[10] Tetsuji Miwa, Michio Jimbo, Michio Jimbo, and E Date. *Solitons: Differential Equations, Symmetries and Infinite-dimensional Algebras*, volume 135. Cambridge University Press, 2000.

[11] Vladimir I. Arnol'd. *Mathematical Methods of Classical Mechanics*, volume 60. Springer Science & Business Media, 2013.

[12] Edwin J. Beggs and Shahn Majid. *Quantum Riemannian Geometry*. Springer, 2020.

[13] Marc Henneaux and Claudio Teitelboim. *Quantization of Gauge Systems*. Princeton university press, 1992.

[14] Mark Hovey. *Model Categories*. Number 63. American Mathematical Soc., 2007.

[15] Gregory M. Kelly. *Basic Concepts of Enriched Category Theory*, volume 64. CUP Archive, 1982.

[16] Mukund Rangamani and Tadashi Takayanagi. *Holographic Entanglement Entropy.* Springer, 2017.

[17] Saunders Mac Lane. *Categories for the Working Mathematician*, volume 5. Springer Science & Business Media, 2013.

[18] Peter T. Johnstone. *Topos Theory.* Dover Publications, 2014.

[19] Charles W. Misner, Kip S. Thorne, and John A. Wheeler. *Gravitation.* Princeton University Press, 2017.

[20] Carlo Rovelli and Francesca Vidotto. *Covariant Loop Quantum Gravity: An Elementary Introduction to Quantum Gravity and Spinfoam Theory.* Cambridge University Press, 2014.

[21] Richard W. Sharpe. *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program*, volume 166. Springer Science & Business Media, 2000.

[22] John C. Baez, Irving E. Segal, and Zhengfang Zhou. *Introduction to Algebraic and Constructive Quantum Field Theory.* Princeton University Press, 2014.

[23] Raoul Bott and Loring W. Tu. *Differential Forms in Algebraic Topology.* Graduate Texts in Mathematics. Springer New York, 1995.

[24] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics.* `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.

[25] Bruce Blackadar. *Operator Algebras: Theory of C∗-Algebras and von Neumann Algebras.* Springer, 2013.

[26] Marek Capinski and Peter E. Kopp. *Measure, Integral and Probability.* Springer Science & Business Media, 2013.

[27] Georgiev Svetlin. *Theory of Distributions.* Springer, 2015.

[28] Gerd Rudolph and Matthias Schmidt. *Differential Geometry and Mathematical Physics: Part II. Fibre Bundles, Topology and Gauge Fields.* Springer, 2017.

[29] Martin Schottenloher. *A Mathematical Introduction to Conformal Field Theory*, volume 759. 2008.

[30] Dusa McDuff and Deitmar Salamon. *Introduction to Symplectic Topology.* Oxford Graduate Texts in Mathematics. Oxford University Press, 2017.

[31] John C. Baez and Peter May. *Towards Higher Categories*, volume 152 of *IMA Volumes in Mathematics and its Applications.* Springer, 2009.

[32] Mikhail. M. Kapranov and Vladimir A. Voevodsky. *2-categories and Zamolodchikov Tetrahedra Equations*, volume 56 of *Proc. Sympos. Pure Math.* Amer. Math. Soc., Providence, RI, 1994.

[33] Geoffrey Compère. *Advanced Lectures on General Relativity*, volume 952. Springer, 2019.

[34] Pavel Etingof, Shlomo Gelaki, Dmitri Nikshych, and Victor Ostrik. *Tensor Categories*, volume 205. American Mathematical Soc., 2016.

[35] David Mumford. *The Red Book of Varieties and Schemes: Includes the Michigan Lectures (1974) on Curves and Their Jacobians*, volume 1358. Springer Science & Business Media, 1999.

[36] Charles A. Weibel. *An Introduction to Homological Algebra*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1994.

[37] Peter J. Hilton and Urs Stammbach. *A Course in Homological Algebra*. Springer.

[38] Jean-Luc Brylinski. *Loop Spaces, Characteristic Classes and Geometric Quantization*. Birkhauser.

[39] Antoine Van Proeyen and Daniel Freedman. *Supergravity*. Cambridge University Press.

[40] William S. Massey. *A Basic Course in Algebraic Topology*. Springer.

[41] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press.

[42] Nadir Jeevanjee. *An Introduction to Tensors and Group Theory for Physicists*. Birkhauser.

[43] Yvonne Choquet-Bruhat, Cecile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds and Physics, Part 1: Basics*. North-Holland.

[44] Yvonne Choquet-Bruhat and Cecile DeWitt-Morette. *Analysis, Manifolds and Physics, Part 2*. North-Holland.

[45] Herbet Goldstein, John L. Safko, and Charles P. Poole. *Classical Mechanics*. Pearson.

[46] Franco Cardin. *Elementary Symplectic Topology and Mechanics*. Springer.

[47] Walter Greiner and Joachim Reinhardt. *Field Quantization*. Springer.

[48] Walter Greiner. *Quantum Mechanics*. Springer.

[49] B. H. Bransden and Charles J. Joachain. *Quantum Mechanics*. Prentice Hall.

[50] Heydar Radjavi and Peter Rosenthal. *Invariant Subspaces*. Dover Publications.

[51] Max Karoubi. *K-Theory: An Introduction*. Springer.

[52] Damien Calaque and Thomas Strobl. *Mathematical Aspects of Quantum Field Theories*. Springer, 2015.

[53] Ivan Kolar, Peter W. Michor, and Jan Slovak. *Normal Operations in Differential Geometry*. Springer.

[54] Stephen B. Sontz. *Principal Bundles: The Classical Case*. Springer.

[55] Stephen B. Sontz. *Principal Bundles: The Quantum Case*. Springer.

[56] William Fulton and Joe Harris. *Representation Theory: A First Course*. Springer.

[57] Peter Petersen. *Riemannian Geometry*. Springer.

[58] Charles Nash and Siddharta Sen. *Topology and Geometry for Physicists*. Dover Publications.

[59] Ian M. Anderson. *The Variational Bicomplex*.

[60] Joel Robbin and Dietmar Salamon. The maslov index for paths. *Topology*, 32(4):827--844, 1993.

[61] Nima Moshayedi. 4-manifold topology, donaldson-witten theory, floer homology and higher gauge theory methods in the BV-BFV formalism. 2021.

[62] Edward Witten. Global anomalies in string theory. In *Symposium on Anomalies, Geometry, Topology*, 6 1985.

[63] F.A. Berezin and M.S. Marinov.

[64] Paul A. M. Dirac. Generalized Hamiltonian dynamics. *Canadian Journal of Mathematics*, 2:129–148, 1950.

[65] Angelo Vistoli. Notes on Grothendieck topologies, fibered categories and descent theory. *arXiv:math/0412512*, 2004.

[66] Emily Riehl and Dominic Verity. The theory and practice of Reedy categories. *Theory and Applications of Categories*, 29, 2013.

[67] Emily Riehl. Homotopical categories: From model categories to $(\infty, 1)$-categories. 2019. arXiv:1904.00886.

[68] Floris Takens. A global version of the inverse problem of the calculus of variations. *Journal of Differential Geometry*, 14(4):543--562, 1979.

[69] John Baez and Alexander Hoffnung. Convenient categories of smooth spaces. *Transactions of the American Mathematical Society*, 363(11):5789--5825, 2011.

[70] John C. Baez and Alissa S. Crans. Higher-dimensional algebra vi: Lie 2-algebras. 2003. arXiv:math/0307263.

[71] John C. Baez and Aaron D. Lauda. Higher-dimensional algebra v: 2-groups. 2003. arXiv:math/0307200.

[72] Edward Witten. Supersymmetry and Morse theory. *J. Diff. Geom*, 17(4):661--692, 1982.

[73] Urs Schreiber. *From Loop Space Mechanics to Nonabelian Strings*. PhD thesis, 2005.

[74] John C. Baez and Urs Schreiber. Higher gauge theory. 2005. arXiv:math/0511710.

[75] Jade Master. Why is homology so powerful? 2020. arXiv:2001.00314.

[76] Marcus Berg, Cécile DeWitt-Morette, Shangjr Gwo, and Eric Kramer. The Pin groups in physics: C, P and T. *Reviews in Mathematical Physics*, 13(08):953--1034, 2001.

[77] Richard Palais. The symmetries of solitons. *Bulletin of the American Mathematical Society*, 34(4):339--403, 1997.

[78] Michael F. Atiyah. Topological quantum field theory. *Publications Mathématiques de l'IHÉS*, 68:175--186, 1988.

[79] Jens Eisert, Christoph Simon, and Martin B Plenio. On the quantification of entanglement in infinite-dimensional quantum systems. *Journal of Physics A: Mathematical and General*, 35(17):3911--3923, 2002.

[80] Benoît Tuybens. Entanglement entropy of gauge theories. 2017.

[81] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338--353, 1965.

[82] John C. Baez, Alexander E. Hoffnung, and Christopher Rogers. Categorified symplectic geometry and the classical string. *Communications in Mathematical Physics*, 293:701--725, 2010.

[83] Charles Rezk. A model for the homotopy theory of homotopy theory. *Transactions of the American Mathematical Society*, 353(3):973--1007, 2001.

[84] Peter May. A note on the splitting principle. *Topology and Its Applications*, 153(4):605--609, 2005.

[85] Irina Markina. Group of diffeomorphisms of the unit circle as a principal $U(1)$-bundle.

[86] Sjoerd E. Crans. Localizations of transfors. 1998.

[87] Tom Leinster. Basic bicategories. 1998. arXiv:math/9810017.

[88] Alexander E. Hoffnung. Spans in 2-categories: A monoidal tricategory. 2011. arXiv:1112.0560.

[89] Eugenia Cheng and Nick Gurski. The periodic table of n-categories for low dimensions ii: Degenerate tricategories. 2007. arXiv:0706.2307.

[90] Mehmet B. Şahinoğlu, Dominic J. Williamson, Nick Bultinck, Michael Mariën, Jutho Haegeman, Norbert Schuch, and Frank Verstraete. Characterizing topological order with matrix product operators. 2014. arXiv:1409.2150.

[91] Dominic J. Williamson, Nick Bultinck, Michael Mariën, Mehmet B. Şahinoğlu, Jutho Haegeman, and Frank Verstraete. Matrix product operators for symmetry-protected topological phases: Gauging and edge theories. *Phys. Rev. B*, 94, 2016.

[92] Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91, 2003.

[93] Aaron D. Lauda and Hendryk Pfeiffer. Open–closed strings: Two-dimensional extended TQFTs and Frobenius algebras. *Topology and its Applications*, 155(7):623--666, 2008.

[94] Domenico Fiorenza. An introduction to the Batalin-Vilkovisky formalism. 2004. arXiv:math/0402057v2.

[95] Stefan Cordes, Gregory Moore, and Sanjaye Ramgoolam. Lectures on 2d Yang-Mills theory, equivariant cohomology and topological field theories. arXiv:hep-th/9411210v2.

[96] Donald C. Ferguson. A theorem of Looman-Menchoff. `http://digitool.library.mcgill.ca/thesisfile111406.pdf`.

[97] Holger Lyre. Berry phase and quantum structure. arXiv:1408.6867.

[98] Florin Belgun. Gauge theory. `http://www.math.uni-hamburg.de/home/belgun/Gauge4.pdf`.

[99] Vladimir Itskov, Peter J. Olver, and Francis Valiquette. Lie completion of pseudogroups. *Transformation Groups*, 16:161--173, 2011.

[100] Richard Borcherds. Lie groups. `https://math.berkeley.edu/~reb/courses/261/`.

[101] Andrei Losev. From Berezin integral to Batalin-Vilkovisky formalism: A mathematical physicist's point of view. 2007.

[102] Edward Witten. Coadjoint orbits of the Virasoro group. *Comm. Math. Phys.*, 114(1):1--53, 1988.

[103] Sidney R. Coleman and Jeffrey E. Mandula. All possible symmetries of the S-matrix. *Phys. Rev.*, 159, 1967.

[104] Emily Riehl. Monoidal algebraic model structures. *Journal of Pure and Applied Algebra*, 217(6):1069--1104, 2013.

[105] Valter Moretti. Mathematical foundations of quantum mechanics: An advanced short course. *International Journal of Geometric Methods in Modern Physics*, 13, 2016.

[106] Antonio Michele Miti. Homotopy comomentum maps in multisymplectic geometry, 2021.

[107] John E. Roberts. Spontaneously broken gauge symmetries and superselection rules. 1974.

[108] Richard Sanders. Commutative spectral triples & the spectral reconstruction theorem.

[109] Jean Gallier. Clifford algebras, Clifford groups, and a generalization of the quaternions, 2008. arXiv:0805.0311.

[110] Bozhidar Z. Iliev. Normal frames for general connections on differentiable fibre bundles. arXiv:math/0405004.

[111] Piotr Stachura. Short and biased introduction to groupoids. arXiv:1311.3866.

[112] Fosco Loregian. Coend calculus. arXiv:1501.02503.

[113] Frederic Schuller. Lectures on the geometric anatomy of theoretical physics. `https://www.youtube.com/channel/UC6SaWe7xeOp31Vo8cQG1oXw`.

[114] Nima Amini. Infinite-dimensional Lie algebras. `https://people.kth.se/~namini/PartIIIEssay.pdf`.

[115] Peter Selinger. Lecture notes on lambda calculus.

[116] Nigel Hitchin. Lectures on special Lagrangian submanifolds. `https://arxiv.org/abs/math/9907034v1`, 1999.

[117] Olivia Caramello. Lectures on topos theory at the university of Insubria. `https://www.oliviacaramello.com/Teaching/Teaching.htm`.

[118] Derek Sorensen. An introduction to characteristic classes. `http://derekhsorensen.com/docs/sorensen-characteristic-classes.pdf`, 2017.

[119] Arun Debray. Characteristic classes. `https://web.ma.utexas.edu/users/a.debray/lecture_notes/u17_characteristic_classes.pdf`.

[120] Pascal Lambrechts.

[121] Chris Tiee. Contravariance, covariance, densities, and all that: An informal discussion on tensor calculus. `https://ccom.ucsd.edu/~ctiee/notes/tensors.pdf`, 2006.

[122] Emily Riehl. Homotopy (limits and) colimits. `http://www.math.jhu.edu/~eriehl/hocolimits.pdf`.

[123] Andreas Gathmann. Algebraic geometry. `https://www.mathematik.uni-kl.de/~gathmann/class/alggeom-2019/alggeom-2019.pdf`.

[124] Will J. Merry. Algebraic topology. `https://www.merry.io/algebraic-topology`.

[125] Stacks project. `https://stacks.math.columbia.edu/`.

[126] The nlab. `https://ncatlab.org/nlab`.

[127] Wikipedia. `https://www.wikipedia.org/`.

[128] Joost Nuiten. Cohomological quantization of local prequantum boundary field theory. Master's thesis, 2013.

Index