# Numerical Analysis in DROP

**v6.38** 20 June 2024

# Introduction

## Framework Glossary

1. <u>Hyperspace Search</u>: Hyperspace search is a search to determine whether the entity is inside the zone of a range, e.g., bracketing search.

2. <u>Hyperpoint Search</u>: Hyperpoint searches are hard searches that search for an exact specific point (to within an appropriately established tolerance).

3. <u>Iterate Nodes</u>: This is the set of the traveled nodes (variate/Objective Function ordered pairs) that contain the trajectory traveled.

4. <u>Iteration Search Primitives</u>: The set of variate iteration routines that generate the subsequent iterate nodes.

5. <u>Compound iterator search scheme</u>: Search schemes where the primitive iteration routine to be invoked at each iteration are evaluated.

6. <u>RunMap</u>: Map that holds the program state at the end of each iteration, in the generic case, this is composed of the Wengert iterate node list, along with the corresponding root finder state.

7. <u>Cost of Primitive (cop)</u>: This is the cost of invocation of a single variate iterator primitive.

## Document Layout

1. Base Framework
2. Search Initialization
   a. Bracketing
   b. Objective Function Failure
   c. Bracketing Start Initialization

# Framework

1. The root search given an objective function and its goal is achieved by iteratively evolving the variate, and involves the following steps:

   - <u>Search initialization and root reachability determination</u>: Searched is kicked off by spawning a root variate iterator for the search initialization process (described in detail in the next section).
   - <u>Absolute Tolerance Determination</u>.
   - <u>Root Search Iteration</u>: The root is searched iteratively according to the following steps:
     1. The iterator progressively reduces the bracket width.
     2. Successive iteration occurs using either a single primitive (e.g., using the bisection primitive), or using a selector scheme that picks the primitives for each step (e.g., Brent's method).
     3. For Open Method, instead of 1 and 2, the routine drives towards convergence iteratively.
   - <u>Search Termination Detection</u>: The search termination occurs typically based on the following:
     - Proximity to the Objective Function Goal
     - Convergence on the variate
     - Exhaustion if the number of iterations

2. The flow behind these steps is illustrated in Figure 1.

3. The "Flow Control Variate" in root search is the "Objective Function Distance to Goal" Metric.

# Search Initialization

1. Broadly speaking, root finding approaches can be divided into a) those that bracket roots before they solve for them, and b) those that don't need to bracket, opting instead to pick a suitable starting point.

2. Depending upon the whether the search is a bracketing or an open method, the search initialization does one the following:
   - o Determine the root brackets for bracketing methods
   - o Locate root convergence zone for open methods

3. Initialization begins by a search for the starting zone. A suitable starting point/zone is determined where, by an appropriate choice for the iterator, you are expected to reach the fixed-point target within a sufficient degree of reliability. Very general-purpose heuristics often help determine the search start zone.

4. Both bracketing and open initializers are hyperspace searches, since they search for something "IN", not "AT".

## Bracketing

1. Bracketing is the process of localizing the fixed point to within a target zone with the least required number of Objective Function calculations. Steps are:
   - o Determine a valid bracketing search start
   - o Choose a suitable bracket expansion
   - o Limit attention to where the Objective Function is defined (more on this below).

2. Figure 2 shows the flow for the Bracketing routine.

3. Bracketing methods require that the initial search interval bracket the root (i.e. the function values at interval end points have opposite signs).

4. Bracketing traps the fixed point between two variate values, and uses the intermediate value theorem and the continuity of the Objective Function to guarantee the presence/existence of the fixed point between them.

5. Unless the objective function is discontinuous, bracketing methods guarantee convergence (although may not be within the specified iteration limit).

6. Typically, they do not require the objective function to be differentiable.

7. Bracketing iteration primitives' convergence is usually linear to super-linear.

8. Bracketing methods preserve bracketing throughout computation and allow user to specify which side of the convergence interval to select as the root.

9. It is also possible to force a side selection after a root has been found, for example, in sequential search, to find the next root.

10. Generic root bracketing methods that treat the objective function as a black box will be slower than targeted ones – so much so that they can constitute the bulk of the time for root search. This is because, to accommodate generic robustness coupled with root-pathology avoidance (oscillating bracket pairs etc.), these methods have to perform a full variate space sweep without any assumptions regarding the location of the roots (despite this most, bracketing algorithms cannot guarantee isolation of root intervals). For instance, naïve examination of the Objective Function's "sign-flips" alone can be misleading, especially if you bracket fixed-points with even numbered multiplicity within the brackets. Thus, some ways of analyzing the Black Box functions (or even the closed form Objective Functions) are needed to better target/customize the bracketing search (of course, parsimony in invoking the number of objective function calls is the main limitation).

11. <u>Soft Bracketing Zone</u>: One common scenario encountered during bracketing is the existence of a soft preferred bracketing zone, one edge of which serves as a "natural edge". In this case, the bracketing run needs to be positioned to be able to seek out starting variate inside soft zone in the direction AWAY from the natural edge.

12. The first step is to determine a valid bracketing search start. One advantage with univariate root finding is that objective function range validity maybe established

using an exhaustive variate scanner search without worrying about combinatorial explosion.

## Objective Function Failure

1. Objective Function may fail evaluation at the specified variate for the following reason:
    o Objective Function is not defined at the specified variate.
    o Objective Function evaluates to a complex number.
    o Objective Function evaluation produces NaN/Infinity/Under-flow/Over-flow errors.
    o In such situations, the following steps are used to steer the variate to a valid zone.
2. <u>Objective Function undefined at the Bracketing Candidate Variate</u>: If the Objective Function is undefined at the starting variate, the starting variate is expanded using the variate space scanner algorithm described above. If the objective Function like what is seen in Figure 3, a valid starting variate will eventually be encountered.
3. <u>Objective Function not defined at any of the Candidate Variates</u>: The risk is that the situation in Figure 4 may be encountered, where the variate space scanner iterator "jumps over" the range over which the objective function is defined. This could be because the objective function may have become complex. In this case, remember that an even power of the objective function also has the same roots as the objective function itself. Thus, solving for an even power of the objective function (like the square) – or even bracketing for it – may help.

## Bracketing Start Initialization

1. Figure 5 shows the flow behind a general-purpose bracket start locator.

2. Once the starting variate search is successful, and the objective function validity is range-bound, then use an algorithm like bisection to bracket the root (as shown in Figure 6 below).

3. However, if the objective function runs out of its validity range under the variate scanner scheme, the following steps need to be undertaken:
   o If the left bracketing candidate fails, bracketing is done towards the right using the last known working left-most bracketing candidate as the "left edge".
   o Likewise, if the right bracketing candidate fails, bracketing is done towards the left using the last known working right-most bracketing candidate as the "right edge".

4. The final step is to trim the variate zone. Using the variate space scanner algorithm, and the mapped variate/Objective Function evaluations, the tightest bracketing zones are extracted (Figure 7).

## Open Search Initialization

1. Non-bracketing methods use a suitable starting point to kick off the root search. As is obvious, the chosen starting point can be critical in determining the fate of the search. In particular, it should be within the zone of convergence of the fixed-point root to guarantee convergence. This means that specialized methods are necessary to determine zone of convergence.

2. When the objective function is differentiable, the non-bracketing root finder often may make use of that to achieve quadratic or higher speed of convergence. If the non-bracketing root finder cannot/does not use the objective function's differentiability, convergence ends up being linear to super-linear.

3. The typical steps for determining the open method starting variate are:
   o Find a variate that is proximal to the fixed point
   o Verify that it satisfies the convergence heuristic

4. Bracketing followed by a choice of an appropriate primitive variate (such as bisection/secant) satisfies both, thus could be a good starting point for open method searches like Newton's method.

5. Depending upon the structure of the Objective Function, in certain cases the chain rule can be invoked to ease the construction of the derivative – esp. in situations where the sensitivity to inputs are too high/low.

## Search/Bracketing Initializer Heuristic Customization

1. Specific Bracketing Control Parameters

2. Left/Right Soft Bracketing Start Hints: The other components may be used from the bracketing control parameters.

3. Mid Soft Bracketing Start Hint: The other components may be used from the bracketing control parameters.

4. Floor/Ceiling Hard Bracketing Edges: The other components may be used from the bracketing control parameters.

5. Left/Right Hard Search Boundaries: In this case, no bracketing is done – brackets are used to verify the roots, search then starts directly.

# Numerical Challenges in Search

1. Bit Cancellation

2. Ill-conditioning (e.g., see high order polynomial roots)

3. "domains of indeterminacy" – existence of sizeable intervals around which the objective function hovers near the target

4. Continuous, but abrupt changes (e.g., near-delta Gaussian objection function)

5. Under-flow/over-flow/round-off errors

6. root multiplicity (e.g., in the context of polynomial roots)

7. Typical solution is to transform the objective function to a better conditioned function – insight into the behavior of the objective can be used to devise targeted solutions.

# Variate Iteration

1.

$$v_{i+1} = I(v_i, \mathfrak{I}_i)$$

where $v_i$ is the ith variate and $\mathfrak{I}_i$ is the root finder state after the ith iteration.

2. <u>Iterate nodes as Wengert variables</u>: Unrolling the traveled iterate nodes during the forward accumulation process, as a Wengert list, is a proxy to the execution time, and may assist in targeted pre-accumulation and check-pointing.

3. <u>Cognition Techniques of Mathematical Functions</u>:
   o Wengert Variate Analysis => Collection of the Wengert variates helps build and consolidate the Objective Function behavior from the variate iterate Wengert nodes – to build a behavioral picture of the Objective Function.
   o Objective Function Neighborhood Behavior => With every Wengert variable, calculation of the set of forward sensitivities and the reverse Jacobians builds a local picture of the Objective Function without having to evaluate it.

4. <u>Check pointing</u>: Currently implemented using a roving variate/OF iterate node "RunMap"; this is also used to check circularity in the iteration process.

5. <u>Compound Iterator RunMap</u>: For compound iterations, the iteration circularity is determined the doublet $(v_i, \mathfrak{I}_i)$, so the Wengert RunMap is really a doublet Multi-Map.

6. <u>Hyperpoint univariate fixed-point search proximity criterion</u>: For hyperpoint checks, the search termination check needs to explicitly accommodate a "proximity to target" metric. This may not be then case for hyperspace checks.

7. <u>Regime crossover indicator</u>: On one side the crossover, the variate is within the fast convergence zone, so you may use faster Open techniques like the Newton's methods. On the other side, continue using the bracketing techniques.

a. Fast side of the crossover must be customizable (including other Halley's method variants); robust side should also be customizable (say False Position).

8. Crossover indicator determination: Need to develop targeted heuristics needed to determine the crossover indicator.

   o Entity that determines the crossover indicator may be determined from the relative variate shift change $\frac{x_{N+1}-x_N}{x_N-x_{N-1}}$ and the relative objective function change $\frac{y_{N+1}-y_N}{y_N-y_{N-1}}$.

9. Types of bracketing primitives:

   • Bracket narrower primitives (Bisection, false position), and interpolator primitives (Quadratic, Ridder).

   • Primitive's COP determinants: Expressed in terms of characteristic compute units.

     a. Number of objective function evaluation (generally expensive).

     b. Number of variate iterator steps needed.

     c. Number of objective function invocation per a given variate iteration step.

   • Bracket narrower primitives => Un-informed iteration primitives, low invocation cost (usually single objective function evaluation), but low search targeting quality, and high COP.

   • Interpolator primitives => Informed iteration primitives, higher invocation cost (multiple objective function evaluations, usually 2), better search targeting quality, and lower COP.

10. Pre-OF Evaluation Compound Heuristic: Heuristic compound variates are less informed, but rely heavily on heuristics to extract the subsequent iterator, i.e., pre-OF evaluation heuristics try to guide the evolution without invoking the expensive OF evaluations (e.g., Brent, Zheng).

11. OF Evaluation Compound Heuristic: These compound heuristics use the OF evaluations as part of the heuristics algorithm to establish the next variate => better informed

# Open Search Method: Newton's Method

1. Newton's method uses the objective function $f$ and its derivative $f'$ to iteratively evaluate the root.

2. Given a well-behaved function $f$ and its derivative $f'$ defined over real $x$, the algorithm starts with an initial guess of $x_0$ for the root.

3. First iteration yields

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

4. This is repeated in

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

till a value $x_n$ that is convergent enough is obtained.

5. If $\alpha$ is a simple root (root with multiplicity 1), and

$$\epsilon_n = x_n - \alpha$$

and

$$\epsilon_{n+1} = x_{n+1} - \alpha$$

respectively, then for sufficiently large $n$, the convergence is quadratic:

$$\epsilon_{n+1} \approx \frac{1}{2} \left| \frac{f''(x_n)}{f'(x_n)} \right| \epsilon_n{}^2$$

6. Newton's method only works when $f$ has continuous derivatives in the root neighborhood.
7. When analytical derivatives are hard to compute, calculate slope through nearby points, but convergence tends to be linear (like secant).
8. If the first derivative is not well behaved/does not exit/undefined in the neighborhood of a particular root, the method may overshoot, and diverge from that root.
9. If a stationary point of the function is encountered, the derivative is zero and the method will fail due to division by zero.
10. The stationary point can be encountered at the initial or any of the other iterative points.
11. Even if the derivative is small but not zero, the next iteration will be a far worse approximation.
12. A large error in the initial estimate can contribute to non-convergence of the algorithm (owing to the fact that the zone is outside of the neighborhood convergence zone).
13. If $\alpha$ is a root with multiplicity

$$m > 1$$

then for sufficiently large $n$, the convergence becomes linear, i.e.,

$$\epsilon_{n+1} \approx \frac{m-1}{m} \epsilon_n$$

14. When there are two or more roots that are close together then it may take many iterations before the iterates get close enough to one of them for the quadratic convergence to be apparent.

15. However, if the multiplicity $m$ of the root is known, one can use the following modified algorithm that preserves the quadratic convergence rate (equivalent to using successive over-relaxation)

$$x_{n+1} = x_n - m \frac{f(x_n)}{f'(x_n)}$$

16. The algorithm estimates $m$ after carrying out one or two iterations, and then use that value to increase the rate of convergence. Alternatively, the modified Newton's method may also be used:

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{f'(x_n)f'(x_n) - f(x_n)f''(x_n)}$$

17. It is easy to show that if

$$f'(x_n) = 0$$

and

$$f''(x_n) \neq 0$$

the convergence in the neighborhood becomes linear. Further, if

$$f'(x_n) \neq 0$$

and

$$f''(x_n) = 0$$

convergence becomes cubic.

18. One way of determining the neighborhood of the root. Define

$$g(x) = x - \frac{f(x)}{f'(x)}$$

$$p_n = g(p_{n-1})$$

where $p$ is a fixed point of $g$, i.e.

$$g \in C[a, b]$$

$k$ is a positive constant,

$$p_0 \in C[a, b]$$

and

$$g(x) \in C[a, b] \; \forall \, x \in C[a, b]$$

19. One sufficient condition for $p_0$ to initialize a convergent sequence $\{p_0\}_{k=0}^{\infty}$, which converges to the root

$$x = p$$

of

$$f(x) = 0$$

is that

$$x \in (p - \delta, p + \delta)$$

and that $\delta$ be chosen so that

$$\frac{f(x_n)f''(x_n)}{f'(x_n)f'(x_n)} \leq k < 1 \ \forall \ x \in (p - \delta, p + \delta)$$

20. It is easy to show that under specific choices for the starting variate, Newton's method can fall into a basin of attraction. These are segments of the real number line

such that within each region iteration from any point leads to one particular root - can be infinite in number and arbitrarily small. Also, the starting or the intermediate point can enter a cycle - the $n$-cycle can be stable, or the behavior of the sequence can be very complex (forming a Newton fractal).

21. Newton's method for optimization is equivalent to iteratively maximizing a local quadratic approximation to the objective function. But some functions are not approximated well by quadratic, leading to slow convergence, and some have turning points where the curvature changes sign, leading to failure. Approaches to fix this use a more appropriate choice of local approximation than quadratic, based on the type of function we are optimizing. Next section demonstrates three such generalized Newton rules. Like Newton's method, they only involve the first two derivatives of the function, yet converge faster and fail less often.

22. One significant advantage of Newton's method is that it can be readily generalized to higher dimensions.

23. Also, Newton's method calculates the Jacobian automatically as part of the calibration process, owing to the reliance on derivatives – in particular, automatic differentiation techniques can be effectively put to use.

# Closed Search Methods

## Secant

1. Secant method results on the replacement of the derivative in the Newton's method with a secant-based finite difference slope.
2. Convergence for the secant method is slower than the Newton's method (approx. order is 1.6); however, the secant method does not require the objective function to be explicitly differentiable.
3. It also tends to be less robust than the popular bracketing methods.

## Bracketing Iterative Search

1. Bracketing iterative root searches attempt to progressively narrow the brackets and to discover the root within.
2. The first set discusses the goal search univariate iterator primitives that are commonly used to iterate through the variate.
3. These goal search iterator primitives continue generating a new pair of iteration nodes (just like their bracketing search initialization counter-parts).
4. Certain iterator primitives carry bigger "local" cost, i.e., cost inside a single iteration, but may reduce global cost, e.g., by reducing the number iterations due to faster convergence.
5. Further, certain primitives tend to be inherently more robust, i.e., given enough iteration, they will find the root within – although they may not be fast.

6. Finally, the case of compound iterator search schemes, search schemes where the primitive iteration routine to be invoked at each iteration is evaluated on-the-fly, are discussed.

7. Iterative searches that maintain extended state across searches pay a price in terms of scalability – the price depending on the nature and the amount of state held (e.g., Brent's method carries iteration selection state, whereas Zheng's does not).

## Univariate Iterator Primitive: Bisection

1. Bisection starts by determining a pair of root brackets $a$ and $b$.
2. It iteratively calculates $f$ at

$$c = \frac{a + b}{2}$$

then uses $c$ to replace either $a$ or $b$, depending on the sign. It eventually stops when $f$ has attained the desired tolerance.

3. Bisection relies on $f$ being continuous within the brackets.
4. While the method is simple to implement and reliable (it is a fallback for less reliable ones), the convergence is slow, producing a single bit of accuracy with each iteration.

## Univariate Iterator Primitive: False Position

1. False position works the same as bisection, except that the evaluation point $c$ is linearly interpolated; $f$ is computed at

$$c = \frac{bf(a) + af(b)}{f(a) + f(b)}$$

where $f(a)$ and $f(b)$ have opposite signs. This holds obvious similarities with the secant method.

2. False position method also requires that $f$ be continuous within the brackets.

3. It is simple enough, more robust than secant and faster than bisection, but convergence is still linear to super-linear.

4. Given that the linear interpolation of the false position method is a first-degree approximation of the objective function within the brackets, quadratic approximation using Lagrange interpolation may be attempted as

$$f(x) = \frac{(x - x_{n-1})(x - x_n)}{(x_{n-2} - x_{n-1})(x_{n-2} - x_n)} f_{n-2} + \frac{(x - x_{n-2})(x - x_n)}{(x_{n-1} - x_{n-2})(x_{n-1} - x_n)} f_{n-1}$$
$$+ \frac{(x - x_{n-2})(x - x_{n-1})}{(x_n - x_{n-2})(x_n - x_{n-1})} f_n$$

where we use the three iterates, $x_{n-2}$, $x_{n-1}$ and $x_n$, with their function values, $f_{n-2}$, $f_{n-1}$ and $f_n$.

5. This reduces the number of iterations at the expense of the function point calculations.

6. Using higher order polynomial fit for the objective function inside the bracket does not always produce roots faster or better, since it may result in spurious inflections (e.g., Runge's phenomenon).

7. Further, quadratic or higher fits may also cause complex roots.

## Univariate Iterator Primitive: Inverse Quadratic

1. Performing a fit of the inverse $\frac{1}{f}$ instead of $f$ avoids the quadratic interpolation problem above. Using the same symbols as above, the inverse can be computed as

$$\frac{1}{f(y)} = \frac{(y - f_{n-1})(y - f_n)}{(f_{n-2} - f_{n-1})(f_{n-2} - f_n)} x_{n-2} + \frac{(y - f_{n-2})(x - f_n)}{(f_{n-1} - f_{n-2})(f_{n-1} - f_n)} x_{n-1}$$
$$+ \frac{(y - f_{n-2})(y - f_{n-1})}{(f_n - x_{n-2})(f_n - f_{n-1})} x_n$$

2. Convergence is faster than secant, but poor when iterates not close to the root, e.g., if two of the function values $f_{n-2}$, $f_{n-1}$ and $f_n$ coincide, the algorithm fails.

## Univariate iterator primitive: Ridder's

1. Ridders' method is a variant on the false position method that uses exponential function to successively approximate a root of $f$.
2. Given the bracketing variates, $x_1$ and $x_2$, which are on two different sides of the root being sought, the method evaluates $f$ at

$$x_3 = \frac{x_1 + x_2}{2}$$

3. It extracts exponential factor $\alpha$ such that $f(x)e^{\alpha x}$ forms a straight line across $x_1, x_2,$ and $x_3$. A revised $x_2$ (named $x_4$) is calculated from

$$x_4 = x_3 + (x_3 - x_1)\frac{sign[f(x_1) - f(x_2)]f(x_3)}{\sqrt{f^2(x_3) - f(x_1)f(x_2)}}$$

2. Ridder's method is simpler than Brent's method, and has been claimed to perform about the same.
3. However, the presence of the square root can render it unstable for many of the reasons discussed above.

# Univariate compound iterator: Brent and Zheng

1. Brent's predecessor method first combined bisection, secant, and inverse quadratic to produce the optimal root search for the next iteration.
2. Starting with the bracket points $a_0$ and $b_0$, two provisional values for the next iterate are computed; the first given by the secant method

$$s = b_k - \frac{b_k - b_{k-1}}{f(b_k) - f(b_{k-1})} f(b_k)$$

and the second by bisection

$$m = \frac{a_k + b_k}{2}$$

3. If $s$ lies between $b_k$ and $m$, it becomes the next iterate $b_{k+1}$, otherwise the $m$ is the next iterate.
4. Then, the value of the new contra-point is chosen such that $f(a_{k+1})$ and $f(b_{k+1})$ have opposite signs.
5. Finally, if

$$|f(a_{k+1})| < |f(b_{k+1})|$$

then $a_{k+1}$ is probably a better guess for the solution than $b_{k+1}$, and hence the values of $a_{k+1}$ and $b_{k+1}$ are exchanged.

6. To improve convergence, Brent's method requires that two inequalities must be simultaneously satisfied.
   a) Given a specific numerical tolerance $\delta$, if the previous step used the bisection method, and if

$$\delta < |b_k - b_{k-1}|$$

the bisection method is performed and its result used for the next iteration. If the previous step used interpolation, the check becomes

$$\delta < |b_{k-1} - b_{k-2}|$$

b) If the previous step used bisection, if

$$|s - b_k| < \frac{1}{2}|b_k - b_{k-1}|$$

then secant is used; otherwise the bisection used for the next iteration. If the previous step performed interpolation

$$|s - b_k| < \frac{1}{2}|b_{k-1} - b_{k-2}|$$

is checked instead.

7. Finally, since Brent's method uses inverse quadratic interpolation, $s$ has to lie between $\frac{3a_k + b_k}{4}$ and $b_k$.

8. Brent's algorithm uses three points for the next inverse quadratic interpolation, or secant rule, based upon the criterion specified above.

9. One simplification to the Brent's method adds one more evaluation for the function at the middle point before the interpolation.

10. This simplification reduces the times for the conditional evaluation and reduces the interval of convergence.

11. Convergence is better than Brent's, and as fast and simple as Ridder's.

# Polynomial Root Search

1. This section carries out a brief treatment of computing roots for polynomials.

2. While closed form solutions are available for polynomials up to degree 4, they may not be stable numerically.

3. Popular techniques such as Sturm's theorem and Descartes' rule of signs are used for locating and separating real roots.

4. Modern methods such as VCA and the more powerful VAS use these with Bisection/Newton methods – these methods are used in Maple/Mathematica.

5. Since the eigenvalues of the companion matrix to a polynomial correspond to the polynomial's roots, common fast/robust methods used to find them may also be used.

6. A number of caveats apply specifically to polynomial root searches, e.g., Wilkinson's polynomial shows why high precision is needed when computing the roots – proximal/other ill-conditioned behavior may occur.

7. Finally, special ways exist to identify/extract multiplicity in polynomial roots – they use the fact that $f(x)$ and $f'(x)$ share the root, and by figuring out their GCD.

# Meta-heuristics

## Introduction

1. <u>Definition</u>: Meta-heuristic is a higher-level procedure or heuristic designed to find, generate, or select a lower level procedure or heuristic (partial search algorithm) that may provide a sufficiently good solution to an optimization problem, especially with incomplete or imperfect information or limited computation capacity (Bianchi, Dorigo, Gambardella, and Gutjahr (2009)).

2. <u>Applicability</u>: Meta-heuristics techniques make only a few assumptions about the optimization problem being addressed, so are usable across a variety of problem (Blum and Roli (2003)).

3. <u>Underpinning Philosophy</u>: Many kinds of meta-heuristics implement some kind of stochastic optimization, so the solution depends upon the random variables being generated. As such it does not guarantee that a globally optimal solution can be found over some class of problems.

4. <u>Search Strategy</u>: By searching over a large set of feasible solutions meta-heuristics often finds good solutions with less computation effort than other algorithms, iterative methods, or simple heuristics (see Glover and Kochenberger (2003), Goldberg (2003), Talbi (2009)).

5. <u>Literature</u>: While theoretical results are available (typically on convergence and the possibility of locating global optimum, see Blum and Roli (2003)), most results on meta-heuristics are experimental, describing empirical results based on computer experiments with the algorithms.
   - While high quality research exists (e.g., Sorensen (2013)), enormously numerous meta-heuristics algorithms published as novel/practical have been of flawed quality – often arising out of vagueness, lack of conceptual elaboration, and ignorance of previous literature (Meta-heuristics (Wiki)).

## Properties and Classification

1. <u>Properties</u>: This comes from Blum and Roli (2003):
    a. Meta-heuristics are strategies that guide the search process.
    b. The goal is to efficiently explore the search space in order to find near-optimal solutions.
    c. Techniques that constitute meta-heuristics range from simple local search procedures to complex learning processes.
    d. Meta-heuristic algorithms are approximate and usually non-deterministic.
    e. Meta-heuristics are not problem-specific.
2. <u>Classification</u>: These are taken from Blum and Roli (2003) and from Bianchi, Dorigo, Gambardella, and Gutjahr (2009):
    a. Classification based on the type of the search strategy
    b. Classification off-of single solution search vs. population-based searches
    c. Classification off-of hybrid/parallel heuristics

## Meta-heuristics Techniques

1. <u>Simple Local Search Improvements</u>: In this family of techniques, the search strategy employed is an improvement on simple local search algorithms; examples include simulated annealing, tabu search, iterated local search, variable neighborhood search, and GRASP (Blum and Roli (2003)).
2. <u>Search Improvements with Learning Strategies</u>: The other type of search strategy has a learning component to the search; meta-heuristics of this type include ant colony optimization, evolutionary computation, and genetic algorithms (Blum and Roli (2003)).
3. <u>Single Solution Searches</u>: These focus on modifying and improving a single candidate solution; single solution meta-heuristics include iterated local search, simulated annealing, variable neighborhood search, and tabu search (Talbi (2009)).

4. <u>Population based Searches</u>: Population based searches maintain and improve multiple candidate solutions using population characteristics to guide the search. These meta-heuristics include evolutionary computation, genetic algorithms, and particle swarm optimization (Talbi (2009)).

5. <u>Swarm Intelligence</u>: Swarm intelligence is the collective behavior of de-centralized, self-organized agents in a particle or a swarm. Ant colony optimization (Dorigo (1992)), particle swarm optimization (Talbi (2009)), artificial bee colony (Karaboga (2010)) are all example algorithms of swarm intelligence.

6. <u>Hybrid meta-heuristic</u>: These combine meta-heuristics with other optimization approaches (these could come from e.g., mathematical programming, constraint programming, machine learning etc.). Components of the hybrid meta-heuristic run concurrently and exchange information to guide the search.

7. <u>Parallel meta-heuristic</u>: This employs parallel programming techniques to run multiple meta-heuristics searches in parallel; these may range from simple distributed schemes to concurrent search runs that interact to improve the overall solution.


## Meta-heuristics Techniques in Combinatorial Problems

1. <u>Combinatorial Optimization Problems</u>: In combinatorial optimization, an optimal solution is sought over a discrete search space. Typically, the search-space of the candidate solution goes faster than exponentially as the problem-size increases, making an exhaustive search for the optimal solution infeasible (e.g., the Travelling Salesman Problem (TSP)).

2. <u>Nature and Types of Combinatorial Problems</u>: Multi-dimensional combinatorial problems (e.g., engineering design problems such as form-finding and behavior-finding (Tomoiaga, Chandris, Sumper, Sudria-Andrieu, and Villafafila-Robles (2013)) suffer from the usual curse of dimensionality, making them infeasible to analytical or exhaustive-search methods.

3. <u>Meta-heuristics applied to Combinatorial Optimization</u>: Popular meta-heuristic algorithms for combinatorial problems include genetic algorithms (Holland (1975)),

scatter search (Glover (1977)), simulated annealing (Kirkpatrick, Gelatt, and Vecchi (1983)), and tabu search (Glover (1986)).

## Key Meta-heuristics Historical Milestones

1. <u>Contributions</u>: Many different meta-heuristics are in existence, and new variants are being continually developed. The following key milestone contributions have been extracted from Meta-heuristics (Wiki).
2. <u>1950s</u>:
    a. Robbins and Munro (1951) work on stochastic optimization methods.
    b. Barricelli (1954) carries out the first simulation of the evolutionary process and uses it on general optimization problems.
3. <u>1960s</u>:
    a. Rastrigin (1963) proposes random search.
    b. Matyas (1965) proposes random optimization.
    c. Nelder and Mead (1965) propose a simplex heuristic, which later shown to converge to non-stationary points on some problems.
    d. Fogel, Owens, and Walsh (1966) propose evolutionary programming.
4. <u>1970s</u>:
    a. Hastings (1970) proposes the Metropolis-Hastings algorithm.
    b. Cavicchio (1970) proposes adaptation of the control parameters for an optimizer.
    c. Kernighan and Lin (1970) propose a graph-partitioning method that is related to variable-depth search and prohibition based tabu search.
    d. Holland (1975) proposes genetic algorithm.
    e. Glover (1977) proposes scatter search.
    f. Mercer and Sampson (1978) propose a meta-plan for tuning the optimizer's parameters by using another optimizer.
5. <u>1980s</u>:
    a. Smith (1980) describes genetic programming.
    b. Kirkpatrick, Gelatt, and Vecchi (1983) propose simulated annealing.

    c.  Glover (1986) proposes tabu search, along with the first mention of the word meta-heuristic (Yang (2011)).

    d.  Moscato (1989) proposes memetic algorithms.

6.  <u>1990s</u>:

    a.  Dorigo (1992) introduces ant colony optimization in his PhD thesis.

    b.  Wolpert and MacReady (1995) prove the no free lunch theorems (these are later extended by Droste, Jansen, and Wegener (2002), Igel and Toussaint (2003), and Auger and Teytaud (2010)).

# References

- Auger, A., and O. Teytaud (2010): Continuous Lunches are Free Plus the Design of Optimal Optimization Algorithms *Algorithmica* **57 (1)** 121-146

- Barricelli, N. A (1954): Esempi Numerici di Processi di Evoluzione *Methodos* 45-68

- Bianchi, L., M. Dorigo, L. M. Gambardella, and W. J. Gutjahr (2009): A Survey on Metaheuristics for Stochastic Combinatorial Optimization *Natural Computing: An International Journal* **8 (2)** 239-287

- Blum, C., and A. Roli (2003): Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison *ACM Computing Surveys* **35 (3)** 268-308

- Cavicchio, D. J. (1970): *Adaptive Search using Simulated Evolution* Technical Report **Computer and Communication Sciences Department, University of Michigan**

- Dorigo, M (1992): *Optimization, Learning, and Natural Algorithms* PhD Thesis **Politecnico di Milano**

- Droste, S., T. Jansen, and I. Wegener (2002): Optimization with Randomized Search Heuristics – The (A)NFL Theorem, Realistic Scenarios, and Difficult Functions *Theoretical Computer Science* **287 (1)** 131-144

- Fogel, L., A. J. Owens, and M. J. Walsh (1966): *Artificial Intelligence Through Simulated Evolution* **Wiley**

- Glover, F. (1977): Heuristics for Integer Programming using Surrogate Constraints *Decision Sciences* **8 (1)** 156-166

- Glover, F. (1986): Future Paths for Integer Programming and Links to Artificial Intelligence *Computers and Operations Research* **13 (5)** 533-549

- Glover, F., and G. A. Kochenberger (2003): *Handbook of Metaheuristics* **57 Springer International Series in Operations Research and Management Science**.

- Goldberg, D. E. (1989): *Genetic Algorithms in Search, Optimization, and Machine Learning* **Kluwer Academic Publishers**

- Hastings, W. K. (1970): Monte Carlo Sampling Methods using Markov Chains and their Applications *Biometrika* **57 (1)** 97-109

- Holland, J. H. (1975): *Adaptation in Natural and Artificial Systems* **University of Michigan Press**

- Igel, C., and M. Toussaint (2003): On Classes of Functions for which the No Free Lunch Results hold *Information Processing Letters* **86 (6)** 317-321

- Karaboga, D. (2010): Artificial Bee Colony Algorithm *Scholarpedia* **5 (3)** 6915

- Kernighan, B. W., and S. Lin (1970): An Efficient Heuristic Procedure for Partitioning Graphs *Bell System Technical Journal* **49 (2)** 291-307

- Kirkpatrick, S., C. D. Gelatt Jr., M. P. Vecchi (1983): Optimization by Simulated Annealing *Science* **220 (4598)** 671-680

- Matyas, J. (1965): Random Optimization *Automation and Remote Control* **26 (2)** 246-253

- Mercer, R. E., and J. R. Sampson (1978): Adaptive Search using a Reproductive Meta-plan *Kybernetes* **7 (3)** 215-228

- Moscato, P. (1989): *On Evolution, Search, Optimization, Genetic Algorithms, and Martial Arts: Towards Memetic Algorithms* Report 826 **Caltech Concurrent Computation Program**

- Nelder, J. A., and R. Mead (1965): A Simplex Method for Function Minimization *Computer Journal* **7** 308-313

- Rastrigin, L. A. (1963): The Convergence of Random Search Method in the Extremal Control of a many Parameter System *Automation and Remote Control* **24 (10)** 1337-1342

- Robbins, S., and H. Munro (1951): A Stochastic Approximation Method *Annals of Mathematical Statistics* **22 (3)** 400-407

- Smith, S. F. (1980): *A Learning System based on Genetic Adaptive Algorithms* PhD Thesis **University of Pittsburgh**

- Sorensen, K. (2013): *Metaheuristics—the metaphor exposed*

- Talbi, E. G. (2009): *Metaheuristics: From Design to Implementation* **Wiley**

- Tomoiaga, B., M. Chandris, A. Sumper, A. Sudria-Andrieu, and R. Villafafila-Robles (2013): Pareto Optimal Reconfiguration of Power Distribution Systems using a Genetic Algorithm based on NSGA-II *Energies* **6 (3)** 1439-1455

- Wolpert, D. H., and W. G. MacReady (1995): *No Free Lunch Theorems for Search* Technical Report SFI-TR-95-02-010 **Santa Fe Institute**

- Yang, X. S. (2011): Metaheuristic Optimization *Scholarpedia* **6 (8)** 11472

# Multi-variate Analysis

1. <u>Mean/Variance Location Dependence</u>: The mean is sensitive to both translation and rotation, unless the distribution is mean centered. The variance along a given fixed direction, however, is not sensitive to rotation of the basis.

   - This also implies that the maximal/minimal variances are invariant to representational basis changes. They are, however, sensitive to scaling, though, as PCA itself is.

2. <u>Dimensional Independence vs. Dimensional Realization Independence</u>: Dimensions are distinct (pressure, temperature etc.), but the realizations in those dimensions need not be. Therefore, correlated unit vectors only apply to actual realizations (and they are NOT scale invariant).

3. <u>Orthogonal Data Set in the Native Basis Representation</u>: If a data set is orthogonal under a given basis, then

$$\langle x_i x_j \rangle = 0$$

   (See Figure 1).

4. <u>Non-orthogonal Data Set in the Native Basis Representation</u>: In this case, the native representation basis results in

$$\langle x_i x_j \rangle \neq 0$$

   (See Figure 2). Thus, an orthogonalization operation needs to be performed such that under the new basis

$$\langle x_i' x_j' \rangle = 0$$

(See Figure 3).

5. <u>Orthogonalization as Principal Components Extraction</u>: As can be seen from figures 2 and 3, under the new schematic axes

$$\langle x_i' x_j' \rangle = 0$$

Further these correspond to the principal components, i.e., orthogonal components where the variance is an extremum.

6. <u>Orthogonalized Representation</u>: Upshot of all these is that, if the representation basis is structured such that

$$\langle x_i x_j \rangle = 0$$

for all

$$i \neq j$$

then these representations automatically correspond to principal components as well.

7. <u>Full Rank Matrix</u>: This is simply an alternate term for multi-collinear matrix.

## Parallels between Vector Calculus and Statistical Distribution Analysis

1. <u>DOT PRODUCT</u> => The notion of dot product is analogous to covariance operation in statistical analysis, i.e., DOT PRODUCT is

$$\vec{x}_i \cdot \vec{x}_j = 0$$

if

$$\vec{x}_i \perp \vec{x}_j$$

and covariance is

$$\langle x_i x_j \rangle = 0$$

if $x_i$ and $x_j$ are orthogonal to each other.

2. <u>Distance Metric</u> => Vector Euclidean distance is $\sum[(x_1 - x_2)^2 + (y_1 - y_2)^2 + \cdots]$ is equivalent to the variance $\sum[(x_i - \mu_i)^2 + \cdots]$. Further, extremizing the Euclidean/Frobenius distance is analogous to variance minimization/maximization techniques.

# Linear Systems Analysis and Transformation

## Matrix Transforms

1. Co-ordinate Rotation and Translation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} + B$$

   where $A$ is the rotating transformer, and $B$ is the translator.

2. Scaling vs. Rotation: Say

$$A = \begin{bmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{bmatrix}$$

   If

$$a_{11} = a_{22} \neq 1$$

   and

$$a_{12} = a_{21} = 0$$

   it becomes pure scaling.

$$a_{11} \neq a_{22}$$

   and

$$a_{12} = a_{21} = 0$$

produces differential elongation/compression and

$$a_{12} \neq 0$$

or

$$a_{21} \neq 0$$

results in rotation.

3. <u>Uses of Gaussian Elimination</u>:
   - Linearization
   - Orthogonalization
   - Inversion
   - Diagonalization
   - Lower/Upper Triangle Decomposition (LU Decomposition)
   - Independent Component Extraction
   - Principal Component Analysis

4. <u>Diagonal Identity Matrix Conception</u>: Given a matrix $A$ what matrix $M$ should it be transformed by to get a diagonal identity matrix, i.e.

$$MA = I$$

Answer is

$$M = A^{-1}$$

Thus, diagonalization is also an inversion operation.

$$Diagonalization == Orthogonalization == Inversion == ICA$$

Diagonalization, of course, is unique only to a diagonal entry, whereas inversion corresponds to a specific choice of the diagonal entry.

## Systems of Linear Equations

1. <u>Importance of Diagonal Dominance in Gauss-Seidel</u>: Is diagonal dominance important because the dominant diagonal's contribution to the RHS drives the given equation's value, and therefore the iterative accuracy?

2. <u>Eigenization Square Matrix Inversion Conceptualization</u>: Given a source matrix

$$F_{SOURCE,0} = \{f_{ij}\}_{i,j=0}^{n-1}$$

and an initialized target inverse identity matrix

$$F_{INV,0} = \{I\}_{n \times n}$$

achieve a suitable set of $p$ transformations simultaneously on $F_{SOURCE}$ and $F_{INV}$ to eventually make

$$F_{SOURCE,p} = \{I\}_{n \times n}$$

so that the corresponding

$$F_{INV,p} = \{f_{INV,ij}\}_{i,j=0}^{n-1}$$

becomes the inverse.

3.  <u>Valid Inversion Rules</u>: These are fairly straight forward application of the Gaussian elimination scheme:

    - Scale a single $F_{SOURCE}$ row by a constant => scale the corresponding $F_{INV}$ row by the same constant.

    - Add/subtract any pair of $F_{SOURCE}$ rows from each other => add/subtract the same pair of $F_{INV}$ rows from each other.

4.  <u>Matrix Inversion using Gaussian Elimination</u>:

    - Scan across the diagonal entries.

    - Scan through the rows corresponding to each diagonal entry.

    - If a given row entry call value is zero, or its index corresponds to that of a diagonal, skip.

    - Calculate the $WorkColFactor$ as

$$WorkColFactor = \frac{DiagonalEntry}{CellValueEntry}$$

    - Scan all the cells in the column of the current cell.

    - Apply the $WorkColFactor$ product to each entry in the current working column of the source matrix.

    - Do the same as above to the inverse matrix.

    - Subtract the entries corresponding to the designated diagonal column from the working column to make the current entry zero.

    - Do the same as above to the inverse matrix as well.

    - After the completion of all such diagonal scans, row scans, and working column scans, re-scan the diagonal again.

    - Scale down each entry of the source matrix by itself, so that the source matrix entries now constitute an identity ($\{I\}_{n \times n}$) matrix.

    - Do the same as above to the inverse matrix as well.

5. <u>Non-invertible Coefficient Matrix, but Solution exists</u>: For unprocessed coefficient matrices, certain conditions (such as zero diagonal entries) may cause the coefficient matrix to be technically non-invertible, but that does not automatically mean that the system is unsolvable – a simply re-casting of the basic linear system set may be all that is required.

6. <u>Linear Basis Re-arrangement</u>: Sometimes, a singular coefficient matrix (with zero determinant, therefore non-invertible) may be re-arranged to create an invertible coefficient matrix. After inversion, it can be re-structured again to extract the inverse (which is just a coefficient Jacobian).

7. <u>Rows/Columns as "Preferred Linear Basis Sequence" for Matrix Manipulation</u>: Consider the solution to

$$AX = Z$$

where $X$ and $Z$ are columns. In this case, the notion of constraint linear representation is maintained exclusively in rows. Therefore, all elimination/scaling basis operations need to be applied on that basis. In considering the solution to

$$AX = Z$$

where $X$ and $Z$ are rows, columns now become the preferred linear basis sequence.

8. <u>Regularization before Gauss Elimination</u>: Before the Gauss Elimination can process, we need to diagonalize the matrix. Row swapping is a more robust way to diagonalize than row accumulation due to a couple of reasons:

   - Matrix Row Swap vs. Row Cumulate => In all cases, row swap can be transformed into row accumulation. When inverting, however, the row swapping AND accumulation should both be done on both the SOURCE and the TARGET matrices.
   - From a core linear operation set point-of-view, row accumulation is the inverse of the eventual of Gauss Elimination, so the danger is that the

diagonalization gains of row swap may be undone during the intermediate stages of Gauss Elimination.

- Even more important is that row swapping simply retains the original information, by just re-arranging the row set.

9. Row Swapping Caveats:

- Always swap rows by retaining the directionality of the scan AND by retaining the scan initial node preceding the swap (to choose the pivot). One way to do this is by starting the scan at $row + 1$ - or from

$$row = 0$$

if the edge has been reached - AND always keeping the scan sequence forward/backward.

- Also ensure that the target swapped row is "valid", i.e., it's post-swapped diagonal entry should be non-zero. If this cannot be achieved through the scan, then that is an error condition.

10. Diagonalization/Inversion Algorithm: Work in terms of an intermediate transform variate $Z$ produced by re-arranging the original coefficient matrix and $Y$ such that the $A$ in

$$AX = Z$$

is now invertible. The following would be the steps:

- First, re-arrange the equation system set to identify a suitable pair $A$ and $Z$ such that $A$ is invertible, and $Z$ is estimated from

$$AX = Z$$

- The re-arranging linear operation set will produce $A$ such that

$$BY = Z \implies AX = BY \implies X = A^{-1}BY$$

11. **More General Inverse Transformation Re-formulation:** Given the coefficient matrix $\{a_{qj}\}_{j,q=0}^{n-1}$, the set of unknown variables $\{x_j\}_{j=0}^{n-1}$, and the RHS $\{y_q\}_{q=0}^{n-1}$, $y_p$, $y_q$, and the corresponding $z_p$ and $z_q$ are given as

$$\sum_{j=0}^{n-1} a_{qj}x_j = y_q$$

$$\sum_{j=0}^{n-1} a_{pj}x_j = y_p$$

and

$$\{z_j = y_j\}_{j=0}^{n-1}$$

On transformation (i.e., adding row $p$ to row $q$), you get

$$\sum_{j=0}^{n-1}(a_{pj} + a_{qj})x_j = y_p + y_q$$

and therefore

$$\{z_j = y_j\}_{j=0;j\neq q}^{n-1}$$

along with

$$z_q = y_p + y_q$$

This clearly implies that

$$\frac{\partial z_q}{\partial y_p} = 1$$

or

$$B_{qp} = B_{qp} + 1$$

## Orthogonalization

1. <u>2D Orthogonalization</u>: In 2D, you need to fix one 1D orthogonal axis to be able to orthogonalize the other.
2. <u>2D Equation System</u>:

$$x_1 = a_{11}s_1 + a_{12}s_2$$

and

$$x_2 = a_{21}s_1 + a_{22}s_2$$

This has 4 unknowns, so one solution for this is as follows: Fix

$$a_{11} = 1$$

and

$$a_{12} = 0$$

This results in 2 unknowns. Setting

$$\langle x_1{}^2 \rangle = 1$$

$$\langle x_2{}^2 \rangle = 1$$

and

$$\langle x_1 x_2 \rangle = \rho$$

you get

$$a_{11}{}^2 + a_{12}{}^2 = 1$$

and

$$a_{21}{}^2 + a_{22}{}^2 = 1$$

resulting in

$$a_{21} = \rho$$

and

$$a_{21} = \sqrt{1 - \rho^2}$$

Thus, all unknowns are determined.

3. <u>nD Orthogonalization</u>:

$$\begin{bmatrix} x_0 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} a_{0,0} & \cdots & a_{0,n-1} \\ \vdots & \ddots & \vdots \\ a_{n-1,0} & \cdots & a_{n-1,n-1} \end{bmatrix} \begin{bmatrix} s_0 \\ \vdots \\ s_{n-1} \end{bmatrix}$$

- Number of diagonal entries $=> n$
- Number of non-diagonal entries $=> n^2 - n$
- Net Number of equations => Number of diagonal entries + Number of non-diagonal entries =>

$$n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$$

4. <u>nD Unknowns Analysis</u>:
   - Number of Unknowns $=> n^2$.
   - Fix the first row, and the number of unknowns becomes $=> n^2 - n$.
   - Fixing the row takes off one equation, so the number of equations $=> \frac{n(n+1)}{2} - 1$
   - Number of equations = Number of Unknowns =>

$$\frac{n(n+1)}{2} - 1 = n^2 - n$$

   results in

$$(n-1)(n-2) = 0$$

## Gaussian Elimination

1. <u>n-D Gaussian Elimination</u>: What does it work? If you fix a row, you can rotate the other rows to eliminate dependence on an ordinate of the fixed row.

2. <u>Row Fixation as a Basis Choice</u>: Row elimination does not automatically make the matrix diagonal or orthogonalize it – all it does is to eliminate dependence on a stochastic variate.

3. <u>Number of Elimination Rotations</u>: The first row fixation results in $n-1$ rotations, the second row fixation results in $n-2$ rotations, and so on. Thus, the total number of rotating transformations is $\frac{(n-1)(n-2)}{2}$ (i.e., the same as the number of unknowns seen earlier). The result of these transformations is a lower/upper triangular matrix.

4. <u>Final Reversing Sweep</u>: An additional reverse sweep would eliminate similar column dependencies as well – resulting in another $\frac{(n-1)(n-2)}{2}$ rotation choices.

5. <u>Number of Sweep Operations</u>: Total result of all the rotations and their corresponding choices =>

$$2 \cdot \frac{(n-1)(n-2)}{2} = (n-1)(n-2)$$

# Rayleigh Quotient Iteration

## Introduction

1. Idea behind Rayleigh Quotient Iteration: **Rayleigh Quotient Iteration** is an eigenvalue algorithm that extends the idea of inverse iteration by using the Rayleigh Quotient to obtain increasingly accurate eigenvalue estimates (Wiki - Rayleigh Quotient Iteration (2018)).

2. Progressive Navigation towards "True" Selection: Rayleigh Quotient Iteration is an iterative method, i.e., it delivers a sequence of approximate solutions that converge to a true solution in the limit. This is true for all algorithms that compute eigenvalues; since the eigenvalues can be irrational numbers, there can be no general method for computing them in a finite number of steps.

3. Practicality of the Iterative Approach: Very rapid convergence is guaranteed and no more than a few iterations are needed in practice to obtain a reasonable solution.

4. Cubic Convergence for Typical Matrices: The Rayleigh Quotient algorithm converges cubically for Hermitian or symmetric matrices, given an initial vector that is sufficiently close to an eigenvector of the matrix that is being analyzed.

## The Algorithm

1. Update Eigenvalue using Rayleigh Quotient: The algorithm is very similar to inverse iteration, but replaces the estimated eigenvalues at the end of each iteration with the Rayleigh Quotient.

2. Initial Guess for Eigenvalue/Eigenvector: Begin by choosing a value $\mu_0$ as an initial eigenvalue guess for the Hermitian matrix $A$. An initial vector $b_0$ must also be supplied as the initial eigenvector guess.

3. Iterative Eigenvalue and Eigenvector: Calculate the subsequent approximation of the eigenvector $b_{i+1}$ by

$$b_{i+1} = \frac{[A - \mu_i I]^{-1} b_i}{\|[A - \mu_i I]^{-1} b_i\|}$$

where $I$ is the identity matrix, and set the next approximation of the eigenvalue to the Rayleigh quotient of the current iteration equal to

$$\mu_i = \frac{b_i{}^* A b_i}{b_i{}^* b_i}$$

4. Deflation Techniques for Successive Eigenvalues: To compute more than one eigenvalue the algorithm can be combined with a deflation technique.

5. Treatment for Very Small Matrices: Note that for very small matrices it is beneficial to replace the matrix inverse with the adjugate, which will yield the same iteration because it is equal to the inverse to an irrelevant scale – specifically, the inverse of the determinant.

6. Advantages of using the Adjugate: The adjugate is easier to compute explicitly than the inverse – though the inverse is easier to apply to a vector for problems that aren't small – and is more numerically sound because it remains well-defined as the eigenvalue changes.

## References

- [Wikipedia – Rayleigh Quotient Iteration (2018)](#)

# Power Iteration

## Introduction

1. <u>Power Iteration – Problem Statement/Definition</u>: In mathematics, **power iteration** – also known as the *power method* – is an eigenvalue algorithm. Given a diagonalizable matrix $A$, the algorithm will produce a number $\lambda$, which is the greatest – in absolute value – eigenvalue of $A$, and a non-zero vector $v$, the corresponding eigenvector of $\lambda$, such that

$$Av = \lambda v$$

The algorithm is also known as the von Mises iteration (von Mises and Pollazek-Geiringer (1929)).

2. <u>Characteristics of the Power Iteration Algorithm</u>: Power iteration is a very simple algorithm, but it may converge slowly. It does not compute a matrix decomposition, and hence can be used when $A$ is a very large sparse matrix.

## The Method

1. <u>Starting Choice for Principal Eigenvector</u>: The power iteration method starts with a vector $b_0$, which may be either an approximation to the dominant eigenvector or a random factor.

2. <u>Recurrence Relation for Power Iteration</u>: The method is described by the recurrence relation

$$b_{k+1} = \frac{Ab_k}{\|Ab_k\|}$$

So, at every iteration, the vector $b_k$ is mulplied by the matrix $A$ and normalized.

3. <u>Necessary Condition for Eigen-component Convergence</u>: If one assumes that $A$ has an eigenvalue that is greater in magnitude than its other eigenvalues and that the starting vector $b_0$ has a non-zero component in the direction of an eigenvector associated with the dominant eigenvalue, then a sub-sequence $\{b_k\}$ converges to an eigenvector associated with the dominant eigenvector.

4. <u>Recast of $\{b_k\}$ using Phase Shift</u>: Without the two assumptions above the sequence does not necessarily converge. In this sequence

$$b_k = e^{i\phi_k}v_1 + r_k$$

where $v_1$ is the eigenvector associated with the dominant eigenvalue, and

$$\|r_k\| \to 0$$

The presence of the term $e^{i\phi_k}$ implies that $\{b_k\}$ does not converge unless

$$e^{i\phi_k} \equiv 1$$

5. <u>Convergence to the Dominant Eigenvalue</u>: Under the two assumptions listed above, the sequence $\{\mu_k\}$ defined by

$$\mu_k = \frac{b_k^T Ab_k}{b_k^T b_k}$$

converges to the dominant eigenvalue.

6. Eigenvector and Eigenvalue Sequences: The vector $b_k$ is the associated eigenvector. Ideally one should use the Rayleigh Quotient in order to get the associated eigenvalue.

7. Spectral Radius using the Rayleigh Quotient: The method can also be used to calculate the spectral radius – the largest eigenvalue of a matrix – by computing the Rayleigh Quotient

$$\frac{b_k{}^T A b_k}{b_k{}^T b_k} = \frac{b_{k+1}{}^T b_k}{b_k{}^T b_k}$$

## Analysis

1. Decomposition to Jordan Canonical Form: Let $A$ be decomposed into its Jordan Canonical Form

$$A = VJV^{-1}$$

where the first column of $V$ is an eigenvector of $A$ corresponding to the dominant eigenvalue $\lambda_1$.

2. The Principal Component Jordan Block: Since the dominant eigenvalue of $A$ is unique, the first Jordan block of $J$ is the $1 \times 1$ matrix $[\lambda_1]$, where $\lambda_1$ is the largest eigenvalue of $A$ in magnitude.

3. Initializing the Principal Component Eigenvector: The starting vector $b_0$ can be written as a linear combination of the columns of $V$:

$$b_0 = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n$$

By assumption, $b_0$ has a non-zero component in the direction of the dominant eigenvalue, so

$$c_1 \neq 0$$

4. $k^{\text{th}}$ Recurrence of the Initial Vector: The computationally recurrence relation for $b_{k+1}$ can be written as:

$$b_{k+1} = \frac{Ab_k}{\|Ab_k\|} = \frac{A^{k+1}b_0}{\|A^{k+1}b_0\|}$$

where the expression $\frac{A^{k+1}b_0}{\|A^{k+1}b_0\|}$ is amenable to the analysis below.

5. Principal Component Dependence of $b_k$:

$$
\begin{aligned}
b_k &= \frac{A^k b_0}{\|A^k b_0\|} = \frac{[VJV^{-1}]^k b_0}{\|[VJV^{-1}]^k b_0\|} = \frac{VJ^k V^{-1} b_0}{\|VJ^k V^{-1} b_0\|} \\
&= \frac{VJ^k V^{-1}(c_1 v_1 + c_2 v_2 + \cdots + c_n v_n)}{\|VJ^k V^{-1}(c_1 v_1 + c_2 v_2 + \cdots + c_n v_n)\|} \\
&= \frac{VJ^k(c_1 \hat{e}_1 + c_2 \hat{e}_2 + \cdots + c_n \hat{e}_n)}{\|VJ^k(c_1 \hat{e}_1 + c_2 \hat{e}_2 + \cdots + c_n \hat{e}_n)\|} \\
&= \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{c_1}{|c_1|} \cdot \frac{v_1 + \frac{1}{c_1} V\left(\frac{1}{\lambda_1}J\right)^k (c_2 \hat{e}_2 + \cdots + c_n \hat{e}_n)}{\left\|v_1 + \frac{1}{c_1} V\left(\frac{1}{\lambda_1}J\right)^k (c_2 \hat{e}_2 + \cdots + c_n \hat{e}_n)\right\|}
\end{aligned}
$$

6. Scaling Down the Principal Eigenvector: As

$$k \to \infty$$

the expression above simplifies to

$$\left(\frac{1}{\lambda_1}J\right)^k = \begin{pmatrix} [\mathbb{I}] & \cdots & & \\ \cdots & \left(\frac{1}{\lambda_1}J_2\right)^k & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \left(\frac{1}{\lambda_1}J_n\right)^k \end{pmatrix} \rightarrow \begin{pmatrix} [\mathbb{I}] & \cdots & \cdots & \cdots \\ \cdots & 0 & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

7. <u>Limit of $\left(\frac{1}{\lambda_1}J_i\right)^k$ as $k \to \infty$</u>: The limit follows from the fact that the eigenvalue of $\left(\frac{1}{\lambda_1}J_i\right)^k$ is less than 1 in magnitude, so

$$\left(\frac{1}{\lambda_1}J_i\right)^k \to 0$$

as

$$k \to \infty$$

8. <u>Expansion of Higher Order Eigen Terms</u>: It follows that

$$\frac{1}{c_1}V\left(\frac{1}{\lambda_1}J\right)^k (c_2\hat{e}_2 + \cdots + c_n\hat{e}_n) \to 0$$

as

$$k \to \infty$$

9. <u>Reduction of $b_k$ for $k \to \infty$</u>: Using this fact, $b_k$ can be written in a form that emphasizes its relationship with $v_1$ when $k$ is large:

$$b_k = \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{c_1}{|c_1|} \frac{v_1 + \frac{1}{c_1}V\left(\frac{1}{\lambda_1}J\right)^k (c_2\hat{e}_2 + \cdots + c_n\hat{e}_n)}{\left\| v_1 + \frac{1}{c_1}V\left(\frac{1}{\lambda_1}J\right)^k (c_2\hat{e}_2 + \cdots + c_n\hat{e}_n)\right\|} = e^{i\phi_k} \frac{c_1}{|c_1|} \frac{v_1}{\|v_1\|} + r_k$$

where

$$e^{i\phi_k} = \left(\frac{\lambda_1}{|\lambda_1|}\right)^k$$

and

$$\|r_k\| \to 0$$

as

$$k \to \infty$$

10. <u>Uniqueness of the $\{b_k\}$ Sequence</u>: The sequence $\{b_k\}$ is bounded, so it contains a convergent sub-sequence. Note that the eigenvector corresponding to the dominant eigenvalue is unique only upto a scaler, so although the sequence $\{b_k\}$ may not converge, $b_k$ is nearly an eigenvector of $A$ for large $k$.

11. <u>Alternative Proof of the Sequence Convergence</u>: Alternatively, if $A$ is diagonalizable, then the following proof yields the same result. Let $\lambda_1, \lambda_2, \cdots, \lambda_m$ be the $m$ eigenvalues – counted with multiplicity – of $A$, and let $v_1, v_2, \cdots, v_m$ be the corresponding eigenvectors. Suppose that $\lambda_1$ is the dominant eigenvalue, so that

$$|\lambda_1| > |\lambda_j|$$

for

$$j > 1$$

12. <u>Suitable Choice for the Initial Vector</u>: As seen earlier, the initial vector $b_0$ is written as

$$b_0 = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n$$

If $b_0$ is chosen randomly – with uniform probability – the

$$c_1 \neq 0$$

with probability 1

13. <u>Power Iteration over Initial Eigenvector</u>: Now

$$A^k b_0 = c_1 A^k v_1 + c_2 A^k v_2 + \cdots + c_n A^k v_n = c_1 {\lambda_1}^k v_1 + c_2 {\lambda_2}^k v_2 + \cdots + c_n {\lambda_n}^k v_n$$

$$= c_1 {\lambda_1}^k \left[ v_1 + \frac{c_2}{c_1} \left( \frac{\lambda_2}{\lambda_1} \right)^k v_2 + \cdots + \frac{c_n}{c_1} \left( \frac{\lambda_n}{\lambda_1} \right)^k v_n \right] \to c_1 {\lambda_1}^k v_1$$

as long as

$$\left| \frac{\lambda_j}{\lambda_1} \right| < 1$$

for

$$j > 1$$

14. <u>Convergence to the Principal Eigenvector</u>: On the other hand,

$$b_k = \frac{A^k b_0}{\|A^k b_0\|}$$

Therefore $b_k$ converges to a multiple of the eigenvector $v_1$

15. <u>Convergence Ratio of Power Iteration</u>: The convergence ratio is geometric with the ratio $\left|\frac{\lambda_2}{\lambda_1}\right|$ where $\lambda_2$ denotes the second principal eigenvalue.

16. <u>Consequence of Close Principal Eigenvectors</u>: Thus, the method converges slowly if there is an eigenvalue close in magnitude to the dominant eigenvalue.

## Applications

1. <u>Identification of the Dominant Eigenvector/Eigenvalue</u>: Although the power iteration method approximates only one eigenvalue of a matrix, it remains useful for several computational problems.

2. <u>Use in Google and Twitter</u>: For instance, Google uses it to calculate the PageRank of documents in their search engine (Ipsen and Wills (2005)), and Twitter uses it to show users recommendations of who to follow (Gupta, Goel, Lin, Sharma, Wang, and Zadeh (2013)).

3. <u>Space Advantage of Power Iteration</u>: The power iteration is especially suitable for sparse matrices, such as the web matrix, or as the matrix-free method that does not require storing the coefficient matrix $A$ explicitly, but can instead access a function evaluating matrix-vector products $Ax$.

4. <u>Power Iteration vs. Arnoldi Method</u>: For non-symmetric matrices that are well-conditioned, the power iteration method can outperform the more complex Arnoldi method.

5. <u>Power Iteration vs. Lanczos/LOBPCG</u>: For symmetric matrices, the power iteration method is rarely used, since its convergence speed can be easily increased without sacrificing the smaller cost per iteration, e.g., Lanczos iteration and LOBPCG.

6. <u>Variation in the Method - Inverse Iteration</u>: Some of the more advanced algorithms can be understood as variations of the power iteration method. For instance, the inverse iteration method applies power iteration to the matrix $A^{-1}$.

7. <u>Sub-space of Eigenvalues - Krylov Subspace</u>: Other algorithms look at the whole subspace generated by the vectors $b_k$. This subspace is known as the Krylov subspace. It can be computed by Arnoldi iteration or Lanczos iteration.

## References

- Gupta, P., A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh (2013): [WTF – The Who-To-Follow Service at Twitter](#)

- Ipsen, I., and R. Wills (2005): [Analysis and Computation of Google's PageRank](#)

- Von Mises, R., and H. Pollaczek-Geiringer (1929): Praktische Verfahren der Gleichingsauflosung *Zeitschrift fur Andewandte Mathematik und Mechanik* **9** 152-164

# Sylvester's Formula

## Overview

1. <u>Analytical Expression of Matrix Function</u>: *Sylvester's Formula*, *Sylvester's Matrix Theorem*, or *Lagrange-Sylvester interpolation* expresses an analytic function $f(A)$ of a matrix $A$ as a polynomial in $A$, in terms of the eigenvectors and eigenvalues of $A$ (Claerbout (1985), Horn and Johnson (1991), Wikipedia (2019))).

2. <u>In Terms of Eigenvalues and Eigenvectors</u>: It states that (Sylvester (1883))

$$f(A) = \sum_{i=1}^{k} f(\lambda_i) A_i$$

where the $\lambda_i$ are the eigenvalues of $A$, and the matrices

$$A_i = \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{1}{\lambda_i - \lambda_j} (A - \lambda_j I)$$

are the corresponding Frobenius covariants of $A$, which are the projection matrix Lagrange polynomials of $A$.

## Conditions

Sylvester's formula applies for any diagonalizable matrix $A$ with $k$ distinct eigenvalues $\lambda_1, \cdots, \lambda_k$ and any function $f$ defined on some subset of complex numbers such that $f(A)$

is well defined. The last condition means that every eigenvalue $\lambda_i$ is in the domain of $f$, and that every eigenvalue $\lambda_i$ with multiplicity

$$m_i > 1$$

is in the interior of the domain with $f$ being $m_i - 1$ times differentiable at $\lambda_i$ (Horn and Johnson (1991)).

## Generalization

1. <u>Buchheim Extension Based on Hermite Polynomials</u>: Sylvester's formula is only valid for diagonalizable matrices; an extension due to Buchheim (1884), based on Hermite interpolating polynomials, covers the general case:

$$f(A) = \sum_{i=1}^{s} \left[ \sum_{j=0}^{n_i-1} \frac{1}{j!} \phi_i^{(j)}(\lambda_i) (A - \lambda_i I)^j \prod_{\substack{j=1 \\ j \neq i}}^{s} (A - \lambda_j I)^{n_j} \right]$$

where

$$\phi_i^{(j)} := \frac{f(t)}{\prod_{\substack{j=1 \\ j \neq i}}^{s} (t - \lambda_j)^{n_j}}$$

2. <u>Concise Expression of Schwerdtfeger</u>: A further concise form is given by Schwerdtfeger (1938) as

$$f(A) = \sum_{i=1}^{s} A_i \sum_{j=0}^{n_i-1} \frac{1}{j!} f^{(j)}(\lambda_i)(A - \lambda_i I)^j$$

# References

- Buchheim, A. (1884): On the Theory of Matrices *Proceedings of the London Mathematical Society* **1-16 (1)** 63-82
- Claerbout, J. F. (1985): *Fundamentals of Geophysical Data Processing* **Blackwell Scientific Publishing**
- Horn, R. A., and C. R. Johnson (1985): *Topics in Matrix Analysis* **Cambridge University Press**
- Sylvester, J. J. (1883): On the Equation to the Secular Inequalities in the Planetary Theory *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **16 (100)** 267-269
- Wikipedia (2019): [Sylvester's Formula](#)

# Numerical Integration

## Introduction and Overview

1. <u>Numerical Estimate of Definite Integrals</u>: In numerical analysis, **numerical integration** comprises a broad family of algorithms for calculating the numerical value of a definite integral, and by extension, the term is also used sometimes to describe the numerical solution to differential equations.

2. <u>Chapter Focus on Definite Integrals</u>: This chapter focuses on the calculation of definite integrals.

3. <u>Numerical Quadrature: One-Dimensional Integrals</u>: The term **numerical quadrature** – often abbreviated to *quadrature* – is more or less a synonym for **numerical integration**, especially as applied to one-dimensional integrals (Wikipedia (2019)).

4. <u>Curbature: Multi-Dimensional Numerical Integrals</u>: Some authors refer to numerical integration over more than one dimension as *curbature*; others take *quadrature* to include higher dimensional integration.

5. <u>Mathematical Specification of the Definite Integral</u>: The basic problem in numerical integration is to compute an approximate solution to a definite integral $\int_a^b f(x)dx$ to a given degree of accuracy.

6. <u>Accurately Estimating the Definite Integral</u>: If $f(x)$ is a smooth function integrated over a small number of dimensions, and the domain of the integration is bounded, there are many methods for approximating the integral to the desired precision.

# Reasons for Numerical Integration

1. Discrete Sampling of the Integrand: There are several reasons for carrying out numerical integration. For instance, the integrand $f(x)$ may only be known at certain points, such as those obtained by sampling. Some embedded systems and other software applications may need numerical integration for this reason.

2. Anti-derivative not an Elementary Function: An expression for the integrand may be known, but it may be difficult or impossible to find an anti-derivative that is an elementary function. An example of such an integrand is

$$f(x) = e^{-x^2}$$

   the anti-derivative of which – the error function times a constant – cannot be written in an elementary form.

3. Series Approximation of the Anti-derivative: It may be possible to find an anti-derivative symbolically, but it may be easier to compute a numerical approximation than to compute the anti-derivative. That may be the case if the anti-derivative is computed as an infinite derivative or product, or it its evaluation requires a special function that is not available.

# Methods for One-Dimensional Integrals

1. <u>Combining Evaluations of the Integrand</u>: Numerical integration methods can generally be described as combining the evaluations of the integrand to get an approximation to the integral.

2. <u>Weighted Sum at the Integration Points</u>: The integrand is evaluated at a finite set of points called the *integration points* and a weighted sum of these values is used to approximate the integral. The integration points and the weights depend on the specified method used and the accuracy required from the approximation.

3. <u>Approximation Error of the Method</u>: An important part of the analysis of any numerical integration method is to study the behavior of the approximation error as a function of the number of integrand evaluations. A method that yields a small error for a small number of evaluations widely considered superior.

4. <u>Reduced Number of Integrand Evaluations</u>: Reducing the number of operations of the integrand reduces the number of arithmetic operations involved, and therefore reduces the total round-off error. Also, each evaluation takes time, and the integrand may be arbitrarily complicated.

5. <u>Conditions for Brute Force Integration</u>: A brute force kind of numerical integration can be done if the integrand is reasonably well-behaved, i.e., it is piece-wise continuous and of bounded variation, by evaluating the integrand with very small increments.

## Quadrature Rules Based on Interpolating Functions

1. <u>Error of Integrating Interpolation Polynomials</u>: A large class of quadrature rules can be derived by constructing interpolating functions that are easy to integrate. Typically, these interpolating functions are polynomials. In practice, some of these polynomials of very high degree tend to oscillate very wildly, only polynomials of low degree are used, typically linear and quadratic.

2. <u>Mid-Point Rule: Zero Degree Polynomial</u>: The simplest function of this type is to let the interpolating function be a constant function – a polynomial of degree $0$ – that passes through the point $\left[\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right]$ This is the *mid-point rule* of the *rectangle rule*:

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{a+b}{2}\right)$$

3. <u>Trapezoidal Rule: First Degree Polynomial</u>: The interpolating function may be a straight line – an affine function, i.e., a polynomial of degree one – passing through the points $[a, f(a)]$ and $[b, f(b)]$. This is called the *trapezoidal rule*:

$$\int_a^b f(x)dx \approx (b-a)\left[\frac{f(a)+f(b)}{2}\right]$$

4. <u>Sub division of the Integration Intervals:</u> For either one of these rules, a more accurate representation can be made by breaking up the interval $[a, b]$ into some number $n$ of sub-intervals, computing an approximation for each sub-interval, then adding up all the results. This is called *composite rule*, *extended rule*, or *iterated rule*.

5. <u>Example: Composite Trapezoidal Rule</u>: For example, the composite trapezoidal rule can be stated as

$$\int_a^b f(x)dx \approx (b-a)\left[\frac{f(a)+f(b)}{2} + \sum_{k=1}^{n-1} f\left(a+k\frac{b-a}{n}\right)\right]$$

where the sub-intervals have the form

$$[a + kh, a + (k + 1)h] \subset [a, b]$$

with

$$h = \frac{b - a}{n}$$

and

$$k = 0, \cdots, n - 1$$

Here the sub-intervals used have the same length $h$, but one could use intervals of varying length $h_k$

6. <u>Definition of Newton-Cotes Formula</u>: Interpolation with polynomials evaluated at equally spaced points in $[a, b]$ yields the Newton-Cotes formula, of which the rectangle rule and the trapezoidal rule are examples. Simpson's rule, which is based on a polynomial of order 2, is also a Newton-Cotes formula.

7. <u>Quadrature Rules with Nesting Property</u>: Quadrature rules with equally spaced points have the very convenient property of *nesting*. The corresponding rule with each interval sub-divided includes all the current points, so those integrand points can be re-used.

8.  <u>Integrand with Variable Interpolant Spaces</u>: If one allows the intervals between the interpolation points to vary, one finds another group of quadrature formulas, such as the Gaussian quadrature formula.

9.  <u>Improved Accuracy with Gaussian Quadrature</u>: Gaussian quadrature rule is typically more accurate than a Newton-Cotes rule, which requires the same number of function evaluations, if the integrand is smooth, i.e., if it is sufficiently differentiable.

10. <u>Other Varying Interval Quadrature Rules</u>: Other quadrature methods with varying intervals include Clenshaw-Curtis quadrature methods – also called Fejer quadrature – and these do nest.

11. <u>Nestability of Gaussian Quadrature Rules</u>: Gaussian quadrature rules do not nest, but the related Gauss-Kronrod quadrature formulas do.

## Generalized Mid-Point Rule Formulation

1.  <u>Generalized Mid-Point Rule Expression</u>: A generalized mid-point rule formula is given by

$$\int_{0}^{1} f(x)dx = \sum_{m=1}^{M} \sum_{n=0}^{\infty} \frac{(-1)^n + 1}{(2M)^{n+1}(n+1)!} \frac{d^n y}{dx^n} \bigg|_{x=\frac{m-\frac{1}{2}}{M}}$$

or

$$\int_0^1 f(x)dx = \underset{N \to \infty}{Limit} \sum_{m=1}^{M} \sum_{n=0}^{N} \frac{(-1)^n + 1}{(2M)^{n+1}(n+1)!} \frac{d^n y}{dx^n} \bigg\|_{x=\frac{m-\frac{1}{2}}{M}}$$

2. <u>Example Expression for Inverse Tangent</u>: For example, substituting

$$M = 1$$

and

$$f(x) = \frac{\theta}{1 + \theta^2 x^2}$$

in the generalized mid-point rule formula, one obtains the equation of the inverse tangent as

$$\tan^{-1} z = i \sum_{n=1}^{\infty} \frac{1}{2n-1} \left[ \frac{1}{\left(1 + \frac{2i}{z}\right)^{2n-1}} - \frac{1}{\left(1 - \frac{2i}{z}\right)^{2n-1}} \right]$$

$$= 2 \sum_{n=1}^{\infty} \frac{1}{2n-1} \frac{a_n(z)}{a_n{}^2(z) + b_n{}^2(z)}$$

where

$$i = \sqrt{-1}$$

is the imaginary unit, and

$$a_1(z) = \frac{2}{z}$$

$$b_1(z) = 1$$

$$a_n(z) = a_{n-1}(z)\left[1 - \frac{4}{z^2}\right] + 4\frac{b_{n-1}(z)}{z}$$

$$b_n(z) = b_{n-1}(z)\left[1 - \frac{4}{z^2}\right] - 4\frac{a_{n-1}(z)}{z}$$

3. <u>Eliminating the Series Odd Terms</u>: Since at each odd $n$ the numerator of the integrand becomes

$$(-1)^n + 1 = 0$$

the generalized mid-point rule formula can be re-organized as

$$\int_0^1 f(x)dx = 2 \sum_{m=1}^{M} \sum_{n=0}^{\infty} \frac{1}{(2M)^{2n+1}(2n+1)!} \frac{d^{2n}y}{dx^{2n}}\bigg|_{x=\frac{m-\frac{1}{2}}{M}}$$

4. <u>Transforming $(a, b)$ Limits into $(0, 1)$</u>: For a function $g(t)$ defined over the interval $(a, b)$ its integral is

$$\int_a^b g(t)dt = \int_0^{b-a} g(\tau + a)d\tau = (b - a) \int_0^1 g([b - a]x + a)dx$$

Therefore, the generalized mid-point integration formula above can be applied assuming that

$$f(x) = g([b - a]x + a)dx$$

## Adaptive Algorithms

1. <u>Lack of Suitable Derivative Points</u>: If $f(x)$ does not sufficient derivatives at all points, or if the derivatives become large, the generalized mid-point quadrature is often insufficient. In such cases, the adaptive algorithm similar to the one outlined in Wikipedia (2019) will perform better.

2. <u>Estimation of the Quadrature Error</u>: Some details of the algorithm require careful thought. For many cases, estimating the error from quadrature over an interval for the

function $f(x)$ is not obvious. One popular solution is to use two different rules for the quadrature, and use their difference as an estimate of the error from the quadrature.

3. <u>Too Large Versus Small Errors</u>: The other problem is deciding what *too large* or *very small* signify.

4. <u>Local Criterion for *Too Large*</u>: A *local* criterion for *too large* is that the quadrature error should not be larger than $t \cdot h$, where $t$, a real number, is the tolerance one wishes to set for the global error. However, if $h$ is too tiny, it may not be worthwhile to make it even smaller even if the quadrature error is apparently large.

5. <u>Global Criterion for *Too Large*</u>: A *global* criterion is that the sum of the errors on all the intervals should be less than $t$.

6. <u>*A-Posteriori* Type of Error Analysis</u>: This type of error analysis is typically called *a-posteriori* since the error is computed after having computed the approximation.

7. <u>Forsythe Heuristics for Adaptive Quadrature</u>: Heuristics for adaptive quadrature are discussed in Forsythe, Malcolm, and Moler (1977).

## Extrapolation Methods

1. <u>Error Dependence on Evaluation Point Count</u>: The accuracy of a quadrature rule of the Newton-Cotes type is generally a function of the number of evaluation points. The result is usually more accurate as the number of evaluation points increases, or, equivalently, the width of the step size between the points decreases.

2. <u>Error Dependence on Step Width</u>: It is natural to ask what the result would be if the step size were allowed to approach zero. This can be answered by extrapolating the result from two or more non-zero step sizes, using series acceleration methods such as Richardson extrapolation.

3. <u>Details of the Extrapolation Methods</u>: The extrapolation function may be either a polynomial or a rational function. Extrapolation methods are described in more detail

by Stoer and Bulirsch (1980) and are implemented in many of the routines in the QUADPACK library.

## A Priori Conservative Error Estimation

1. <u>Bounded First Derivative over $[a, b]$</u>: Let $f$ have a bounded first derivative over $[a, b]$, i. e.,

$$f \in \mathbb{C}^1(a, b)$$

The mean-value theorem for $f$ where

$$x \in (a, b]$$

gives

$$(x - a)\frac{df(\xi_x)}{dx} = f(x) - f(a)$$

for some

$$\xi_x \in (a, x]$$

depending on $x$.

2.  <u>Integrating Over $x$ on Both Sides</u>: On integrating in $x$ from $a$ to $b$ on both sides and taking the absolute values, one obtains

$$\left| \int_a^b f(x)dx - (b-a)f(a) \right| \leq \left| \int_a^b (x-a) \frac{df(v_x)}{dx} dx \right|$$

3.  <u>Approximating the Right-Hand Side</u>: The integral on the right-hand side can be further approximated by bringing the absolute value into the integrand, and replacing the term $\frac{df}{dx}$ by an upper bound:

$$\left| \int_a^b f(x)dx - (b-a)f(a) \right| \leq \frac{(b-a)^2}{2} \sup_{a \leq x \leq b} \left| \frac{df}{dx} \right|$$

where the supremum has been used for the approximation.

4.  <u>Corresponding Approximation Applied to the LHS</u>: Hence, if the integral $\int_a^b f(x)dx$ is approximated by the quadrature $(b-a)f(a)$, the error is no greater than the right-hand side above.

5.  <u>Converting the RHS into a Riemann Sum</u>: This can be converted into an error analysis for the Riemann sum, giving an upper bound of

$$\frac{n^{-1}}{2} \sup_{0 \le x \le 1} \left|\frac{df}{dx}\right|$$

for the error term of that particular approximation. Note that this precisely is the error obtained for the example

$$f(x) = x$$

6. <u>Strict Upper Bounds on the Error</u>: Using more derivatives, and by tweaking the quadrature, a similar analysis can be done using a Taylor series, using a partial sum with a remainder term, for $f$. This analysis gives a strict upper bound on the error, if the derivatives of $f$ are available.

7. <u>Algorithmic Proofs and Verified Calculations</u>: The integration method can be combined with interval arithmetic to produce computer proofs and *verified* calculations.

## Integrals Over Infinite Intervals

1. <u>Standard Techniques for Unbounded Intervals</u>: Several methods exist for approximate integration over unbounded intervals. The standard technique involves specially derived quadrature rules, such as Gauss-Hermitian quadrature for integrals on the whole real line, and Gauss-Laguerre quadrature for the integrals on the positive reals (Leader (2004)).

2. <u>Changing Variables to Bounded Intervals</u>: Monte Carlo methods can be used, or a change of variables to a finite interval can be applied; e.g., for the whole line, one could use

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-1}^{+1} f\left(\frac{1}{1-t^2}\right)\frac{1+t^2}{(1-t^2)^2} dt$$

and for semi-infinite intervals one could use

$$\int_{a}^{+\infty} f(x)dx = \int_{0}^{+1} f\left(a + \frac{t}{1-t}\right)\frac{1}{(1-t)^2} dt$$

$$\int_{-\infty}^{a} f(x)dx = \int_{0}^{+1} f\left(a - \frac{1-t}{t}\right)\frac{1}{t^2} dt$$

as possible transformations.

## Multi-dimensional Integrals

1. <u>Fubini's Theorem: Curse of Dimensionality</u>: The quadrature rules discussed so far are al designed to compute one-dimensional integrals. To compute integrals in multiple

dimensions, one approach is to phrase the multiple integrals as repeated one-dimensional integrals by applying the Fubini's theorem – the tensor product rule. This approach requires the function evaluations to grow exponentially as the number of dimensions increases. Three methods described below are known to overcome this so-called *curse of dimensionality*.

2. <u>Multi-dimensional Cubature Integration Rules</u>: A great many additional techniques for forming multi-dimensional cubature integration rules for a variety of weighting functions are given in Stroud (1971).

## Monte Carlo

1. <u>Potential Accuracy Improvement from MC</u>: Monte Carlo and quasi-Monte Carlo methods are easy to apply to multi-dimensional integrals. They may yield greater accuracy for the same number of function evaluations than repeated integrations using one-dimensional methods.

2. <u>Markov Chain Monte Carlo Algorithms</u>: A large class of useful Monte Carlo methods are the so-called Markov Chain Monte Carlo algorithms, which include the Metropolis-Hastings algorithms and Gibbs sampling.

## Sparse Grids

Sparse grids were originally developed by Smolyak for the quadrature of high-dimensional functions. The method is always based on a one-dimensional quadrature

rule, but performs a more sophisticated combination of univariate results. However, whereas the tensor product rule guarantees that the weight of all the quadrature points will be positive, Smolyak's rule does not guarantee that the weights will be positive.

## Bayesian Quadrature

Bayesian quadrature is a statistical approach to the numerical problem of computing integrals and falls under the field of probabilistic numerics. It can provide a full handling of the uncertainty over the range of the integral expressed as a Gaussian process posterior variance. It is also known to provide very fast convergence rates which can be up to exponential in the number of quadrature points $n$ (Briol, Oates, Girolami, and Osborne (2015)).

## Connections to Differential Equations

1.  The ODE Initial Value Problem: The problem of evaluating the integral

$$F(x) = \int_a^x f(u)du$$

can be reduced to an initial value problem for an ordinary differential equation by applying the first part of the fundamental theorem of calculus.

2. <u>Differential Form of the ODE</u>: By differentiating both sides of the above with respect to the argument $x$, it can be seen that the function $F$ satisfies

$$\frac{dF(x)}{dx} = f(x)$$

$$F(a) = 0$$

3. <u>Applying the ODE Solution Schemes</u>: Methods for ordinary differential equations, such as Runge-Kutta schemes, can be applied to the re-stated problem and thus be used to evaluate the integral. For instance, the standard fourth order Runge-Kutta method applied to the differential equation yields the Simpson's rule from above.

4. <u>Separating the Independent/Dependent Variables</u>: The differential equation

$$F(x) = f(x)$$

has a special form; the right-hand side contains only the dependent variable (here $x$) and not the independent variable (here $F$). This simplifies the theory and the algorithms considerably.

## References

- Briol, F. X., C. J. Oates, M. Girolami, and M. A. Osborne (2015): [Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees](#) **arXiv**

- Forsythe, G. E., M. A. Malcolm, and C. B. Moler (1977): *Computer Methods for Mathematical Computation* **Prentice Hall** Englewood Cliffs NJ

- Leader, J. J. (2004): *Numerical Analysis and Scientific Computation* **Addison Wesley**

- Stoer, J., and R. Bulirsch (1980): *Introduction to Numerical Analysis* **Springer-Verlag** New York

- Stroud, A. H. (1971): *Approximate Calculation of Multiple Integrals* **Prentice Hall** Englewood Cliffs NJ

- Wikipedia (2019): [Numerical Integration](#)

# Gaussian Quadrature

## Introduction and Overview

1.  <u>Quadrature Rule in Numerical Analysis</u>: In numerical analysis, a **quadrature rule** is the approximation of the definite integral of a function, usually stated as a weighted sum of the function values at the specified points within the domain of integration (Wikipedia (2019)).

2.  <u>n-Point Gaussian Quadrature Rule</u>: An n-point **Gaussian Quadrature Rule**, named after Carl Friedrich Gauss, is a quadrature rule constructed to yield the exact result for polynomials of degree $2n - 1$ or less using a suitable choice of nodes $x_i$ and weights $w_i$ for

$$i = 1, \cdots, n$$

3.  <u>Gaussian Quadrature Nodes and Weights</u>: The most common domain of integration for such a rule is taken as $[-1, +1]$, so the rule can be stated as

$$\int_{-1}^{+1} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

which is exact for polynomials of degree $2n - 1$ or less. This exact rule is known as the Gauss-Legendre quadrature rule.

4.  <u>Approximating $f(x)$ by a $2n - 1$ Polynomial</u>: The quadrature rule will only be an approximation to the integral above if $f(x)$ is well approximated by a polynomial of degree $2n - 1$ or less in $[-1, +1]$

5. Integrands with End-Point Singularities: The Gauss-Legendre quadrature rule is not typically used for integrable functions with end-point singularities.

6. Alternative Specification of the Quadrature Rules: Instead, of the integrand can be written as

$$f(x) = (1 - x)^\alpha (1 + x)^\beta g(x)$$

$$\alpha, \beta > -1$$

where $g(x)$ is well-approximated by a low-degree polynomial, then the alternative nodes $x_i'$ and weights $w_i'$ will usually give more accurate quadrature rules.

7. The Gauss-Jacobi Quadrature Rules: These are known as Gauss-Jacobi quadrature rules, i.e.,

$$f(x) = (1 - x)^\alpha (1 + x)^\beta g(x) \approx \sum_{i=1}^{n} w_i' f(x_i')$$

8. Gauss-Chebyshev Quadrature Weights: Common weights include $\frac{1}{\sqrt{1-x^2}}$ - referred to as Chebyshev-Gauss – and $\sqrt{1 - x^2}$.

9. Gauss-Laguerre and Gauss-Hermite Quadrature Rules: One may also want to integrate over semi-infinite intervals – the Gauss-Laguerre quadrature – or infinite intervals – the Gauss-Hermite quadrature.

10. Quadrature Nodes as Roots of Orthogonal Polynomials: It can be shown (Stoer and Bulirsch (2002), Press, Teukolsky, Vetterling, and Flannery (2007)) that the quadrature nodes $x_i$ are the roots of a polynomial belonging to a class of orthogonal polynomials, i.e., they belong to the class that is orthogonal with respect to a weighted inner-product. This is a key observation for computing Gauss quadrature nodes and weights.

# Gauss-Legendre Quadrature

1. <u>Legendre Polynomials as Associated Orthogonals</u>: For the simplest integration problem stated above, i.e., where $f(x)$ is well-approximated by polynomials in $[-1, +1]$, the associated orthogonal polynomials are Legendre polynomials, denoted by $P_n(x)$.

2. <u>Weights of the Legendre Terms</u>: With $n^{th}$ polynomial normalized to give

$$P_n(1) = 1$$

the $i^{th}$ Gauss node, $x_i$, is the $i^{th}$ root of $P_n$, and the weights are given by the formula (Abramowitz and Stegun (2007))

$$w_i = \frac{2}{(1 - x_i)^2 [P_n'(x_i)]^2}$$

3. <u>Low Order Nodes and Weights</u>: Some low order quadrature rules are tabulated below over $[-1, +1]$. The next section contains other intervals.

| Number of Points | Points $x_i$ | Approximate $x_i$ | Weight $w_i$ | Approximate $w_i$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 2 | 2 |
| 2 | $\pm\sqrt{\dfrac{1}{3}}$ | $\pm 0.57735$ | 1 | 1 |
| 3 | 0 | 0 | $\dfrac{8}{9}$ | 0.888889 |
| | $\pm\sqrt{\dfrac{3}{5}}$ | $\pm 0.774597$ | $\dfrac{5}{9}$ | 0.555556 |

| 4 | $\pm\sqrt{\dfrac{3}{7} - \dfrac{2}{7}\sqrt{\dfrac{6}{5}}}$ | $\pm 0.339981$ | $\dfrac{18 + \sqrt{30}}{36}$ | $0.652145$ |
| | $\pm\sqrt{\dfrac{3}{7} + \dfrac{2}{7}\sqrt{\dfrac{6}{5}}}$ | $\pm 0.861136$ | $\dfrac{18 - \sqrt{30}}{36}$ | $0.347855$ |
| 5 | $0$ | $0$ | $\dfrac{128}{225}$ | $0.568889$ |
| | $\pm\dfrac{1}{3}\sqrt{5 - 2\sqrt{\dfrac{10}{7}}}$ | $\pm 0.538469$ | $\dfrac{322 + 13\sqrt{70}}{900}$ | $0.478629$ |
| | $\pm\dfrac{1}{3}\sqrt{5 + 2\sqrt{\dfrac{10}{7}}}$ | $\pm 0.906180$ | $\dfrac{322 - 13\sqrt{70}}{900}$ | $0.236927$ |

## Change of Interval

1. <u>Interval Change - $[a, b]$ to $[-1, +1]$</u>: The integral over $[a, b]$ must be changed to the integral over $[-1, +1]$ before applying the Gaussian quadrature rule. The change of interval can be done in the following way:

$$\int_a^b f(t)dt = \frac{b-a}{2} \int_{-1}^{+1} f\left(\frac{b-a}{2}x + \frac{b+a}{2}\right)dx$$

2. <u>Application of Gaussian Quadrature Rules</u>: Applying the Gaussian quadrature rule then results in the following approximation:

$$\int_a^b f(t)dt \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f\left(\frac{b-a}{2}x_i + \frac{b+a}{2}\right)$$

## Other Forms

1. <u>Generalization of the Integrand Quadrature</u>: The integration problem can be expressed in a slightly more general way by introducing a positive weight function $\omega$ into the integrand, and allowing an interval other than $[-1, +1]$. That is, the problem is to calculate $\int_a^b \omega(x)f(x)dx$ for some choices of $a$, $b$, $c$, and $\omega$.

2. <u>Parameter Choices for the Quadrature Generation</u>: For

$$a = -1$$

$$b = 1$$

and

$$\omega(x) = 1$$

the problem is the same as the one considered above. Other choices lead to other integration rules, some of which are tabulated below.

| Interval | $\omega(x)$ | Orthogonal Polynomials |
|---|---|---|
| $[-1, +1]$ | 1 | Legendre Polynomials |
| $(-1, +1)$ | $(1-x)^\alpha(1+x)^\beta, \alpha, \beta > -1$ | Jacobi Polynomials |
| $(-1, +1)$ | $\dfrac{1}{\sqrt{1-x^2}}$ | Chebyshev Polynomials (First Kind) |

85

| | | |
|---|---|---|
| $[-1, +1]$ | $\sqrt{1 - x^2}$ | Chebyshev Polynomials (Second Kind) |
| $[0, \infty)$ | $e^{-x}$ | Laguerre Polynomials |
| $[0, \infty)$ | $x^\alpha e^{-x}, \alpha > -1$ | Generalized Laguerre Polynomials |
| $(-\infty, +\infty)$ | $e^{-x^2}$ | Hermite Polynomials |

## Fundamental Theorem

1. n-Degree Polynomial Driving the Quadrature: Let $p_n$ be a non-trivial polynomial of degree $n$ such that

$$\int_a^b \omega(x) x^k p_n(x) dx = 0$$

for all

$$k = 0, \cdots, n - 1$$

2. Nodes as Roots of the Polynomial: If the $n$ nodes $x_i$ are picked to be the zeros of $p_n$, then there exist $n$ weights $w_i$ which make the Gauss quadrature computed integral exact for polynomials $h(x)$ of degree $2n - 1$ or less. Furthermore, all these nodes $x_i$ lie in the open interval $(a, b)$ (Stoer and Bulirsch (2002)).

3. Chosen as the Orthogonal Polynomial: The polynomial $p_n$ is said to be an orthogonal polynomial of degree $n$ associated with the weight function $\omega(x)$. It is unique to a constant normalization factor.

4. Decomposing $h(x)$ into Orthogonal Polynomials: The idea underlying the proof is that, because of its sufficiently low degree, $h(x)$ can be divided by $p_n(x)$ to produce a quotient $q(x)$ strictly lower than $n$, and a remainder of still lower degree, so that both will be orthogonal to $p_n(x)$, by the defining property of $p_n(x)$. Thus,

$$\int_a^b \omega(x)h(x)dx = \int_a^b \omega(x)r(x)dx$$

5. Quadrature Over Reduced Degree Polynomial: Because of the choice of nodes $x_i$, the corresponding relation

$$\sum_{i=1}^n w_i h(x_i) = \sum_{i=1}^n w_i r(x_i)$$

also holds. The exactness of the computed integral $h(x)$ then follows from the exactness of $r(x)$, i.e., for polynomials of degree $n$ or less.

## General Formula for the Weights

1. Generic Expression for the Quadrature Weights: The weights can be expressed as

$$w_i = \frac{a_n}{a_{n-1}} \frac{\int_a^b \omega(x)p_{n-1}{}^2(x)dx}{p_n{}'(x_i)p_{n-1}(x_i)}$$

where $a_k$ is the coefficient of $x^k$ in $p_n(x)$.
2. Lagrange Polynomial Form for $r(x)$: To prove this, note that using the Lagrange interpolation, one can express $r(x)$ in terms of $r(x_i)$ as

$$r(x) = \sum_{i=1}^{n} r(x_i) \prod_{\substack{1 \le j \le n \\ j \ne i}} \frac{x - x_j}{x_i - x_j}$$

because $r(x)$ has a degree less than $n$ and is thus fixed by the value it attains at $n$ different points.

3. <u>Re-evaluation of the $r(x)$ Quadrature</u>: Multiplying both sides by $\omega(x)$ and integrating from $a$ to $b$ yields

$$\int_a^b \omega(x) r(x) dx = \sum_{i=1}^{n} r(x_i) \int_a^b \omega(x) \prod_{\substack{1 \le j \le n \\ j \ne i}} \frac{x - x_j}{x_i - x_j} dx$$

4. <u>Quadrature Weights Using Lagrange Polynomials</u>: The weights $w_i$ are thus given by

$$w_i = \int_a^b \omega(x) \prod_{\substack{1 \le j \le n \\ j \ne i}} \frac{x - x_j}{x_i - x_j} dx$$

5. <u>Quadrature Weights Using Orthogonal Polynomials</u>: This integral expression for $w_i$ can be expressed in terms of the orthogonal polynomials $p_n(x)$ and $p_{n-1}(x)$ as follows.

6. <u>Orthogonal Polynomial from Lagrange Numerator</u>: One can write

$$\prod_{\substack{1 \le j \le n \\ j \ne i}} (x - x_j) = \frac{\prod_{1 \le j \le n}(x - x_j)}{x - x_i} = \frac{p_n(x)}{a_n(x - x_i)}$$

where $a_n$ is the coefficient of $x^n$ in $p_n(x)$.

7. <u>Orthogonal Derivative from Lagrange Denominator</u>: Taking the limit of $x$ to $x_i$ yields, using L'Hopital's rule

$$\prod_{\substack{1 \le j \le n \\ j \ne i}} (x_i - x_j) = \frac{p_n'(x_i)}{a_n}$$

8. <u>Weights from Polynomials and Derivatives</u>: The integral expression for the weights can thus be written as

$$w_i = \frac{1}{p_n'(x_i)} \int_a^b \omega(x) \frac{p_n(x)}{x - x_i} dx$$

9. <u>Reducing the Degree of Integrand</u>: In the integrand, writing

$$\frac{1}{x - x_i} = \frac{1 - \left(\frac{x}{x_i}\right)^k}{x - x_i} + \left(\frac{x}{x_i}\right)^k \frac{1}{x - x_i}$$

yields

$$\int_a^b \omega(x) x^k \frac{p_n(x)}{x - x_i} dx = x_i^k \int_a^b \omega(x) \frac{p_n(x)}{x - x_i} dx$$

provided

$$k \le n$$

because $\dfrac{1 - \left(\frac{x}{x_i}\right)^k}{x - x_i}$ is a polynomial of degree $k - 1$ which is then orthogonal to $p_n(x)$.

10. <u>Product of Lower Degree Polynomials</u>: So, if $t(x)$ is a polynomial of at most degree $n$, one has

$$\int_a^b \omega(x) \frac{p_n(x)}{x - x_i} dx = \frac{1}{t(x_i)} \int_a^b \omega(x) \frac{t(x)p_n(x)}{x - x_i} dx$$

11. $\frac{p_n(x)}{x-x_i}$ as $n-1$ Degree Polynomial: The integral on the right-hand side can be evaluated for

$$t(x) = p_{n-1}(x)$$

as follows. Because $\frac{p_n(x)}{x-x_i}$ is a polynomial of degree $n-1$, one has

$$\frac{p_n(x)}{x - x_i} = a_n x^{n-1} + s(x)$$

where $s(x)$ is a polynomial of degree $n-2$.

12. $n-1$ Polynomial in the Weight Integral: Since $s(x)$ is orthogonal to $p_{n-1}(x)$, one has

$$\int_a^b \omega(x) \frac{p_n(x)}{x - x_i} dx = \frac{a_n}{p_{n-1}(x_i)} \int_a^b \omega(x) p_{n-1}(x) x^{n-1} dx$$

13. $x^{n-1}$ in Terms of $p_{n-1}$: One can then write

$$x^{n-1} = \left[ x^{n-1} - \frac{p_{n-1}(x)}{a_{n-1}} \right] + \frac{p_{n-1}(x)}{a_{n-1}}$$

The term in the brackets is a polynomial of degree $n-2$, which is therefore orthogonal to $p_{n-1}(x)$.

14. Weight Integral in Terms of $p_{n-1}^2$: The integral can thus be written as

90

$$w_i = \frac{a_n}{a_{n-1}p_{n-1}(x_i)} \int_a^b \omega(x)p_{n-1}{}^2(x)dx$$

15. <u>Recovery of the Postulated Expression</u>: According to

$$w_i = \frac{1}{p_n{}'(x_i)} \int_a^b \omega(x)\frac{p_n(x)}{x - x_i}dx$$

the weights are obtained by dividing the weight integral above by $p_n{}'(x_i)$, and that yields the expression

$$w_i = \frac{a_n}{a_{n-1}} \frac{\int_a^b \omega(x)p_{n-1}{}^2(x)dx}{p_n{}'(x_i)p_{n-1}(x_i)}$$

16. <u>Alternate Expression Using the $p_{n+1}$ Term</u>: $w_i$ can also be expressed in terms of the orthogonal polynomials $p_n(x)$ and $p_{n+1}(x)$. In a 3-term recurrence relation (see below)

$$p_{n+1}(x_i) = \rho_a p_n(x_i) + \rho_b p_{n-1}(x_i)$$

the $p_n(x_i)$ term vanishes, so $p_{n-1}(x_i)$ in

$$w_i = \frac{a_n}{a_{n-1}} \frac{\int_a^b \omega(x)p_{n-1}{}^2(x)dx}{p_n{}'(x_i)p_{n-1}(x_i)}$$

can be replaced by $\frac{p_{n+1}(x_i)}{\rho_b}$.

## Proof that the Weights are Positive

1. <u>Remainder – Lagrange Squared Term Polynomial</u>: Consider the following polynomial of degree $2n - 2$

$$l(x) = \prod_{\substack{1 \leq j \leq n \\ j \neq i}} \left( \frac{x - x_j}{x_i - x_j} \right)^2$$

   where, as above, the $x_j$ are the roots of the polynomial $p_n(x)$.

2. <u>Gaussian Quadrature for the above Polynomial</u>: Clearly

$$l(x_j) = \delta_{ij}$$

   Since the degree $l(x)$ is less than $2n - 1$, the Gaussian quadrature formula involving the weights and the nodes obtained from $p_n(x)$ applies.

3. <u>Weights Have to be Positive</u>: Since

$$l(x_j) = 0$$

   for $j$ not equal to $i$, one has

$$\int_a^b \omega(x)l(x)dx = \sum_{j=1}^N w_j l(x_j) = \sum_{j=1}^N w_j \delta_{ij} = w_i > 0$$

   Since both $\omega(x)$ and $l(x)$ are non-negative functions, it follows that

$$w_i > 0$$

## Computation of Gaussian Quadrature Rules

There are many algorithms for computing the nodes $x_i$ and the weights $w_i$ of Gaussian quadrature rules. The popular are the Golub-Welsch algorithm requiring $\mathcal{O}(n^2)$ operations, Newton's method for solving

$$p_n(x) = 0$$

using the three-term recurrence relation for evaluation requiring $\mathcal{O}(n^2)$ operations, and asymptotic formulas for large $n$ requiring $\mathcal{O}(n)$ operations.

## Recurrence Relation

1. Recurrence Relation for Orthogonal Polynomials: Orthogonal polynomials $p_r$ with

$$[\![p_r, p_s]\!] = 0$$

for

$$r \neq s$$

for a scalar product $[\![\cdot]\!]$,

$$Degree(p_r) = r$$

and leading coefficient one – i.e., monic orthogonal polynomials – satisfy the recurrence relation

$$p_{r+1}(x) = (x - a_{r,r})p_r(x) - a_{r,r-1}p_{r-1}(x) - \cdots - a_{r,0}p_0(x)$$

where the scalar product is defined as

$$[\![p_r(x), p_s(x)]\!] = \int_a^b \omega(x)p_r(x)p_s(x)dx$$

for

$$r = 0, \cdots, n - 1$$

where $n$ is the upper bound on the degree – which can be taken to infinity – and where

$$a_{r,s} = \frac{[\![xp_r, p_s]\!]}{[\![p_r, p_s]\!]}$$

2. <u>Proof of Recurrence using Induction</u>: First of all, the polynomials defined by the recurrence relation starting with

$$p_0(x) = 1$$

have a leading coefficient of one and the correct degree (i.e. 0). Setting the starting polynomial by $p_0$, the orthogonality of $p_r$ can be demonstrated by induction.

3. <u>Proof for the First Term</u>: For

$$r = s = 0$$

one has

$$[\![p_1, p_0]\!] = (x - a_{0,0})[\![p_0, p_0]\!] = [\![xp_0, p_0]\!] - a_{0,0}[\![p_0, p_0]\!] = [\![xp_0, p_0]\!] - [\![xp_0, p_0]\!]$$
$$= 0$$

4. <u>Proof for an Arbitrary Term</u>: Now, if $p_0, \cdots, p_r$ are orthogonal, so is $p_{r+1}$, because in

$$[\![p_{r+1}, p_s]\!] = [\![xp_r, p_s]\!] - a_{r,r}[\![p_r, p_s]\!] - a_{r,r-1}[\![p_{r-1}, p_s]\!] - \cdots - a_{r,0}[\![p_0, p_0]\!]$$

all scalar products vanish except for the first one and the one where $p_s$ meets the same orthogonal polynomial. Therefore,

$$[\![p_{r+1}, p_s]\!] = [\![xp_r, p_s]\!] - a_{r,s}[\![p_r, p_s]\!] = [\![xp_r, p_s]\!] - [\![xp_r, p_s]\!] = 0$$

5. <u>Reduction to Three Term Recurrence</u>: However, if the scalar product satisfies

$$[\![xp_r, p_s]\!] = [\![p_r, xp_s]\!]$$

- which is the case for Gaussian Quadrature – the above recurrence relation reduces to a three-term recurrence relation.

6. <u>Serial Zeroing of Recurrence Terms</u>: For

$$s < r - 1$$

$xp_s$ is a polynomial of degree less than or equal to $r - 1$. On the other hand, $p_r$ is orthogonal to every polynomial of degree less than or equal to $r - 1$. Therefore, one has

$$[\![xp_r, p_s]\!] = [\![p_r, xp_s]\!] = 0$$

and

$$a_{r,s} = 0$$

for

$$s < r - 1$$

7. <u>Full Form of the Recurrence Relation</u>: The recurrence relation then simplifies to

$$p_{r+1}(x) = (x - a_{r,r})p_r(x) - a_{r,r-1}p_{r-1}(x)$$

or, with the convention

$$p_{-1}(x) \equiv 0$$

$$p_{r+1}(x) = (x - a_r)p_r(x) - b_r p_{r-1}(x)$$

where

$$a_r \doteq \frac{[\![xp_r, p_r]\!]}{[\![p_r, p_r]\!]}$$

$$b_r \doteq \frac{[\![xp_r, p_{r-1}]\!]}{[\![p_{r-1}, p_{r-1}]\!]} = \frac{[\![p_r, p_r]\!]}{[\![p_{r-1}, p_{r-1}]\!]}$$

where the final step occurs because

$$[\![xp_r, p_{r-1}]\!] = [\![p_r, xp_{r-1}]\!] = [\![p_r, p_r]\!]$$

since $xp_r$ differs from $p_r$ by a degree less than $r$.

## The Golub-Welsch Algorithm

1. <u>Three-Term Jacobi Recurrence Matrix</u>: The three-term recurrence relation can be written in the matrix form

$$J\tilde{P} = x\tilde{P} - P_n(x) \times \hat{e}_n$$

where

$$\tilde{P} = [P_0(x), \cdots, P_{n-1}(x)]^T$$

$\hat{e}_n$ is the $n^{th}$ standard basis vector, i.e.

$$\hat{e}_n = [0, \cdots, 0, 1]^T$$

and $J$ is the so-called Jacobi matrix.

$$J = \begin{bmatrix} a_0 & 1 & 0 & \cdots & \cdots & \cdots \\ b_1 & a_1 & 1 & 0 & \cdots & \cdots \\ 0 & b_2 & a_2 & 1 & 0 & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & 0 & b_{n-2} & a_{n-2} & 1 \\ \cdots & \cdots & \cdots & 0 & b_{n-1} & a_{n-1} \end{bmatrix}$$

2. <u>The Golub-Welsch Algorithm</u>: The zeroes $x_j$ of the polynomials upto degree $n$, which are used as nodes for the Gaussian quadrature, can be found by computing the eigenvalues of this tri-diagonal matrix. This procedure is known as the Golub-Welsch algorithm.

3. <u>Elements of the Tri-diagonal Matrix</u>: For computing the weights and the nodes, it is preferable to consider the symmetric tridiagonal matrix $\mathcal{J}$ with elements

$$\mathcal{J}_{i,i} = J_{i,i} = a_{i-1}$$

$$i = 1, \cdots, n$$

$$\mathcal{J}_{i-1,i} = \mathcal{J}_{i,i-1} = \sqrt{J_{i,i-1}J_{i-1,i}} = \sqrt{b_{i-1}}$$

$$i = 2, \cdots, n$$

4. <u>Quadrature Nodes from the Eigenvalues</u>: $J$ and $\mathcal{J}$ are similar matrices, and therefore have the same eigenvalues – the quadrature nodes.

5. <u>Quadrature Weights Extracted from Eigenvectors</u>: The weights can be computed from the corresponding eigenvectors. If $\phi_j$ is a normalized eigenvector – i.e., an eigenvector with Euclidean norm equal to one – associated with the eigenvalue $x_j$, the corresponding weight can be computed from the first component of this eigenvector, namely:

$$w_j = \mu_0 \left[ \phi_j(1) \right]^2$$

where $\mu_0$ is the integral of the weight function

$$\mu_0 = \int_a^b \omega(x)dx$$

Gil, Segura, and Temme (2007) contain further details.

**Error Estimates**

1. <u>Stoer-Bulirsch (2002) Error Estimate</u>: Using the analysis presented in Stoer and Bulirsch (2002), the error of a Gaussian quadrature can be stated as follows. For an integrand that has $2n$ continuous derivatives,

$$\int_a^b \omega(x)f(x)dx - \sum_{i=1}^n w_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n)!} [\![p_n, p_n]\!]$$

for some $\xi$ in $(a, b)$, where $p_n$ is the monic orthogonal polynomial of degree $n$, and

$$[\![f(x), g(x)]\!] = \int_a^b \omega(x)f(x)g(x)dx$$

2. <u>Kahaner, Moler, and Nash (1989) Error Estimate</u>: In the important special case of

$$\omega(x) = 1$$

one has the error estimate (Kahaner, Moler, and Nash (1989))

$$\int_a^b \omega(x)f(x)dx - \sum_{i=1}^n w_i f(x_i) = \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi)$$

$$a < \xi < b$$

3. <u>Conservative Nature of the Estimate</u>: Stoer and Bulirsch (2002) remark that this error estimate is inconvenient in practice, since it may be difficult to estimate the order $2n$ derivatives. Furthermore, the actual error may be much less that the bound established by the derivative.

4. <u>Error Estimate Using Different Orders</u>: Another approach is to use two Gaussian quadrature rules of different orders, and to estimate the error as the difference

between these two results. For this purpose, the Gauss-Kronrod quadrature rules can be useful.

## Gauss-Kronrod Rules

1. <u>Sub-divided Points do not Coincide</u>: If the interval $[a, b]$ is sub-divided, the Gaussian evaluation points of the new sub-interval never coincide with the previous evaluation points – except at zero for odd numbers – and thus the integrand must be evaluated at every point.

2. <u>Extensions to Gauss Quadrature Rules</u>: *Gauss-Kronrod rules* are extensions to Gauss quadrature rules generated by adding $n + 1$ points to a $n$-point rule in such a way that the resulting rule is of the order $2n + 1$. This allows for computing higher-order estimates while re-using the function values of the lower-order estimates.

3. <u>Estimation of the Quadrature Error</u>: The difference between the Gauss quadrature rule and its Kronrod extension is often used an estimate of the approximation error.

## Gauss-Lobatto Rules

1. <u>Gaussian vs. Lobatto Quadrature Rules</u>: Also known as *Lobatto quadrature* (Abramowitz and Stegun (2007)), Gauss-Lobatto quadrature is named after the Dutch mathematician Rehuel Lobatto. It is similar to the Gaussian quadrature, except for the following differences:
   a. The integration points include the end-points of the integration interval.
   b. It is accurate for polynomials up to degree $2n - 3$, where $n$ is the number of integration points (Quatteroni, Sacco, and Saleri (2000)).

2. <u>Gauss-Lobatto Quadrature Function Expression</u>: Lobatto quadrature of function $f(x)$ in an interval $[-1, +1]$ is

$$\int_{-1}^{+1} f(x)dx = \frac{2}{n(n-1)}[f(-1) + f(+1)] + \sum_{i=2}^{N} w_i f(x_i) + R_N$$

3. Gauss-Lobatto Quadrature Abscissa: The abscissae $x_i$ is the $(i-1)^{st}$ zero of $P_{n-1}'(x)$.

4. Gauss-Lobatto Quadrature Weights:

$$w_i = \frac{2}{n(n-1)[P_{n-1}(x_i)]^2}$$

$$x_i \neq \pm 1$$

5. Gauss-Lobatto Quadrature Error Remainder:

$$R_N = \frac{-n(n-1)^3 2^{2n-1}[(n-2)!]^4}{(2n-1)[(2n-2)!]^3} f^{(2n-2)}(\xi)$$

$$-1 < \xi < +1$$

6. Low Order Gauss-Lobatto Quadrature Weight Sample:

| Number of Points $n$ | Points $x_i$ | Weights $w_i$ |
|---|---|---|
| 3 | 0 | $\frac{4}{3}$ |
| | $\pm 1$ | $\frac{1}{3}$ |
| 4 | $\pm\sqrt{\frac{1}{5}}$ | $\frac{5}{6}$ |
| | $\pm 1$ | $\frac{1}{6}$ |

| | | |
|---|---|---|
| 5 | $0$ | $\dfrac{32}{45}$ |
| | $\pm\sqrt{\dfrac{3}{7}}$ | $\dfrac{49}{90}$ |
| | $\pm 1$ | $\dfrac{1}{10}$ |
| 6 | $\pm\sqrt{\dfrac{1}{3} - \dfrac{2\sqrt{7}}{21}}$ | $\dfrac{14 + \sqrt{7}}{30}$ |
| | $\pm\sqrt{\dfrac{1}{3} + \dfrac{2\sqrt{7}}{21}}$ | $\dfrac{14 - \sqrt{7}}{30}$ |
| | $\pm 1$ | $\dfrac{1}{15}$ |
| 7 | $0$ | $\dfrac{256}{525}$ |
| | $\pm\sqrt{\dfrac{5}{11} - \dfrac{2}{11}\sqrt{\dfrac{5}{3}}}$ | $\dfrac{124 + 7\sqrt{15}}{350}$ |
| | $\pm\sqrt{\dfrac{5}{11} + \dfrac{2}{11}\sqrt{\dfrac{5}{3}}}$ | $\dfrac{124 - 7\sqrt{15}}{350}$ |
| | $\pm 1$ | $\dfrac{1}{21}$ |

7. <u>Implementation of the Adaptive Variant</u>: An adaptive variant of this algorithm with 2 interior nodes (Gander and Gautschi (2000)) is found in GNU Octave and in MATLAB as *quadl* and *intergate*, respectively.

**References**

- Abramowitz, M., and I. A. Stegun (2007): *Handbook of Mathematics Functions* **Dover Book on Mathematics**

- Gander, W., and W. Gautschi (2000): Adaptive Quadrature – Revisited *Bit Numerical Mathematics* **40 (1)** 84-101

- Gil, A., J. Segura, and N. M. Temme (2007): *Numerical Methods for Special Functions* **Society for Industrial and Applied Mathematics** Philadelphia

- Kahaner, D., C. Moler, and S. Nash (1989): *Numerical Methods and Software* **Prentice Hall**

- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007): *Numerical Recipes: The Art of Scientific Computing 3rd Edition* **Cambridge University Press** New York

- Quatteroni, A., R. Sacco, and F. Saleri (2000): *Numerical Mathematics* **Springer Verlag** New York

- Stoer, J., and R. Bulirsch (2002): *Introduction to Numerical Analysis 3rd Edition* **Springer**

- Wikipedia (2019): [Gaussian Quadrature](#)

# Gauss-Kronrod Quadrature

## Introduction and Overview

1. <u>Adaptive Method for Numerical Integration</u>: The *Gauss-Kronrod quadrature formula* is an adaptive method for numerical integration.
2. <u>Improved Accuracy by Re-using Less Accurate Results</u>: It is a variant of the Gaussian quadrature, in which the evaluation points are chosen so that an accurate approximation can be computed by re-using the information produced by the computation of a less accurate approximation (Wikipedia (2019)).
3. <u>Multi-Order Quadrature Rules and Errors</u>: It is an example of what is called a nested quadrature rule; for the same set of function evaluation points, it has two quadrature rules, one higher order and one lower order – the latter is called an *embedded* rule. The difference between these two approximations is used to estimate the calculation error of the integral.

## Description

1. <u>Numerical Approximation of Definite Integrals</u>: The problem in numerical integration is to approximate $\int_a^b f(x)dx$
2. <u>Use of n Point Gaussian Quadrature</u>: Such integrals can be approximated, for example, by n-point Gaussian quadrature

$$\int_a^b f(x)dx \approx \sum_{i=1}^n w_i f(x_i)$$

where $x_i$ and $w_i$ are the points and the weights used to evaluate the function $f(x)$.

3. <u>Consequences of Non-Matching Nodes</u>: If the interval $[a, b]$ is sub-divided, the Gauss evaluation points of the new sub-intervals never coincide with the previous evaluation points – except at the mid-point for odd numbers of evaluation points – and thus the integrand must be evaluated at every point.

4. <u>Node Extensions using Stieltjes Polynomials</u>: Gauss-Kronrod formulas are extensions of the Gauss quadrature formulas generated by adding $n + 1$ points to a $n$ point rule in such a way that the resulting rule is of order $2n + 1$ (Laurie (1997)); the corresponding Gauss rule is order $2n - 1$. The extra points are zeros of Stieltjes polynomials.

5. <u>Kronrod Extension as an Error Estimate</u>: This allows for computing higher-order error estimates while re-using the function values of a lower order estimate. The difference between a Gauss quadrature rule and its Kronrod extension is used often as an estimate of the approximation error.

## Example

1. <u>7 Point Gauss Plus 15 Point Kronrod</u>: A popular example combines a 7-point Gauss rule with a 15-point Kronrod rule (Kahaner, Moler, and Nash (1989)). Because the Gauss points are incorporated into the Kronrod points, a total of only 15 function evaluations are needed.

2. <u>$(G7, K15)$ on $[-1, +1]$</u>:

| Gauss Nodes | Weights |
|---|---|
| $\pm$0.94910 79123 42759 | 0.12948 49661 68870 |
| $\pm$0.74153 11855 99394 | 0.27970 53914 89277 |
| $\pm$0.40584 51513 77397 | 0.38183 00505 05119 |
| $\pm$0.00000 00000 00000 | 0.41795 91836 73469 |

| Kronrod Nodes | Weights |
|---|---|
| ±0.99145 53711 20813 | 0.02293 53220 10529 |
| ±0.94910 79123 42759 | 0.06309 20926 29979 |
| ±0.86486 44233 59769 | 0.10479 00103 22250 |
| ±0.74153 11855 99394 | 0.14065 32597 15525 |
| ±0.58608 72354 67691 | 0.16900 47266 39267 |
| ±0.40584 51513 77397 | 0.19035 05780 64785 |
| ±0.20778 49550 07898 | 0.20443 29400 75298 |
| ±0.00000 00000 00000 | 0.20948 21410 84728 |

3. <u>Use in Quadrature Error Estimate</u>: The integral is then estimated by the Kronrod rule $K15$ and the error can be estimated as $|G7 - K15|$.

4. <u>Enhancements to the Quadrature Algorithm</u>: Patterson (1968) showed how to find further extensions of this type. Monegato (1978) and Piessens, de Doncker-Kapenga, Uberhuber, and Kahaner (1983) proposed improved algorithms. Finally, the most efficient algorithm was proposed by Laurie (1997).

5. <u>Tabulation of Quadruple Precision Coefficients</u>: Quadruple Precision (34 decimal digits) coefficients for $(G7, K15)$, $(G10, K21)$, $(G15, K31)$, $(G20, K41)$, and others are computed and tabulated in Holoborodko (2011).

## Implementations

Routines for Gauss-Kronrod quadrature are provided by the QUADPACK library, the GNU Scientific Library, the NAG Numerical Libraries R, the C++ Boost Library, and DROP.

## References

- Holoborodko, P. (2011): [Gauss-Kronrod Quadrature Nodes and Weights](#)

- Kahaner, D., C. Moler, and S. Nash (1989): *Numerical Methods and Software* **Prentice Hall**

- Laurie, D. (1997): Calculation of Gauss-Kronrod Quadrature Rules *Mathematics of Computation* **66 (219)** 1133-1145

- Monegato, G. (1978): Some Remarks on the Construction of Extended Gaussian Quadrature Rules *Mathematics of Computation* **32 (141)** 247-252

- Patterson, T. N. L. (1968): The Optimum Addition of Points to Quadrature Formulae *Mathematics of Computation* **22 (104)** 847-856

- Piessens, R., E. de Doncker-Kapenga, C. W. Uberhuber, and D. K. Kahaner (1983): *QUADPACK – A Subroutine Package for Automatic Integration* **Springer-Verlag**

- Wikipedia (2019): [Gauss-Kronrod Quadrature Formula](#)

# Gamma Distribution

## Overview

1. <u>Parametrization of the Gamma Distribution</u>: The *gamma distribution* is a two-parameter family of continuous probability distributions. The Erlang distribution, the exponential distribution, and the chi-square distribution are all special cases of the gamma distribution. There are three different parameterizations in common use:

   - With a shape parameter $k$ and a scale parameter $\theta$
   - With a shape parameter

$$k = \alpha$$

   and an inverse scale parameter

$$\beta = \frac{1}{\theta}$$

   called a rate parameter
   - With a shape parameter $k$ and a mean parameter

$$\mu = k\theta = \frac{\alpha}{\beta}$$

   In each of these forms, both parameters are positive and real numbers (Wikipedia (2019)).

2. <u>As a Constrained Maximum Entropy Distribution</u>: The gamma distribution is a maximum entropy probability distribution – both with respect to a uniform base measure and with respect to a $\frac{1}{x}$ base measure – for a random variable $X$ for which

$$\mathbb{E}[X] = k\theta = \frac{\alpha}{\beta}$$

is fixed and greater than zero, and

$$\mathbb{E}[\log X] = \psi(k) + \log\theta = \psi(\alpha) - \log\beta$$

is fixed. Here $\psi$ is the digamma function (Park and Bera (2009)).

## Gamma Distribution – Central Measures Table

| Parameters | $k > 0$ *Shape* <br><br> $\theta > 0$ *Scale* | $\alpha > 0$ *Shape* <br><br> $\beta > 0$ *Rate* |
|---|---|---|
| **Support** | $x \in (0, \infty)$ | $x \in (0, \infty)$ |
| **PDF** | $\dfrac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ | $\dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ |
| **CDF** | $\dfrac{1}{\Gamma(k)} \gamma\left(k, \dfrac{x}{\theta}\right)$ | $\dfrac{1}{\Gamma(\alpha)} \gamma(k, \beta x)$ |
| **Mean** | $k\theta$ | $\dfrac{\alpha}{\beta}$ |

| Median | No Closed Form | No Closed Form |
|---|---|---|
| Mode | $(k-1)\theta$ for $k \geq 1$ | $\frac{\alpha-1}{\beta}$ for $\alpha \geq 1$ |
| Skewness | $\dfrac{2}{\sqrt{k}}$ | $\dfrac{2}{\sqrt{\alpha}}$ |
| Excess Kurtosis | $\dfrac{6}{k}$ | $\dfrac{6}{\alpha}$ |
| Entropy | $k + \log\theta + \log\Gamma(k) + (1-k)\psi(k)$ | $\alpha - \log\beta + \log\Gamma(\alpha) + (1-\alpha)\psi(\alpha)$ |
| MGF | $(1-\theta t)^{-k}$ for $t < \frac{1}{\theta}$ | $\left(1-\frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$ |
| CDF | $(1-i\theta t)^{-k}$ | $\left(1-i\dfrac{t}{\beta}\right)^{-k}$ |

## Definitions

1. <u>Usage of the $k/\theta$ Parametrization</u>: The parametrization using $k$ and $\theta$ appeears to be more common in econometrics and other applied fields where, for example, the gamma distribution is frequently used to model waiting times. For instance, in life testing, the waiting time until death is a random variable that is frequently modeled with a gamma distribution (Hogg, McKean, and Craig (2013)).

2. <u>Usage of the $\alpha/\beta$ Parametrization</u>: The parametrization with $\alpha$ and $\beta$ is more common in Bayesian statistics, where the gamma distribution is used as a conjugate prior for various types of inverse scale – aka rate – parameters, such as the $\lambda$ of the exponential or the Poisson distribution (Gopalan, Hofman, and Blei (2014)) – or, for that matter, the $\beta$ of the gamma distribution itself. The closely related inverse gamma

distribution is used as a conjugate prior for the scale parameters, such as the variance of a normal distribution.

3. Erlang Distribution - $k$ Dependent Exponentials: If $k$ is a positive integer, then the distribution represents an Erlang distribution, i.e., the sum of $k$ independent exponentially distributed random variables, each of which has a mean of $\theta$.

## Characterization Using Shape $\alpha$ and Rate $\beta$

1. $\alpha/\beta$ Parameterized Gamma Distribution Representation: The gamma distribution can be represented in terms of a shape parameter

$$\alpha = k$$

and an inverse scale parameter

$$\beta = \frac{1}{\theta}$$

called a rate parameter. A random variable $X$ that is gamma distributed with shape $\alpha$ and rate $\beta$ is denoted

$$X \sim \Gamma(\alpha, \beta) \equiv Gamma(\alpha, \beta)$$

2. $\alpha/\beta/\theta$ Parameterized Gamma Distribution PDF: The corresponding probability distribution in the shape-rate parametrization is

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

for

$$x > 0$$

and

$$\alpha, \beta > 0$$

$$f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}$$

for

$$x > 0$$

and

$$\alpha, \beta > 0$$

Here $\Gamma(\alpha)$ is the gamma function. Both parametrizations are common because either can be more convenient depending on the situation.

3. _$\alpha/\beta/\theta$ Parameterized Gamma Distribution CDF_: The cumulative distribution function is the regularized gamma function

$$F(x;\ \alpha, \beta) = \int_0^u f(u;\ \alpha, \beta) du = \frac{1}{\Gamma(\alpha)} \gamma(k, \beta x)$$

where $\gamma(k, \beta x)$ is the lower incomplete gamma function.

4. Independent Integer $k$ - Erlang Distribution: If $\alpha$ is a positive integer, i.e., the distribution is an Erlang distribution, the cumulative distribution function has the following series expansion (Papoulis and Pillai (2002))

$$F(x;\ \alpha, \beta) = 1 - \sum_{i=0}^{\alpha-1} \frac{(\beta x)^i}{i!} e^{-\beta x} = e^{-\beta x} \sum_{i=\alpha}^{\infty} \frac{(\beta x)^i}{i!}$$

**Characterization using Shape $k$ and Scale $\theta$**

1. *k/θ* Parameterized Gamma Distribution: A random variable $X$ that is gamma distributed with shape $k$ and rate $\theta$ is denoted

$$X \sim \Gamma(k,\theta) \equiv Gamma(k,\theta)$$

2. *k/θ* Parameterized Gamma Distribution PDF: The probability density function in the shape-scale parametrization is

$$f(x;\ k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

for

$$x > 0$$

and

$$k,\theta > 0$$

Here $\Gamma(k)$ is the gamma function evaluated at $k$.

3. *k/θ* Parameterized Gamma Distribution CDF: The cumulative distribution function is the regularized gamma function

$$F(x; k, \theta) = \int_0^u f(u; k, \theta)du = \frac{1}{\Gamma(k)}\gamma\left(k, \frac{x}{\theta}\right)$$

where $\gamma(k, \beta x)$ is the lower incomplete gamma function.

4. <u>Independent Integer $k$ - Erlang Distribution</u>: It can also be expressed as follows; if $k$ is a positive integer, i.e., the distribution is an Erlang distribution (Papoulis and Pillai (2002))

$$F(x; k, \theta) = 1 - \sum_{i=0}^{k-1} \frac{\left(\frac{x}{\theta}\right)^i}{i!}e^{-\frac{x}{\theta}} = e^{-\frac{x}{\theta}}\sum_{i=k}^{\infty} \frac{\left(\frac{x}{\theta}\right)^i}{i!}$$

## Properties – Skewness

The skewness of the gamma distribution depends only on its shape parameter $k$ , and is equal to $\frac{6}{k}$.

## Properties – Median Calculation

1. <u>Expression for the Median Value</u>: Unlike the mode and the mean which have readily calculable formulas based on the parameters, the median does not have an easy closed form equation. The median for this distribution is defined as the value $\nu$ such that

$$\frac{1}{\Gamma(k)\theta^k} \int_0^\nu x^{k-1} e^{-\frac{x}{\theta}} = \frac{1}{2}$$

2. <u>Banneheka-Ekanayake Approximation for the Median</u>: A formula for approximating the median for any gamma distribution, when the mean is known, has been derived based on the fact that $\frac{\mu}{\mu-\nu}$ is approximately a linear function of $k$ when

$$k \geq 1$$

(Banneheka and Ekanayake (2009)). The approximation formula is

$$\nu = \frac{3k - 0.8}{3k + 0.2}$$

where

$$\mu = k\theta$$

is the mean.

3. <u>Chen and Rubin Median Bounds</u>: A rigorous treatment of the problem of choosing asymptotic expansion and the bounds for the median of the gamma distribution was handled first by Chen and Rubin (1986), who proved that

$$m - \frac{1}{3} < \lambda(m) < m$$

where $\lambda(m)$ denote $\lambda(m)$ the median of the $Gamma(m, 1)$ distribution.

4. <u>Choi Series Expansion for Median</u>: Choi (1994) later showed that the first five terms in the asymptotic expansion of the median are

$$\lambda(m) = m - \frac{1}{3} + \frac{8}{405m} + \frac{184}{25515m^2} + \frac{2248}{3444525m^3} - \cdots$$

by comparing the median to the Ramanujan's $\theta$ function. Later it was shown that $\lambda(m)$ is a convex function of $m$ (Berg and Pedersen (2008)).

## **Properties – Summation**

If $X_i$ has a $Gamma(k_i, \theta)$ distribution for

$$i = 1, \cdots, N$$

where all distributions have the same scale parameter $\theta$, then

$$\sum_{i=1}^{N} X_i = Gamma\left(\sum_{i=1}^{N} k_i, \theta\right)$$

provided all $X_i$ are independent. Mathai (1982) and Moschopoulus (1984) treat the case where $X_i$ are independent but have different scale parameters. The gamma distribution exhibits infinite divisibility.

## Properties – Scaling

If

$$X \sim Gamma(k, \theta)$$

then, for any

$$k > 0$$

$$cX \sim Gamma(k, c\theta)$$

by moment generating functions, or equivalently

$$cX \sim Gamma\left(k, \frac{\beta}{c}\right)$$

Indeed, it is known that if $X$ is an exponential random variable with rate $\lambda$ then $cX$ is an exponential random variable with rate $\frac{\lambda}{c}$; the same thing is true with gamma variates.

## Properties – Exponential Family

The gamma distribution is part of the two-parameter exponential family with natural parameters $k - 1$ and $-\frac{1}{\theta}$ - equivalently $\alpha - 1$ and $-\beta$ - and natural statistics $X$ and $\log X$. If the shape parameter $k$ is held fixed, the resulting one-parameter family of distributions is a natural exponential family.

## Properties – Logarithmic Expectation and Variance

It can be shown that

$$\mathbb{E}[\log X] = \psi(\alpha) - \log \beta$$

or, equivalently,

$$\mathbb{E}[\log X] = \psi(k) + \log \theta$$

where $\psi$ is the digamma function. Likewise,

$$\mathbb{V}[\log X] = \psi_{(1)}(\alpha) = \psi_{(1)}(k)$$

where $\psi_{(1)}$ is the polygamma function. This can be derived using the exponential family formula for the moment generating function of the sufficient statistic, because one of the sufficient statistics of the gamma distribution is $\log X$.

## Properties – Information Entropy

The information entropy is

$$\mathbb{H}[X] = \mathbb{E}[-\log p(X)] = \mathbb{E}[-\alpha \log \beta + \log \Gamma(\alpha) - (\alpha - 1) \log X + \beta X]$$
$$= \alpha - \log \beta + \log \Gamma(\alpha) + (1 - \alpha)\psi(\alpha)$$

In the $k, \theta$ parameterization, the information entropy is given by

$$\mathbb{H}[X] = k + \log\theta + \log\Gamma(k) + (1-k)\psi(k)$$

## Properties – Kullback-Liebler Divergence

1. $\alpha/\beta$ Parametrization of the Divergence: The Kullback-Liebler divergence – KL-divergence - of $Gamma(\alpha_p, \beta_p)$ - the true distribution – from $Gamma(\alpha_q, \beta_q)$ - the *approximating* distribution – is given as

$$D_{KL}(\alpha_p, \beta_p; \ \alpha_q, \beta_q)$$
$$= (\alpha_p - \alpha_q)\,\psi(\alpha_p) - \log\Gamma(\alpha_p) + \log\Gamma(\alpha_q) + \alpha_q(\log\beta_p - \log\beta_q)$$
$$+ \alpha_p\frac{\beta_p - \beta_q}{\beta_p}$$

2. $k/\theta$ Parametrization of the Divergence: The Kullback-Liebler divergence – KL-divergence - of $Gamma(k_p, \theta_p)$ - the true distribution – from $Gamma(k_q, \theta_q)$ - the *approximating* distribution – is given as

$$D_{KL}(k_p, \theta_p; \ k_q, \theta_q)$$
$$= (k_p - k_q)\,\psi(k_p) - \log\Gamma(k_p) + \log\Gamma(k_q) + k_q(\log\theta_q - \log\theta_p)$$
$$+ k_p\frac{\theta_p - \theta_q}{\theta_p}$$

## Properties – Laplace Transform

The Laplace Transform of the Gamma PDF is

$$F(s) = (1 + \theta s)^{-k} = \left(\frac{\beta}{s + \beta}\right)^{\alpha}$$

## Related Distributions – General

1. <u>Relation between Gamma and Exponential Sums</u>: If $X_1, \cdots, X_n$ are $n$ independent and identically distributed random variables following an exponential distribution with rate parameter $\lambda$ then

$$\sum_{i=1}^{N} X_i \sim Gamma(n, \lambda)$$

2. <u>Exponential and Chi Square Distributions</u>: If

$$X \sim Gamma\left(1, \frac{1}{\lambda}\right)$$

- the shape-scale parametrization – then $X$ has an exponential distribution with the rate parameter $\lambda$. If

$$X \sim Gamma\left(\frac{\nu}{2}, 2\right)$$

- shape-scale parametrization – then $X$ is identical to $\chi^2(\nu)$ the chi-square distribution with $\nu$ degrees of freedom. Conversely, if

$$Q \sim \chi^2(\nu)$$

and $c$ is a positive constant, then

$$cQ \sim Gamma\left(\frac{\nu}{2}, 2c\right)$$

3. <u>Erlang and Poisson *Arrival* Process</u>: If $k$ is an integer, the gamma distribution is an Erlang distribution and is the probability distribution of the waiting time until the $k^{th}$ arrival in a one-dimensional Poisson process with intensity $\frac{1}{\theta}$. If

$$X \sim \Gamma(k \in \mathbb{Z}, \theta)$$

$$Y \sim Poisson\left(\frac{x}{\theta}\right)$$

then

$$P(X > x) = P(Y < k)$$

4. <u>Relation to Maxwell-Boltzmann Distribution</u>: If $X$ has a Maxwell-Boltzmann distribution with parameter $a$ then

$$X^2 \sim \Gamma\left(\frac{3}{2}, 2a^2\right)$$

5. <u>Relation to Generalized Gamma Distribution</u>: If

$$X \sim Gamma(k, \theta)$$

then $\sqrt{X}$ follows a generalized gamma distribution with parameters

$$p = 2$$

$$d = 2k$$

124

and

$$a = \sqrt{\theta}$$

More generally, if

$$X \sim Gamma(k, \theta)$$

then $X^q$ for

$$q > 0$$

follows the generalized gamma distribution with parameters

$$p = \frac{1}{q}$$

$$d = \frac{k}{q}$$

and

$$a = \theta^q$$

6. <u>Relation to Inverse Gamma Distribution</u>: If

$$X \sim Gamma(k, \theta)$$

then

$$\frac{1}{X} \sim Gamma\left(k, \frac{1}{\theta}\right)$$

7. <u>Ratio of Independent Random Gamma Variables</u>: If

$$X_k \sim Gamma(\alpha_k, \theta_k)$$

are independent, then

$$\frac{\alpha_2 \theta_2 X_1}{\alpha_1 \theta_1 X_2} \sim F(2\alpha_1, 2\alpha_2)$$

Equivalently

$$\frac{X_1}{X_2} \sim \beta'\left(\alpha_1, \alpha_2, 1, \frac{\theta_1}{\theta_2}\right)$$

Alternatively, if

$$X_k \sim Gamma(\alpha_k, \beta_k)$$

are independent, then

$$\frac{\alpha_2 \beta_1 X_1}{\alpha_1 \beta_2 X_2} \sim F(2\alpha_1, 2\alpha_2)$$

or, equivalently,

$$\frac{X_1}{X_2} \sim \beta'\left(\alpha_1, \alpha_2, 1, \frac{\beta_2}{\beta_1}\right)$$

8.  Distribution of the Ratio $\frac{X}{X+Y}$: If

$$X \sim Gamma(\alpha, \theta)$$

127

and

$$Y \sim Gamma(\beta, \theta)$$

are independently distributed, then $\frac{X}{X+Y}$ has a beta distribution with parameters $\alpha$ and $\beta$, and $\frac{X}{X+Y}$ is independent of $X + Y$, which is $Gamma(\alpha + \beta, \theta)$ distributed.

9. <u>Unit Scale Gamma Variables Sum</u>: If

$$X_i \sim Gamma(\alpha_i, 1)$$

are independently distributed, then the vector $\left(\frac{X_1}{S}, \cdots, \frac{X_n}{S}\right)$, where

$$S = X_1 + \cdots + X_n$$

follows a Dirichlet distribution with parameters $\alpha_1, \cdots, \alpha_n$.

10. <u>Large $k$ Convergence to Normal</u>: For large $k$, the gamma distribution converges to a normal distribution with mean

$$\mu = k\theta$$

and variance

$$\sigma^2 = k\theta^2$$

11. <u>Conjugate Prior of Gaussian Precision</u>: The gamma distribution is a conjugate prior for the precision of a normal distribution with known mean.

12. <u>Multivariate Generalization of Gamma Distribution</u>: The Wishart distribution is a multi-variate generalization of the gamma distribution, where the samples are positive-definite matrices rather than positive real numbers.

13. <u>Other Generalizations of the Gamma Distribution</u>: The gamma distribution is a special case of the generalized gamma distribution, the generalized integer gamma distribution, and the generalized inverse gamma distribution. Further, the gamma distribution is a member of the family of Tweedie exponential dispersion models.

14. <u>Negative Binomial Distribution as a Discrete Analogue</u>: Among the discrete distributions, the negative binomial distribution is sometimes considered the discrete analogue of the gamma distribution.

## Properties – Compound Gamma

If the shape parameter of the gamma distribution in known, but the inverse-sale parameter is unknown, then a gamma distribution for the inverse-scale forms a conjugate prior. The compound distribution which results from integrating out the inverse-scale has a closed form solution, known as the compound gamma distribution (Dubey (1970)). If, instead, the shape parameter is known but the mean is unknown, with the prior of the mean being given by another gamma distribution, it then results in $K$-distribution.

# Statistical Inference – Maximum Likelihood Parameter Estimation

1. <u>Likelihood Estimator – Joint Observation Setup</u>: The likelihood function for $N$ i.i.d. observations $x_1, \cdots, x_n$ is

$$\mathcal{L}(k, \theta) = \prod_{i=1}^{N} f(x_i; k, \theta)$$

from which the log-likelihood function maybe calculated as

$$l(k, \theta) = \log \mathcal{L}(k, \theta) = (k-1) \sum_{i=1}^{N} \log x_i - \sum_{i=1}^{N} \frac{x_i}{\theta} - Nk \log \theta - N \log \Gamma(k)$$

2. <u>Maximization across Scale Parameter Space</u>: Finding the maximum with respect to $\theta$ and setting it equal to zero yields the maximum likelihood estimator of the $\theta$ parameter:

$$\hat{\theta} = \frac{1}{Nk} \sum_{i=1}^{N} x_i$$

Substituting this into the log-likelihood function gives

$$l(k, \theta) = (k - 1) \sum_{i=1}^{N} \log x_i - Nk - Nk \log \sum_{i=1}^{N} \frac{x_i}{Nk} - N \log \Gamma(k)$$

3. <u>Maximization across Shape Parameter Space</u>: Finding the maximum with respect to $k$ by taking the derivative and setting it to zero yields

$$\log k - \psi(k) = \log \sum_{i=1}^{N} \frac{x_i}{N} - \frac{1}{N} \sum_{i=1}^{N} \log x_i$$

4. <u>Iterative Root Search - Starting Point</u>: There is no closed-form solution for $k$. The function is numerically well-behaved, so if a numerical solution is desired, it can be found using, for example, the Newton's method. An initial value of $k$ can be found using the method of moments, or using the approximation

$$\log k - \psi(k) \approx \frac{1}{2k} \left( 1 + \frac{1}{6k + 1} \right)$$

5. <u>Iterative Root Search - Variate Increment</u>: Letting

$$s = \log \sum_{i=1}^{N} \frac{x_i}{N} - \frac{1}{N} \sum_{i=1}^{N} \log x_i$$

131

results in a $k$ that is approximately

$$k \approx \frac{3 - s + \sqrt{(s-3)^2 + 24s}}{12s}$$

which is within 1.5% of the correct value (Minka (2002)). An explicit form for the Newton-Raphson update for this initial guess is (Choi and Wette (1969))

$$k \to k - \frac{\log k - \psi(k) - s}{\frac{1}{k} - \psi'(k)}$$

## Closed-Form Estimators

1. <u>Consistent Estimators for Shape/Scale</u>: Consistent closed-form estimators for $k$ and $\theta$ exist that are derived from the likelihood of the generalized gamma distribution (Ye and Chen (2017)). The estimate for the shape $k$ is

$$\hat{k} = \frac{N \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i \log x_i - \sum_{i=1}^{N} \log x_i \sum_{i=1}^{N} x_i}$$

and the estimate for the scale $\theta$ is

$$\hat{\theta} = \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i \log x_i - \sum_{i=1}^{N} \log x_i \sum_{i=1}^{N} x_i\right)$$

If the rate parametrization is used, the estimate of $\beta$ is

$$\hat{\beta} = \frac{1}{\hat{\theta}}$$

2.  Bias Correction for Shape/Scale: These estimators are not strictly maximum
    likelihood estimators, but are instead referred to a mixed-type log-moment estimators.
    They have, however, similar efficiency as the maximum likelihood estimators.
    Although these estimators are consistent, they have a small bias. A bias corrected
    variant of the estimator for the scale parameter $\theta$ is

$$\hat{\theta} = \frac{N}{N-1}\theta$$

The bias correction for the shape parameter $k$ is given as (Francisco, Ramos, and
Ramos (2019))

$$\tilde{k} = \hat{k} - \frac{1}{N}\left[3\hat{k} - \frac{2}{3}\frac{\hat{k}}{1+\hat{k}} - \frac{4}{5}\frac{\hat{k}}{\left(1+\hat{k}\right)^2}\right]$$

133

# Bayesian Minimum Mean-Squared Error

1.  $\theta$ Posterior from Scale-invariant Prior: With known $k$ and unknown $\theta$, the posterior density function for $\theta$ using the standard scale-invariant prior for $\theta$ is

$$P(\theta \mid k, x_1, \cdots, x_N) \; \alpha \; \frac{1}{\theta} \prod_{i=1}^{N} f(x_i; k, \theta)$$

2.  Expression for Joint Probability: Denote

$$y = \sum_{i=1}^{N} x_i$$

then

$$P(\theta \mid k, x_1, \cdots, x_N) = C(x_i)\theta^{-Nk-1} e^{-\frac{y}{\theta}}$$

3.  Integration across the $\theta$ Space: Integration with respect to $\theta$ can be carried out using a change of variables

$$g = \frac{1}{\theta}$$

revealing that $\frac{1}{\theta}$ is gamma distributed with parameters

$$\alpha_g = Nk$$

and

$$\beta_g = y$$

$$\int_0^\infty \theta^{-Nk-1+m} e^{-\frac{y}{\theta}} d\theta = \int_0^\infty g^{Nk-1-m} e^{-gy} dg = y^{-(Nk-m)} \Gamma(Nk - m)$$

4. <u>Posterior Distribution Mean and Variance</u>: The moments can be computed by taking the ratio of $\mathbb{E}[x^m]$ by $\mathbb{E}[x^0]$ as

$$\mathbb{E}[x^m] = \frac{\Gamma(Nk - m)}{\Gamma(Nk)} y^m$$

which shows that the mean $\pm$ standard deviation estimate for the posterior estimate for $\theta$ is

$$\frac{y}{Nk - 1} \pm \sqrt{\frac{y^2}{(Nk - 1)^2(Nk - 2)}}$$

## Bayesian Inference Conjugate Prior

1. <u>Gamma Distribution as Conjugate Prior</u>: In Bayesian inference, the gamma distribution is the conjugate prior to many likelihood distributions: Poisson, exponential, normal with known mean, Pareto, gamma with known shape $\sigma$, inverse gamma with known shape parameter, and Gompertz with known scale parameter.

2. <u>Conjugate Prior for the Gamma Distribution</u>: The gamma distribution's conjugate prior is (Fink (1997))

$$p(k, \theta \mid u, v, r, s) = \frac{1}{Z} \frac{u^{k-1} e^{-\frac{v}{\theta}}}{\Gamma^r(k)\theta^{ks}}$$

where $Z$ is the normalizing constant, which has no closed-form solution.

3. <u>Bayesian Update of Parameter Priors</u>: The posterior distribution can be found by updating the parameters as follows:

$$p' = p \prod_i x_i$$

$$q' = q + \sum_i x_i$$

$$r' = r + n$$

$$s' = s + n$$

where $n$ is the number of observations, and $x_i$ is the $i^{th}$ observation.

## Occurrence and Applications

1. <u>Insurance Claims and Rainfall Accumulations</u>: The gamma distribution has been used to model the size of the insurance claims (Boland (2007)) and rainfall accumulation (Aksoy (2000)). This means that the aggregated insurance claims and the amount of rainfall accumulated in a reservoir are modeled by a gamma process – much like the exponential distribution generates a Poisson process.
2. <u>Error in Multi-level Poisson Regression</u>: The gamma distribution is also used to model the errors in multi-level Poisson regression models, because the combination of the Poisson distribution and the gamma distribution is a negative binomial distribution.

3. <u>Multi-path Signal Processing</u>: In wireless communication, the gamma distribution is used to model the multi-path fading of the signal power; the Rayleigh and the Poisson distributions are also used.

4. <u>Age Incidence of Cancer Distribution</u>: In oncology, the age incidence of the cancer distribution often follows the gamma distribution, where the shape and the scale parameters predict, respectively, the mean number of driver events and the time interval between them (Belikov (2017)).

5. <u>Inter-spike Neurological Interval Distribution</u>: In neurosciences, the gamma distribution is often used to describe the distribution of inter-spike intervals (Robson and Troy (1987), Wright, Winter, Forster, and Bleeck (2014)).

6. <u>Copy Number in Bacterial Gene Protein Expression</u>: In bacterial gene expression, the copy number of a constitutively expressed protein often follows a gamma distribution, where the shape and the scale parameters are, respectively, the mean number of bursts per cell cycle and the mean number of protein molecules produced by a single mRNA during its lifetime (Friedman, Cai, and Xie (2006)).

7. <u>Signal Recognition in Genomics</u>: In genomics, the gamma distribution has been applied in the peak calling step, i.e., in the recognition of signal, in ChIP-chip (Reiss, Facciotti, and Baliga (2008)), and ChIP-seq (Mendoza-Parra, Nowicka, van Gool, and Gronemeyer (2013)) data analysis.

8. <u>Use as a Conjugate Prior</u>: The gamma distribution is used as a conjugate prior in Bayesian statistics. It is the conjugate prior for the precision, i.e., the inverse of the variance, for a normal distribution. It is also the conjugate prior for the exponential distribution.

## Computational Methods – Generating Gamma Distributed Random Variables

1. <u>Principal Methodology behind Gamma Generation</u>: Give the scaling property above, it is enough to generate gamma variables with

$$\theta = 1$$

as a conversion can be done for any $\beta$ with simple division. Suppose one wishes to generate random variables from $Gamma(n + \delta, 1)$ where $n$ is a non-negative integer and

$$0 < \delta < 1$$

Using the fact that the $Gamma(1, 1)$ is the same as an $Exponential(1)$ distribution, and given that it is straightforward to generate exponential variables, it may be concluded that if $U$ is uniformly distributed on $(0, 1]$ then $-\log U$ is distributed as $Gamma(1, 1)$, i.e., using the inverse transform sampling. Using the "$\alpha$ addition" property of the gamma distribution, this result may be expanded as

$$-\sum_{k=1}^{n} \log U_k \sim \Gamma(n, 1)$$

where $U_k$ are all uniformly distributed on $(0, 1]$ and independent. All that is left now is to generate a variable distributed as $Gamma(\delta, 1)$ for

$$0 < \delta < 1$$

and apply the "$\alpha$ addition" property once more. This is the most difficult part.

2. <u>Literature Coverage on Gamma Generation</u>: Random generation of gamma variables is discussed in detail by Devroye (1986), noting that none are uniformly fast for all shape parameters. For small values of the shape parameters, the algorithms are often not valid (Devroye (1986)). For arbitrary values of the shape parameter, one can apply the Ahrens and Dieter (1982) modified acceptance-rejection method algorithm GD for

$$k \geq 1$$

or the transformation method (Ahrens and Dieter (1974)) when

$$0 < k < 1$$

Also applicable are Marsaglia's squeeze method (Marsaglia (1977)) and the Chang and Feast algorithm GKM3 (Chang and Feast (1979)).

3. <u>Ahrens-Dieter Acceptance-Rejection Method</u>: The following is a version of the Ahrens-Dieter acceptance-rejection method (Ahrens and Dieter (1982)):

   a. Generate $U$, $V$, and $W$ as i.i.d. uniform $(0, 1]$ variables.

   b. If

$$U \leq \frac{e}{e + \delta}$$

then

$$\xi = V^{\frac{1}{\delta}}$$

and

$$\eta = W\xi^{\delta-1}$$

Otherwise

$$\xi = 1 - \log V$$

and

$$\eta = W\xi^{-\delta}$$

c.  If

$$\eta > \xi^{\delta-1}e^{-\xi}$$

then go to step 1.

    d. $\xi$ is distributed as $Gamma(\delta, 1)$

4. <u>Summary of the Ahrens-Dieter Scheme</u>: A summary of this is

$$\theta\left(\xi - \sum_{i=1}^{\lfloor k \rfloor} \log U_i\right) \sim \Gamma(k, \theta)$$

where $\lfloor k \rfloor$ is the integer part of $k$, $\xi$ is generated via the algorithm above with

$$\xi = \{k\}$$

- the fractional part of $k$ - and the $U_k$ are all independent.

5. <u>Devroye Critique of Ahrens-Dieter</u>: While the approach above is technically correct, Devroye notes that it is linear in $k$ and in general not a good choice. Instead, he recommends using either rejection-based or table-based methods, depending on the context (Devroye (1986)).

6. <u>Marsaglia's Transformation-Rejection Method</u>: As an example, Marsaglia's simple transformation-rejection method relies on one normal and one uniform random number (Marsaglia and Tsang (2000)):

    a.

$$d = a = \frac{1}{3}$$

$$c = \frac{1}{\sqrt{9d}}$$

b.

$$v = (1 + cx)^3$$

where $x$ is standard normal.

c. If

$$v > 0$$

and

$$\log UNI < 0.5x^2 + d - dv + d \log v$$

return $d \cdot v$

d. Go back to step 2

7. <u>Acceptance Rate Dependence on $k$</u>: With

$$1 \le a = \alpha = k$$

the above algorithm generates a gamma distributed random number that is approximately constant with $k$ in time. The acceptance rate does depend on $k$, with an acceptance rate of 0.95, 0.98, and 0.99 for $k$ values of 1, 2, and 4 respectively. For

$$k < 1$$

one can use

$$\gamma_\alpha = \gamma_{1+\alpha} U^{\frac{1}{\alpha}}$$

to boost $k$ to be usable with this method.

## References

- Ahrens, J. H., and U. Dieter (1974): Computer Methods for Sampling from Gamma, Beta, Poisson, and Binomial Distributions *Computing* **12 (3)** 223-246
- Ahrens, J. H., and U. Dieter (1982): Generating Gamma Variates by a Modified Rejection Technique *Communications of the ACM* **25 (1)** 47-54
- Aksoy, H. (2000): Use of Gamma Distribution in Hydrological Analysis *Turkish Journal of Engineering and Environmental Science* **24** 419-428

- Banneheka, B. M. S. G., and G. E. M. U. P. D. Ekanayake (2009): A New Point Estimator for the Median of the Gamma Distribution **14** 95-103

- Belikov, A. V. (2017): The Number of Key Carcinogenic Events that can be predicted from Cancer Incidence **7 (1)** 12170

- Berg, C., and H. Pedersen (2008): Convexity of the Median in the Gamma Distribution **arXiV**

- Boland, P. J. (2007): *Statistical and Probabilistic Methods in Actuarial Science* **Chapman and Hall CRC Press**

- Chen, J., and H. Rubin (1986): Bounds for the Difference between the Mean and the Median of Gamma and Poisson Distributions *Statistics and Probability Letters* **4 (6)** 281-283

- Cheng, R. C. H., and G. M. Feast (1979): Some Simple Gamma Variate Generators *Journal of the Royal Statistical Society C* **28 (3)** 290-295

- Choi, S. C., and R. Wette (1969): Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and their Bias *Technometrics* **11 (4)** 683-690

- Choi, K. P. (1994): On the Medians of the Gamma Distributions and an Equation of Ramanujan *Proceedings of the American Mathematical Society* **121 (1)** 245-251

- Devroye, L. (1986): *Non-Uniform Random Variate Generation* **Springer-Verlag** New York

- Dubey, S. D. (1970): Compound Gamma, Beta, and F Distributions *Metrika* **16** 27-31

- Fink, D. (1997): Compendium of Conjugate Priors

- Francisco, L., P. L. Ramos, and E. Ramos (2019): A Note on the Bias of Closed-Form Estimators for the Gamma Distribution Derived from Likelihood Equations *American Statistician* **73 (2)** 195-199

- Friedman, N., L. Cai, and X. S. Xie (2006): Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression *Physical Review Letters* **97** 168302

- Gopalan, P., J. M. Hofman, and D. M. Blei (2014): Scalable Recommendation with Poisson Factorization **arXiV**

- Hogg, R. V., J. McKean, and A. T. Craig (2013): *Introduction to Mathematical Statistics 7th Edition* **Pearson**

- Marsaglia, G. (1977): The Squeeze Method for generating Gamma Variates *Computers and Mathematics with Applications* **3 (4)** 321-325

- Marsaglia, G., W. W. Tsang (2000): A Simple Method for Generating Gamma Variates *ACM Transactions on Mathematical Software* **26 (3)** 363-372

- Mathai, A. M. (1982): Storage Capacity of a Dam with Gamma Type Inputs *Annals of the Institute of Statistical Mathematics* **34 (3)** 591-597

- Mendoza-Parra, M. A., M. Nowicka, W. van Gool, and H, Gronemeyer (2013): [Characterizing ChIP-seq binding patterns by model-based peak shape deconvolution](#)

- Minka, T. P. (2002): [Estimating a Gamma Distribution](#)

- Moschopoulus, P. G. (1985): The Distribution of the Sum of the Independent Random Gamma Variables *Annals of the Institute of Statistical Mathematics* **37 (3)** 541-544

- Papoulis A., and S. U. Pillai (2002): *Probability, Random Variables, and Stochastic Processes 4th Edition* **McGraw-Hill**

- Park, S. Y., and A. K. Bera (2009): Maximum Entropy Auto-regressive Conditional Heteroscedasticity Model *Journal of Econometrics* **150 (2)** 219-230

- Reiss, D. J., M. T. Facciotti, and N. S. Baliga (2008): Model-based Deconvolution of Genome-wide DNA Binding *Bioinformatics* **24 (3)** 396-403

- Robson, J. G., and J. B. Troy (1987): Nature of the Maintained Discharge of the Q, the X, and the Y Retinal Ganglion Cells of the Cat *Journal of the Optical Society of America A* **4 (12)** 2301-2307

- Wikipedia (2019): [Gamma Distribution](#)

- Wright, M. C. M., I. M. Winter, J. J. Forster, and S. Bleeck (2014): Response to Best-Frequency Tone Bursts in the Ventral Cochlear Nucleus is governed by Ordered Inter-spike Interval Statistics *Hearing Research* **317** 23-32

- Ye, Z. S., and N. Chen (2017): Closed-Form Distributions for the Gamma Distribution Derived from the Likelihood Equations *American Statistician* **71 (2)** 177-181

# Chi-Squared Distribution

## Overview

1. <u>Definition of the Chi-Squared Distribution</u>: The chi-squared distribution – also called chi-squared or $\chi^2$ distribution – with $k$ degrees of freedom is the distribution of the sum of the squares of $k$ independent standard normal random variables (Wikipedia (2019)).

2. <u>Specialization of the Gamma Distribution</u>: The chi-squared distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in the construction of confidence intervals (Mood, Graybill, and Boes (1974), Johnson, Klotz, and Balakrishnan (1994), Abramowitz and Stegun (2007), National Institute of Standards and Technology (2019)).

3. <u>Non-central Chi-Squared Distribution</u>: When it is being distinguished from the more general non-central chi-squared distribution, this distribution is sometimes called the central chi-squared distribution.

4. <u>Common Uses of the Distribution</u>: The chi-squared distribution is used in the common chi-squared tests for the goodness of distribution of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in the confidence interval estimation for the population standard deviation of a normal distribution from a sample standard deviation.

5. <u>Analysis of Variance by Ranks</u>: Many other statistical tests also use this distribution,, such as Friedman's analysis of variance by ranks.

## Definition

1. <u>Chi-Squared Distribution – Mathematical Definition</u>: If $Z_1, \cdots, Z_k$ are independent standard normal random variable, then the sum of their squares

$$Q = \sum_{i=1}^{k} Z_i^{\,2}$$

is distributed according to the chi-squared distribution with $k$ degrees of freedom.

2. <u>Chi-Squared Representation and Parametrization</u>: This is usually denoted as

$$Q \sim \chi^2(k)$$

or

$$Q \sim \chi_k^{\,2}$$

The chi-squared distribution has one parameter $k$, a positive integer that specifies the number of degrees of freedom – the number of 's.

# Introduction

1. <u>Primary Use in Hypothesis Testing</u>: The chi-squared distribution is used primarily in hypothesis testing.
2. <u>Not Used in Direct Modeling</u>: Unlike the more widely known distribution such as the normal distribution and the exponential distribution, the chi-squared distribution is not as often applied in direct modeling of natural phenomenon.
3. <u>Typical Hypothesis Test Use Cases</u>: It arises in the following hypothesis tests, among others:
   a. Chi-squared test of independence in contingency tables.
   b. Chi-squared test of goodness of fit of observed data to hypothetical distributions
   c. Likelihood-ratio test for nested models
   d. Log-rank test in survival analysis
   e. Cochran-Mantel-Haenszel test for stratified contingency tables
4. <u>Other Uses of the Distribution</u>: It is also a component of the definition of the t-distribution and F-distribution used in t-tests, analysis of variance, and regression analysis.
5. <u>Relationship to the Normal Distribution</u>: The primary reason that the chi-squared distribution is used extensively in hypothesis testing is its relationship to the normal distribution.
6. <u>Test-Statistic Sample Size Behavior</u>: Many hypothesis tests use a test statistic, such as the t-statistic in a t-test. For these hypothesis tests, as the sample size $n$ increases, the sampling distribution of the test statistic approaches a normal distribution – central limit theorem.

7. <u>Asymptotic Test-Statistic Sampling</u>: Because the test statistic – such as $t$ – is asymptotically normally distributed, provided that the sample size is sufficiently large, the distribution used for hypothesis testing may be approximated by a normal distribution.

8. <u>Usage with Underlying Normal Distributions</u>: Testing hypothesis using a normal distribution is well-understood and relatively easy. The simplest chi-squared distribution is the square of a standard normal distribution. So wherever a normal-distribution could be used for a hypothesis test, a chi-squared distribution could be used.

9. <u>Random Draw from a Normal Distribution</u>: Specifically, suppose that $Z$ is a standard normal random variable, with mean $0$ and variance $1$,

$$Z \sim \mathcal{N}(0,1)$$

A sample drawn at random from $Z$ is a sample from the standard normal distribution.

10. <u>Random Draw from Chi-Squared Distribution</u>: Define a new random variable $Q$. To generate a random sample from $Q$, take a sample from $Z$ and square the value. The distribution of the squared values is given by the random variable

$$Q = Z^2$$

The distribution of the random variable $Q$ is an example of a chi-squared distribution

$$Q \sim \chi_1^2$$

11. <u>Degrees of Freedom of the Distribution</u>: The subscript 1 indicates that this particular chi-squared is constructed from only one standard normal distribution. A chi-squared distribution constructed by squaring a single standard normal distribution is said to have 1 degree of freedom.

12. <u>Chi-Squared Distribution Asymptotic Approach</u>: Thus, as the sample size for a hypothesis test increases, the distribution of the test-statistic approaches a normal distribution, and the distribution of the square of the test-statistic approaches a chi-squared distribution.

13. <u>Extreme Values under the Distribution</u>: Just as the extreme values of the normal distribution have low probability – and give small $p$-values – extreme values of the chi-squared distribution have low probability.

14. <u>Likelihood Ratio Class of Tests</u>: An additional reason that the chi-squared distribution is widely used is that it is a member of the class of likelihood ratio tests LRT (Westfall (2013)).

15. <u>Neyman-Pearson Lemma</u>: LRT's have several desirable properties; in particular, LRT's commonly provide highest power to reject NULL hypothesis.

16. <u>Advantages of using t-distribution</u>: However, the normal and the chi-squared distribution are valid only asymptotically. For this reason, it is better to use the t-distribution rather than the normal approximation or the chi-squared approximation for small sample size.

17. <u>Power of Exact Binomial Test</u>: Similarly, in the analysis of contingency tables, the chi-squared approximation will be poor for small sample size, and it is preferable to use Fisher's exact test. Ramsey (1988) shows that the exact binomial test is always more powerful than the normal approximation.

18. <u>Binomial, Normal, and Chi-Squared</u>: Lancaster (1969) showed the connections between the binomial, the normal, and the chi-squared distributions as follows. De Moivre and Laplace established that a binomial distribution could be approximated by a normal distribution. Specifically, they showed the asymptotic normality of the random variable

$$\chi = \frac{m - Np}{\sqrt{Npq}}$$

where $m$ is the observed number of successes in $N$ trials, the probability of success is $p$ and

$$q = 1 - p$$

19. <u>Chi-Squared Distribution Variable</u>: Squaring both sides of the equation gives

$$\chi^2 = \frac{(m - Np)^2}{Npq}$$

20. <u>Re-factoring the Chi-Squared Variable</u>: Using

$$N = Np + N(1 - p)$$

$$N = m + (N - m)$$

and

$$q = 1 - p$$

this equation simplifies to

$$\chi^2 = \frac{(m - Np)^2}{Np} + \frac{(N - m - Np)^2}{Nq}$$

21. <u>Multivariate Generalization by Pearson</u>: The expression on the right is of the form that Pearson would generalize to

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

where $\chi^2$ is the Pearson's cumulative test statistic, which asymptotically approaches a $\chi^2$ distribution, $O_i$ is the number of observations of type $i$,

$$E_i = Np_i$$

is the expected theoretical frequency of the type $i$, asserted by the NULL hypothesis that the fraction of type $i$ in the population is $p_i$, and $n$ is the number of cells in the table.

22. <u>Reducing Binomial to Normal - $\chi^2$</u>: In the case of a binomial outcome – flipping a coin – the binomial distribution may be approximated by a normal distribution for sufficiently large $n$. Because the square of a standard normal distribution is the chi-squared distribution with 1 degree of freedom, the probability of results such as 1

head in 10 trials can be approximated by either the normal or the chi-squared distribution.

23. Extension to Multiple Categorical Variables: However, many problems involve more than two possible outcomes of a binomial, and instead require 3 or more categories, which leads to the multinomial distribution.

24. Chi Squared as Approximating Multinomial Distribution: Just as de Moivre and Laplace sought for and found the normal distribution approximation to the binomial, Pearson sought for and found a multivariate normal approximation to the multinomial distribution. Pearson showed that the chi-squared distribution, the sum of multiple normal distributions, was such as approximation to the multinomial distribution (Lancaster (1969)).

## Probability Density Function

The probability density function of the chi-squared distribution is

$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} & x > 0 \\ 0 & otherwise \end{cases}$$

where $\Gamma\left(\frac{k}{2}\right)$ denotes the gamma function, which has closed-form values for integer $k$.

## Cumulative Distribution Function

1. Explicit Expression for the CDF: The cumulative distribution function is

$$F(x; k) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} = p\left(\frac{k}{2}, \frac{x}{2}\right)$$

   where $\gamma\left(\frac{k}{2}, \frac{x}{2}\right)$ is the lower incomplete gamma function and $p\left(\frac{k}{2}, \frac{x}{2}\right)$ is the regularized gamma function.

2. Expression for the Special Case $k = 2$: The special case of

$$k = 2$$

   of this function has a simple form:

$$F(x; 2) = 1 - e^{-\frac{x}{2}}$$

   and the integer recurrence of the gamma function makes it easy to compute for other small even $k$.

3. Tables for the $\chi^2$ CDF: Tables for the chi-squared cumulative distribution function are widely available and the function is included in many spreadsheets and all statistical packages.

4. <u>Chernoff Bounds on CDF Tails</u>: Letting

$$z \equiv \frac{x}{k}$$

Chernoff bounds on the lower and the upper tails of the CDF may be obtained (Dasgupta and Gupta (2003)).

5. <u>Chernoff Bounds for $0 < z < 1$</u>: For the cases when

$$0 < z < 1$$

– which include all of the cases when this CDF is less than half –

$$F(zk; k) \leq (ze^{1-z})^{\frac{k}{2}}$$

6. <u>Chernoff Bounds for $z > 1$</u>: The tail bound for cases when

$$z > 1$$

similarly is

$$1 - F(zk; k) \leq (ze^{1-z})^{\frac{k}{2}}$$

## Additivity

1. <u>Sum of Independent Chi-squared Variables</u>: It follows from the definition of the chi-squared distribution that the sum of the independent chi-squared variables is also chi-squared distributed.

2. <u>Additivity over Degrees of Freedom</u>: Specifically, if $\{X_i\}_{i=1}^n$ are independent chi-squared variables with $\{k_i\}_{i=1}^n$ degrees of freedom, respectively, then

$$Y = X_1 + \cdots + X_n$$

is chi-squared with $k_1 + \cdots + k_n$ degrees of freedom.

## Sample Mean

1. <u>Chi-Squared Distribution Sample Mean</u>: The sample mean of $n$ i.i.d. chi-squared variables of degree $k$ is distributed according to a gamma distribution with shape $\alpha$ and scale $\theta$ parameters:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim Gamma\left(\alpha = \frac{nk}{2}, \theta = \frac{2}{n}\right)$$

where

$$X_i \sim \chi^2$$

2.  <u>Asymptotic Reduction to Normal Distribution</u>: Asymptotically, give that for a scale parameter $\alpha$ going to infinity, a Gamma distribution converges to a normal distribution with expectation

$$\mu = \alpha \cdot \theta$$

and variance

$$\sigma^2 = \alpha \cdot \theta^2$$

the sample mean converges towards

$$\underset{n \to \infty}{Limit} \bar{X} \to \mathcal{N}\left(\mu = k, \sigma^2 = 2\frac{k}{n}\right)$$

3. Asymptotics using CLT: Note that one would have obtained the same result instead invoking the central limit theorem, noting that for each chi-squared variable of degree $k$, the expectation is $k$ and the variance is $2k$ – and hence the variance of the mean $\bar{X}$ is

$$\sigma^2 = 2\frac{k}{n}$$

**Entropy**

1. Expression for Differential Entropy: The differential entropy is given by

$$h = \int_0^\infty f(x; k) \log f(x; k)\, dx = \frac{k}{2} + \log\left[2\Gamma\left(\frac{k}{2}\right) + \left(1 - \frac{k}{2}\right)\psi\left(\frac{k}{2}\right)\right]$$

where $\psi(x)$ is the digamma function.

2. Chi-squared as a MaxEnt Distribution: The chi-squared distribution is the maximum entropy probability distribution for a random variable $X$ for which

$$\mathbb{E}[X] = k$$

and

$$\mathbb{E}[\log X] = \psi\left(\frac{k}{2}\right) + \log 2$$

are fixed.

3. <u>Log Moment of Gamma Distributions</u>: Since the chi-squared is in the family of gamma distributions, this can be derived by substituting the appropriate values in the expectation of the moments of gamma.

## Non-central Moments

The moments about zero of a chi-squared distribution with $k$ degrees of freedom are given (Simon (2002))

$$\mathbb{E}[X^m] = k(k+2)(k+4)\cdots(k+2m-2) = 2^m \frac{\Gamma\left(m+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

## Cumulants

The cumulants are readily obtained by a formal power series expansion of the logarithm of the characteristic function $\kappa_n = 2^{n-1}(n-1)!\,k$

## Asymptotic Properties

1. <u>Degrees of Freedom Based Determinants of Normality</u>: By the central limit theorem, because the chis-squared distribution is the sum of $k$ independent normal variables with finite mean and variance, it converges to a normal distribution for large $k$. For many practical purposes, for

$$k > 50$$

the distribution is sufficiently close to a normal distribution (Hunter, Box, and Hunter (1970)).

2. <u>Speed of Convergence to Normality</u>: Specifically, if

$$X \sim \chi^2(k)$$

then as $k$ tends to infinity, the distribution of $\frac{X-k}{\sqrt{2k}}$ tends to a standard normal distribution. However, convergence is slow because the skewness is $\sqrt{\frac{8}{k}}$ and the excess kurtosis is $\frac{12}{k}$

3. Schemes for Faster Approach to Normality: The sampling distribution of $\log \chi^2$ converges to normality much faster than the sampling distribution of $\chi^2$ (Bartlett and Kendall (1946)), as the logarithm removes much of the asymmetry (Pillai (2016)). Other functions of the chi-squared distribution converge more rapidly to a normal distribution.

4. Chi-squared Derivative Function #1: Some examples are: If

$$X \sim \chi^2(k)$$

then $\sqrt{2X}$ is approximately normally distributed with mean $\sqrt{2k-1}$ and unit variance (Johnson, Klotz, and Balakrishnan (1994)).

5. Chi-squared Derivative Function #2: If

$$X \sim \chi^2(k)$$

then $\left(\frac{X}{k}\right)^{\frac{1}{3}}$ is approximately normally distributed with mean $1 - \frac{2}{9k}$ and variance $\frac{2}{9k}$. This is known as the Wilson-Hilferty transformation (Wilson and Hilferty (1931), Johnson, Klotz, and Balakrishnan (1994)).

## Relation to Other Distributions

1. Normal Distribution: As

$$k \to \infty$$

$$\frac{\chi_k{}^2 - k}{\sqrt{2k}} \to \mathcal{N}(0, 1)$$

2. Non-central Chi-squared Distribution:

$$\chi^2(k) \sim {\chi'}^2(k)$$

is a non-central chi-squared distribution with the non-centrality parameter

$$\lambda = 0$$

3. Double Chi-squared Distribution #1: If

$$Y \sim F(v_1, v_2)$$

then

$$X = \mathop{Limit}_{v_2 \to \infty} v_1 Y$$

has the double chi-squared distribution $\chi^2(v_1)$

4. <u>Double Chi-squared Distribution #2</u>: As a special case, if

$$Y \sim F(1, v_2)$$

then

$$X = \mathop{Limit}_{v_2 \to \infty} Y$$

has the double chi-squared distribution $\chi^2(1)$

5. <u>Chi-squared Distribution as a Norm</u>: The squared norm of $k$ standard normally distributed variables is a chi-squared distribution with $k$ degrees of freedom:

$$\left\| N_{i=1,\cdots,k}(0, 1) \right\|^2 \sim \chi^2(k)$$

6. <u>Gamma Distribution</u>: If

$$X \sim \chi^2(v)$$

and

$$c > 0$$

then

$$cX \sim \Gamma\left(k = \frac{v}{2}, \theta = 2c\right)$$

which is a gamma distribution.

7. <u>Chi Distribution</u>: If

$$X \sim \chi_k{}^2$$

then

$$\sqrt{X} \sim \chi_k$$

the chi-distribution.

8. <u>Exponential Distribution</u>: If

$$X \sim \chi^2(2)$$

then

$$X \sim e^{\frac{1}{2}}$$

which is an exponential distribution.

9. <u>Rayleigh Distribution</u>: If

$$X \sim Rayleigh(1)$$

which is a Rayleigh distribution, then

$$X^2 \sim \chi^2(2)$$

10. <u>Maxwell Distribution</u>: If

$$X \sim Maxwell(1)$$

which is a Maxwell distribution, then

$$X^2 \sim \chi^2(3)$$

11. <u>Inverse Chi-squared Distribution</u>: If

$$X \sim \chi^2(\nu)$$

then

$$\frac{1}{X} \sim Inv - \chi^2(\nu)$$

which is the inverse chi-squared distribution.

12. <u>Type 3 Pearson Distribution</u>: The chi-squared distribution is a special case of Type-3 Pearson distribution.

13. <u>Beta Distribution</u>: If

$$X \sim \chi^2(\nu_1)$$

and

$$Y \sim \chi^2(\nu_2)$$

are independent, then

$$\frac{X}{X+Y} \sim Beta\left(\frac{v_1}{2} + \frac{v_2}{2}\right)$$

which is the beta distribution.

14. <u>Chi-squared Distribution from Uniform</u>: If

$$X \sim U(0,1)$$

which is a uniform distribution, then

$$-2\log X \sim \chi^2(2)$$

15. <u>Laplace Distribution Transformation</u>: $\chi^2(6)$ is a transformation of the Laplace distribution. If

$$X_i \sim Laplace\,(\mu, \beta)$$

then

$$\sum_{i=1}^{n} \frac{2|X_i - \mu|}{\beta} \sim \chi^2(2n)$$

16. <u>Generalized Normal Distribution Version #1</u>: If $X_i$ follows a generalized normal distribution version 1 with parameters $\mu$, $\alpha$, and $\beta$ then

$$\sum_{i=1}^{n} \frac{2|X_i - \mu|^{\beta}}{\alpha} \sim \chi^2\left(\frac{2n}{\beta}\right)$$

(Backstrom and Fischer (2018))

17. <u>Transformed Pareto Distribution</u>: Chi-squared distribution is a transformation of the Pareto distribution.

18. <u>Student t Distribution from Chi-Squared</u>: Student-t distribution is a transformation of the chi-squared distribution. Student-t distribution can be obtained from a chi-squared distribution and a normal distribution.

19. <u>Non-central Beta Distribution</u>: Non-central beta distribution can be obtained as a transformation of the chi-squared distribution and non-central chi-squared distribution.

20. <u>Non-central -t Distribution</u>: Non-central t-distribution can be obtained from normal distribution and chi-squared distribution.

21. <u>Multidimensional Gaussian Random Vectors</u>: If $Y$ is a $k$ dimensional Gaussian random vector with mean $\mu$ and rank $k$ covariance matrix $C$, then

$$X = [Y - \mu]^T C^{-1} [Y - \mu]$$

is chi-squared with $k$ degrees of freedom.

22. <u>Non central Chi squared Distribution</u>: The sum of squares of statistically independent unit variance Gaussian variables which do *not* have mean zero yields a generalization of the chi-squared distribution called the non-central chi-squared distribution.

23. <u>Rank-reduced Chi-squared Distribution</u>: If $Y$ is a vector of $k$ i.i.d. standard normal variables and $A$ is a $k \times k$ symmetric idempotent matrix with rank $k - n$, then the quadratic form $Y^T A Y$ is chi-squared with $k - n$ degrees of freedom.

24. <u>Special Normal Chi-squared Distribution</u>: If $C$ is a $p \times p$ positive semi-definite covariance matrix with strictly positive diagonal covariance entries, then for

$$X \sim \mathcal{N}(0, C)$$

and $w$ a random $p$-vector independent of $X$ such that

$$w_1 + \cdots + w_p = 1$$

and

$$w_i \geq 0$$

$$i = 1, \cdots, p$$

it then holds that

$$\frac{1}{\left[\frac{w_i}{X_i}\right]^T C \left[\frac{w_i}{X_i}\right]} \sim {\chi_1}^2$$

(Pillai (2016)).

25. <u>Relation to Non-Gaussian Distributions</u>: Thus, the chi-squared is also naturally related to other distributions arising from the Gaussian.

26. <u>Ratio of Independent $\chi^2$ Distributions</u>: In particular, if $Y$ is $F$-distributed,

$$Y \sim F(k_1, k_2)$$

if

$$Y = \frac{X_1/k_1}{X_2/k_2}$$

where

$$X_1 \sim \chi^2(k_1)$$

and

$$X_2 \sim \chi^2(k_2)$$

are statistically independent.

27. Sum of Correlated $\chi^2$ Distributions: If

$$X_1 \sim \chi^2(k_1)$$

and

$$X_2 \sim \chi^2(k_2)$$

are statistically independent, then

$$X_1 + X_2 \sim \chi^2(k_1 + k_2)$$

If $X_1$ and $X_2$ are not independent, then $X_1 + X_2$ is not chi-squared distributed.

## Generalizations

1. <u>Construction of the Chi-squared Distributions</u>: The chi-squared distribution is obtained as the sum of $k$ independent, zero-mean unit-variance Gaussian random variables.

2. <u>Generalizing the Chi-squared Distribution</u>: Generalizations of this distribution can be obtained by summing the squares of other types of Gaussian random variables. Several such distributions are described below.

## Linear Combination

If $X_1, \cdots, X_n$ are chi-squared random variables and

$$a_1, \cdots, a_n \in \mathbb{R}^{>0}$$

then a closed expression for the distribution is

$$X = \sum_{i=1}^{n} a_i X_i$$

is not known. It may be, however, approximated efficiently using the property of characteristic functions of chi-squared random variables (Bausch (2013)).

## Non-Central Chi-Squared Distribution

The non-central chi-squared distribution is obtained from the sum of squares of independent Gaussian random variables having unit variance and *non-zero* means.

## Generalized Chi-Squared Distribution

The generalized chi-squared distribution is obtained from the quadratic form $z^T A z$ where $z$ is a zero-means Gaussian vector having an arbitrary covariance matrix, and $A$ is an arbitrary matrix.

## Gamma, Exponential, and Related Distribution

1. <u>Parametrized Specialization of the Gamma Distribution</u>: The chi-squared distribution

$$X \sim \chi_k{}^2$$

is a special case of gamma distribution, in that

$$X \sim \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$$

using the rate parameterization of the gamma distribution – or

$$X \sim \Gamma\left(\frac{k}{2}, 2\right)$$

using the scale parameterization of the Gamma distribution – where $k$ is an integer.

2. <u>Exponential Distribution from Chi-squared</u>: Because the exponential distribution is also a special case of the gamma distribution, one also has that if

$$X \sim \chi_2{}^2$$

then

$$X \sim \sqrt{e}$$

is an exponential distribution.

3. <u>Erlang Specialization of the Gamma Distribution</u>: The Erlang distribution is also a special case of the gamma distribution, and thus one also has that if

$$X \sim \chi_k{}^2$$

with even $k$, then $X$ is Erlang distributed with shape parameter $\frac{k}{2}$ and scale parameter $\frac{1}{2}$.

## Occurrence and Applications

1. $\chi^2$ Test and Variance Estimation: The chi-squared distribution has numerous applications in inferential statistics, for instance in chi-squared tests, and in estimating variances.

2. Regression Slope and Population Mean: It enters the problem of estimating the mean of a normally distributed population and the problem of estimating the slope of a regression line via its role in the Student-t distribution.

3. Analysis of Variance using F-distribution: It enters all analysis of variance problems via its role in the F-distribution, which is the distribution of the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom.

4. Central $\chi^2$ from Gaussian Distribution: Following are some of the most common situations in which the chi-squared distribution arises from a Gaussian distributed sample. If $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2 \chi_{n-1}^2$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

5. $\chi^2$ Variants in Normal Distribution: The box below shows some statistics based on

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i{}^2)$$

$$i = 1, \cdots, k$$

independent random variables that have probability distributions related to chi-squared distribution.

| Name | Distribution |
|---|---|
| Chi-squared Distribution | $\displaystyle\sum_{i=1}^{k}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ |
| Non-central Chi-squared Distribution | $\displaystyle\sum_{i=1}^{k}\left(\frac{X_i}{\sigma_i}\right)^2$ |
| Chi Distribution | $\displaystyle\sqrt{\sum_{i=1}^{k}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2}$ |

| Non-central Chi Distribution | $\sqrt{\displaystyle\sum_{i=1}^{k}\left(\frac{X_i}{\sigma_i}\right)^2}$ |
|---|---|

6. Use in Magnetic Resonance Imaging: The chi-squared distribution is also often encountered in magnetic resonance imaging (den Dekker and Sijbers (2014)).

## Table of $\chi^2$ Values vs. $p$-Values

1. Review of the $p$-Value Definition: The $p$-Value is the probability of observing a test statistic *at least* as extreme in the specified distribution (here chi-squared).

2. Computing the $p$-Value from CDF: Thus, since the cumulative distribution function (CDF) for the appropriate degrees of freedom gives the probability of having obtained a value *less extreme* than this point, subtracting CDF from 1 gives the $p$-Value.

3. Establishing Statistical Significance using $p$-Values: A low $p$-Value, below the chosen significance level, indicates statistical significance, i.e., sufficient evidence to reject the NULL hypothesis.

4. Typical Hypothesis Testing Significance Threshold: A significance level of 0.05 is often used as the cut-off between significant and not-significant results.

5. $p$-Values for $\chi^2$ with 10 Degrees of Freedom: The table below gives a number of $p$-Values matching to $\chi^2$ for the first 10 degrees of freedom (Pennsylvania State University (2016)).

| Degrees of | $\chi^2$ Value |
|---|---|

| Freedom | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.004 | 0.02 | 0.06 | 0.15 | 0.46 | 1.07 | 1.84 | 2.71 | 3.84 | 6.63 | 10.830 |
| 2 | 0.100 | 0.21 | 0.45 | 0.71 | 1.39 | 2.41 | 3.22 | 4.61 | 5.99 | 9.21 | 13.820 |
| 3 | 0.350 | 0.58 | 1.01 | 1.42 | 2.37 | 3.66 | 4.64 | 6.25 | 7.81 | 11.34 | 16.270 |
| 4 | 0.710 | 1.06 | 1.65 | 2.20 | 3.36 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 | 18.470 |
| 5 | 1.140 | 1.61 | 2.34 | 3.00 | 4.35 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 | 20.520 |
| 6 | 1.630 | 2.20 | 3.07 | 3.83 | 5.35 | 7.23 | 8.56 | 10.64 | 12.49 | 16.81 | 22.460 |
| 7 | 2.170 | 2.83 | 3.82 | 4.67 | 6.35 | 8.38 | 9.80 | 12.02 | 14.07 | 18.48 | 24.320 |
| 8 | 2.730 | 3.49 | 4.59 | 5.53 | 7.34 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 | 26.120 |
| 9 | 3.320 | 4.17 | 5.38 | 6.39 | 8.34 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 | 27.880 |
| 10 | 3.940 | 4.87 | 6.18 | 7.27 | 9.34 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 | 29.590 |
| **p-Value** | **0.950** | **0.90** | **0.80** | **0.70** | **0.50** | **0.30** | **0.20** | **0.10** | **0.05** | **0.01** | **0.001** |

6. <u>p-Value from Inverse CDF Function</u>: The above values can also be calculated using the quantile function – also known as *inverse CDF* or *ICDF* – of the chi-squared distribution – e.g., for a p-Value of 0.05 and 7 degrees of freedom no gets 14.06714, or 14.07 as in the able above.

## Summary Expressions

| Notation | $\chi^2(k)$ or $\chi_k{}^2$ |
|---|---|
| **Parameters** | $k \in \mathbb{N}^+$ (degrees of freedom) |
| **Support** | $x \in (0, +\infty)$ if $k = 1$  $x \in [0, +\infty)$ otherwise |
| **Probability Density Function** | $\dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$ |
| **Cumulative Density Function** | $\dfrac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$ |
| **Mean** | $k$ |
| **Median** | $\approx k\left(1 - \dfrac{2}{9k}\right)^3$ |
| **Mode** | $\max(k - 2, 0)$ |

| | |
|---|---|
| **Variance** | $2k$ |
| **Skewness** | $\sqrt{\dfrac{8}{k}}$ |
| **Excess Kurtosis** | $\dfrac{12}{k}$ |
| **Entropy** | $\dfrac{k}{2} + \log\left[2\Gamma\left(\dfrac{k}{2}\right) + \left(1 - \dfrac{k}{2}\right)\psi\left(\dfrac{k}{2}\right)\right]$ |
| **Moment Generating Function** | $(1 - 2t)^{-\frac{k}{2}}$ for $t < \dfrac{1}{2}$ |
| **Characteristic Function** | $(1 - 2it)^{-\frac{k}{2}}$ Sanders (2011) |
| **Probability Generating Function** | $(1 - 2\log t)^{-\frac{k}{2}}$ for $0 < t < \sqrt{e}$ |

# References

- Abramowitz, M., and I. A. Stegun (2007): *Handbook of Mathematics Functions* **Dover Book on Mathematics**

- Backstrom, T., and J. Fischer (2018): Fast Randomization for Distributed Low Bit-rate Coding of Speech and Audio *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **26 (1)** 19-30

- Bartlett, M. S., and D. G. Kendall (1946): The Statistical Analysis of Variance Heterogeneity and the Logarithmic Transformation *Supplement to the Journal of the Royal Statistical Society* **8 (1)** 128-138

- Bausch, J. (2013): On the Efficient Calculation of a Linear Combination of Chi-Square Random Variables with an Application in Counting String Vacua **arXiv**

- Dasgupta, S. D. K., and A. K. Gupta (2003): An Elementary Proof of a Theorem of Johnson and Lindenstrauss *Random structures and Algorithms* **22 (1)** 60-65

- den Dekker, A. J., and j. Sijbers (2014): Data Distributions in Magnetic Resonance Images: A Review *European Journal of Medical Physics* **30 (7)** 725-741

- Hunter, W. G., G. E. P. Box, and J. S. Hunter (1978): *Statistics for Experimenters* **Wiley**

- Johnson, N. L., S. Klotz, and N. Balakrishnan (1994): *Continuous Univariate Distributions 1 2nd Edition* **John Wiley and Sons**

- Lancaster, H, O. (1969): *The Chi-Squared Distribution* **Wiley**

- Mood, A., A. F. Graybill, and D. C. Boes (1974): *Introduction to the Theory of Statistics 3rd Edition* **McGraw-Hill**

- National Institute for Standards and Technology (2019): Chi-Square Distribution

- Pennsylvania State University (2016): Chi-squared Distribution

- Pillai, N. S. (2016): An Unexpected Encounter with Cauchy and Levy *Annals of Statistics* **44 (5)** 2089-2097

- Ramsey, P. H. (1988): Evaluating the Normal Approximation to the Binomial Test *Journal of Educational Statistics* **13 (2)** 173-182

- Sanders, M. A. (2011): Characteristic Function of the Central Chi-squared Distribution

- Simon, M. K. (2002): *Probability Distributions involving Gaussian Random Variables* **Springer** New York, NY.

- Westfall, P. H. (2013): *Understanding Advanced Statistical Methods* **CRC Press** Boca Raton, FL

- Wikipedia (2019): Chi-squared Distribution

- Wilson, E. B., and M. M. Hilferty (1931): The Distribution of Chi-squared *Proceedings of the National Academy of Sciences* **17 (12**) 684-688.

# Non-central Chi-Square Distribution

## Overview

1. <u>Generalization of the Chi-Square Distribution</u>: The *non-central chi-square distribution* is a generalization of the chi-square distribution (Wikipedia (2019)).
2. <u>NULL Chi Square Distribution Tests</u>: It often arises in the power analysis of statistical tests in which the NULL distribution – perhaps asymptotically – is a chi-square distribution; important examples of such tests are the likelihood ratio tests.

## Background

1. <u>Non-central Chi-square Distribution</u>: Let $(X_1, \cdots, X_k)$ be $k$ independent, normally distributed random variables with means $\mu_i$ and unit variances. Then the random variable $\sum_{i=1}^{k} X_i^2$ is distributed according to the non-central chi-square distribution.
2. <u>Non-central Chi-square Distribution Parameters</u>: It has 2 parameters - $k$ which specifies the number of degrees of freedom, i.e., the number of $X_i$, and $\lambda$ which is related to the mean of the random variables $X_i$ by

$$\lambda = \sum_{i=1}^{k} \mu_i^2$$

$\lambda$ is sometimes called the non-centrality parameter; some references define $\lambda$ in other ways, such as half of the above sum, or its square root.

3. <u>$\chi'^2$ as $\mathcal{N}(\mu, I_k)$ Squared Mean</u>: This distribution arises in multivariate statistics as a derivative of multivariate normal distributions. While the central chi-square

distribution is the squared norm of a random vector with $\mathcal{N}(0_k, I_k)$ distribution, i.e., squared distance from the origin to a point taken at random from that distribution, the non-central chi-square is the squared norm of a random vector with $\mathcal{N}(\mu, I_k)$ distribution, Here $0_k$ is a vector of length $k$,

$$\mu = (\mu_1, \cdots, \mu_k)$$

and $I_k$ is the identity matrix of size $k$.

## Non-central Chi-Square Distribution Table

| Parameters | $k > 0$ Degrees of Freedom $\lambda > 0$ Non-centrality Parameter |
|---|---|
| **Support** | $x \in [0, +\infty)$ |
| **CDF** | $\dfrac{1}{2} e^{-\frac{x+\lambda}{2}} \left(\dfrac{x}{\lambda}\right)^{\frac{k}{4}-\frac{1}{2}} I_{\frac{k}{2}-1}\left(\sqrt{\lambda x}\right)$ |
| **PDF** | $1 - Q_{\frac{k}{2}}\left(\sqrt{\lambda}, \sqrt{x}\right)$ with Marcum Q-function $Q_M(a, b)$ |
| **Mean** | $k + \lambda$ |
| **Variance** | $2(k + 2\lambda)$ |
| **Skewness** | $\dfrac{2^{\frac{3}{2}}(k + 2\lambda)}{(k + 2\lambda)^{\frac{3}{2}}}$ |
| **Excess Kurtosis** | $12 \dfrac{(k + 4\lambda)}{(k + 2\lambda)^2}$ |
| **MGF** | $\dfrac{e^{\frac{\lambda t}{1-2t}}}{(1-2t)^{\frac{k}{2}}}$ for $2t < 1$ |
| **CF** | $\dfrac{e^{\frac{i\lambda t}{1-2it}}}{(1 - 2it)^{\frac{k}{2}}}$ |

## Definition

1.  <u>Non-central Chi-square PDF</u>: The probability density function is given by

$$f_X(x; k, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\frac{\lambda}{2}}}{i!} \left(\frac{\lambda}{2}\right)^i f_{Y_{k+2i}}(x)$$

   where $Y_q$ is distributed as chi-square with $q$ degrees of freedom.

2.  <u>Intuition behind the PDF Expression</u>: From the above representation, the non-central chi-square distribution is seen to be a Poisson-weighted mixture of central chi-square distributions. Suppose that a random variable $J$ has a Poisson distribution with mean $\frac{\lambda}{2}$, and the conditional distribution of $Z$ given

$$J = i$$

   is chi-square with $k + 2i$ degrees of freedom. Then the unconditional distribution of $Z$ is non-central chi-square with $k$ degrees of freedom, and non-centrality parameter $\lambda$.

3.  <u>Bessel Function Based PDF Expression</u>: Alternatively, the PDF can be written as

$$f_X(x; k, \lambda) = \frac{1}{2} e^{-\frac{x+\lambda}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k}{4}-\frac{1}{2}} I_{\frac{k}{2}-1}\left(\sqrt{\lambda x}\right)$$

   where $I_v(y)$ is a modified Bessel function of the first kind given by

$$I_v(y) = \sum_{m=0}^{\infty} \frac{1}{m! \, \Gamma(m + v + 1)} \left(\frac{y}{2}\right)^{2m+v}$$

4. <u>Hypergeometric Function Based PDF</u>: Using the relation between the Bessel functions and the hyper-geometric functions, the PDF can also be written as (Muirhead (2005))

$$f_X(x; k, \lambda) = \frac{1}{2} e^{-\frac{\lambda}{2}} {}_0^1F\left(; \frac{k}{2}; \frac{\lambda x}{4}\right) \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}$$

Siegel (1979) discusses the case

$$k = 0$$

specifically – zero degrees of freedom – in which the distribution has a discrete component at zero.

## Properties – Moment Generating Function

The moment generating function is given by

$$M(t; k, \lambda) = \frac{e^{\frac{\lambda t}{1-2t}}}{(1 - 2t)^{\frac{k}{2}}}$$

for

$$2t < 1$$

## Properties – Moments

1. <u>First Four Non-central Moments</u>: The first few raw moments are

$$\mu'_1 = k + \lambda$$

$$\mu'_2 = (k + \lambda)^2 + 2(k + 2\lambda)$$

$$\mu'_3 = (k + \lambda)^3 + 6(k + \lambda)(k + 2\lambda) + 8(k + 3\lambda)$$

$$\mu'_4 = (k + \lambda)^4 + 12(k + \lambda)^2(k + 2\lambda) + 4(11k^2 + 44k\lambda + 36\lambda^2) + 48(k + 4\lambda)$$

2. <u>Second, Third, and Fourth Central Moments</u>: The first few central moments are

$$\mu_2 = 2(k + 2\lambda)$$

$$\mu_3 = 8(k + 3\lambda)$$

$$\mu_4 = 12(k + \lambda)^2 + 48(k + 4\lambda)$$

3. <u>Cumulant/Arbitrary Non-central Moment</u>: The $n^{th}$ cumulant is

$$K_n = 2^{n-1}(n - 1)!\,(k + n\lambda)$$

Hence

$$\mu'_n = 2^{n-1}(n - 1)!\,(k + n\lambda) + \sum_{j=1}^{n-1} 2^{j-1} \frac{(n - 1)!}{(n - j)!}(k + j\lambda)\mu'_{n-j}$$

## Cumulative Distribution Function

1. <u>Explicit Expression for the CDF</u>: Using the relation between the central and the non-central chi-square distributions, the cumulative CDF can be written as

$$P(x; k, \lambda) = e^{-\frac{\lambda}{2}} \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\lambda}{2}\right)^i Q(x; k + 2i)$$

where $Q(x; k)$ is the cumulative distribution function of the central chi-square distribution with $k$ degrees of freedom, which is given by

$$Q(x; k) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

and $\gamma(k, z)$ is the lower incomplete gamma function.

2. <u>CDF Based Marcum Q Function</u>: The Marcum Q-function $Q_M(a, b)$ can also be used to represent the CDF (Nuttall (1975)) as

$$P(x; k, \lambda) = 1 - Q_{\frac{k}{2}}\left(\sqrt{\lambda}, \sqrt{x}\right)$$

## Approximation – including for Quantiles

1.  <u>Abdel-Aty Non-central CDF Approximation</u>: Abdel-Aty (1954) derives – as a first approximation – a non-central Wilson-Haferty approximation. $\left(\frac{\chi'^2}{k+\lambda}\right)^{\frac{1}{3}}$ is approximately normally distributed as $\mathcal{N}\left(1 - \frac{2}{9f}, \frac{2}{9f}\right)$, i.e.,

$$P(x; k, \lambda) \approx \Phi\left(\frac{\left(\frac{\chi'^2}{k+\lambda}\right)^{\frac{1}{3}} - \left(1 - \frac{2}{9f}\right)}{\sqrt{\frac{2}{9f}}}\right)$$

where

$$f = \frac{(k+\lambda)^2}{k+2\lambda} = k + \frac{\lambda^2}{k+2\lambda}$$

which is quite accurate and well-adapting to non-centrality. Also

$$f = f(\lambda, k)$$

becomes

$$f = k$$

for

$$\lambda = 0$$

which is the central chi-square case.

2. <u>Sankaran Family of Normal Approximations</u>: Sankaran (1963) discusses a number of closed form approximations for the cumulative distribution function. In an earlier paper (Sankaran (1959)) he states and derives the following approximation:

$$P(x; k, \lambda) \approx \Phi\left(\frac{\left(\frac{x}{k+\lambda}\right)^h - \left(1 + hp(h - 1 - 0.5(2 - h)mp)\right)}{h\sqrt{2p}(1 + 0.5mp)}\right)$$

where

$$h = 1 - \frac{2}{3}\frac{(k+\lambda)(k+3\lambda)}{(k+2\lambda)^2}$$

$$f = \frac{k+2\lambda}{(k+\lambda)^2}$$

$$m = (h - 1)(1 - 3h)$$

This and the other approximations are discussed in Johnson, Klotz, and Balakrishnan (1995). For a given probability, these formulas are easily inverted to provide the corresponding approximation for $x$, to compute the approximation quantiles.

## Derivation of the PDF

1. <u>Spherically Symmetric Independent Gaussian Variates</u>: The derivation of the probability density function is done most easily by performing the following steps. First, since $X_1, \cdots, X_n$ have unit variances, their joint distribution is spherically symmetric up to a location shift.

2. <u>Dependence on the Squared Distribution Means</u>: The spherical symmetry implies that the distribution of

$$X = X_1{}^2 + \cdots + X_n{}^2$$

depends only on the means through the squared length

$$\lambda = \mu_1{}^2 + \cdots + \mu_n{}^2$$

Without loss of generality, one can therefore take

$$\mu_1 = \sqrt{\lambda}$$

and

$$\mu_2 = \mu_k = 0$$

3. Density Contributions from Positive/Negative: The next step is to derive the density of

$$X = X_1^2$$

for the

$$k = 1$$

case. Simple transformations of the random variables show that

$$f_X(x; 1, \lambda) = \frac{1}{2\sqrt{x}} \left[ \phi(\sqrt{x} - \sqrt{\lambda}) + \phi(\sqrt{x} + \sqrt{\lambda}) \right] = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x+\lambda}{2}} \cosh \sqrt{\lambda x}$$

4. Poisson Weighted Taylor Series Terms: Expand the $cosh$ term in a Taylor series. This gives the Poisson weighted mixture representation of the density, still for

$$k = 1$$

The indices on the chi-square random variables in the series above are $1 + 2i$ in this case.

5. <u>Connection to the Central Chi Square</u>: Finally, for the general case. It has been assumed, without loss of generality, that $X_2, \cdots, X_k$ are standard normal, and so $X_2^2 + \cdots + X_k^2$ has a *central* chi-square distribution with $k - 1$ degrees of freedom, independent of $X_1^2$. Using the Poisson-weighted mixture representation for $X_1^2$, and the fact that the sum of chi-square distributed random variables is also chi-square, completes the result. The indices in the series are

$$1 + 2i + k - 1 = k + 2i$$

as required.

**Related Distributions**

1. <u>Central Chi-square as a Special Case</u>: If $V$ is chi-squared distributed as

$$V \sim \chi_k^2$$

then $V$ is also non-central chi-square distributed

$$V \sim \chi'^2_k(0)$$

2. <u>Non central $F$ Distribution</u>: If

$$V_1 \sim \chi'^2_{k_1}(\lambda)$$

and

$$V_2 \sim \chi'^2_{k_2}(\lambda)$$

and $V_1$ is independent of $V_2$, then a non-central $F$ distributed variables is developed as

$$\frac{V_1/k_1}{V_2/k_2} = F'_{k_1,k_2}(\lambda)$$

3. <u>Poisson Conditioned Chi Square</u>: If

$$J \sim Poisson\ (\lambda)$$

then

$$\chi_{k+2J}{}^{2} \sim {\chi'}_{k}{}^{2}(\lambda)$$

4.  <u>Transformation to the Rice Distribution</u>: If

$$V \sim {\chi'}_{k}{}^{2}(\lambda)$$

then $\sqrt{V}$ takes the Rice distribution with the parameter $\sqrt{\lambda}$.

5.  <u>Approximation using the Normal Distribution</u>: As shown in Muirhead (2005), if

$$V \sim {\chi'}_{k}{}^{2}(\lambda)$$

then

$$\frac{V - (k + \lambda)}{\sqrt{2(k + 2\lambda)}} \to \mathcal{N}(0, 1)$$

in distribution as either

$$k \to \infty$$

or

$$\lambda \to \infty$$

6. <u>Independent Non central Chi Square Sum</u>:

$$V_1 \sim {\chi'}_{k_1}{}^2(\lambda_1)$$

and

$$V_2 \sim {\chi'}_{k_2}{}^2(\lambda_2)$$

and $V_1$ is independent of $V_2$, then

$$W = V_1 + V_2 \sim {\chi'}_k{}^2(\lambda_1 + \lambda_2)$$

where

$$k = k_1 + k_2$$

In general, for a finite set of

$$V_i \sim {\chi'}_{k_i}^{\ 2}(\lambda_i) \; i \in \{1, \cdots, N\}$$

the sum of the non-central chi-square distributed random variables

$$Y = \sum_{i=1}^{N} V_i$$

has the distribution

$$Y \sim {\chi'}_{k_y}^{\ 2}(\lambda_y)$$

where

$$k_y = \sum_{i=1}^{N} k_i$$

$$\lambda_y = \sum_{i=1}^{N} \lambda_i$$

This can be seen using the moment generating functions as follows:

$$M_Y(t) = M_{Y \sum_{i=1}^{N} V_i}(t) = \prod_{i=1}^{N} M_{Y_i}(t)$$

by the independence of the $V_i$ random variables. It remains to plug-in the MGF for the non-central chi-square distributions into the product and compute the new MGF. Alternatively, it can be seen via the interpretation in the background section above as sum of squares of independently distributed random variables with variance of $1$ and the specified means.

7. <u>Complex Non-central Chi-square</u>: The complex non-central chi-square has applications in radio communications and radar systems. Let $(z_1, \cdots, z_k)$ be independent scalar random variables of circular symmetry, means of $\mu_i$ and unit variances.

$$\mathbb{E}[|z_i - \mu_i|^2] = 1$$

Then the real random variable

$$S = \sum_{i=1}^{k} |z_i|^2$$

is distributed according to the complex non-central

199

$$f_S(S) = \frac{1}{2} e^{-(S+\lambda)} \left(\frac{S}{\lambda}\right)^{\frac{k-1}{2}} I_{k-1}(2\sqrt{Sx})$$

where

$$\lambda = \sum_{i=1}^{k} |\mu_i|^2$$

## Transformations

1. <u>Sankaran Cumulant Analysis and Transformations</u>: Sankaran (1963) discusses transformations of the form

$$z = \sqrt{\frac{X-b}{k+\lambda}}$$

He analyzes the expansions of the cumulants of $z$ up to the term $\mathcal{O}((k+\lambda)^{-4})$ and shows that the following choices of $b$ produce reasonable results.

    a.

$$b = \frac{k - 1}{2}$$

makes the second cumulant of $z$ approximately independent of $\lambda$

b.

$$b = \frac{k - 1}{3}$$

makes the third cumulant of $z$ approximately independent of $\lambda$

c.

$$b = \frac{k - 1}{4}$$

makes the fourth cumulant of $z$ approximately independent of $\lambda$

2.  <u>Variance Stabilizing Transformation of ${\chi'_k}^2(\lambda)$</u>: Also, a simpler transformation

$$z_1 = \sqrt{X - \frac{k - 1}{2}}$$

can be used as a variance stabilizing transformation that produces a random variable with a mean

$$z_2 = \sqrt{X + \frac{k-1}{2}}$$

and variance $\mathcal{O}((k+\lambda)^{-2})$. Usability of these transformations may be hampered by the need to take square roots of negative numbers.

## Use in Tolerance Intervals

Two-sided normal regression tolerance intervals can be obtained based on the non-central chi-square distribution (Young (2010)). This enables the calculation of a statistical interval within which, with some confidence level, a specified proportion of the sampled population falls.

## References

- Abdel-Aty, S. H. (1954): Approximate Formulae for the Percentage Points and the Probability Integral of the Non-central $\chi^2$ Distribution *Biometrika* **41 (3-4)** 538-540
- Johnson, N. L., S. Klotz, and N. Balakrishnan (1995): *Continuous Univariate Distributions 1 2nd Edition* **John Wiley and Sons**
- Muirhead, R. (2005): *Aspects of Multivariate Statistical Theory 2nd Edition* **Wiley**

- Nuttall, A. H. (1975): Some Integrals Involving the $Q_M$ Function *IEEE Transactions on Information Theory* **21 (1)** 95-96

- Sankaran, M. (1959): On the Non-central $\chi^2$ Distribution *Biometrika* **46 (1-2)** 235-237

- Sankaran, M. (1963): Approximations to the Non-central $\chi^2$ Distribution *Biometrika* **50 (1-2)** 199-204

- Siegel, A. F. (1979): The Non-central Chi-square Distribution with Zero Degrees of Freedom and Testing for Uniformity *Biometrika* **66 (2)** 381-386

- Wikipedia (2019): [Non-central Chi-square Distribution](#)

- Young, D. S. (2010): tolerance: An R Package for estimating Tolerance Intervals *Journal of Statistical Software* **36 (5)** 1-39

# Exponential Distribution

## Overview

1. <u>Definition of an Exponential Distribution</u>: The *exponential distribution* or *negative exponential distribution* is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate (Wikipedia (2023)).

2. <u>Key Traits of Exponential Distribution</u>: Exponential distribution is a particular case of gamma distribution. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless. In addition to being used for the analysis of Poisson point processes it is found in various other contexts.

3. <u>Member of Exponential Distribution Family</u>: The exponential distribution is not the same as the class of exponential families of distributions. This is a large class of probability distributions that includes the exponential distribution as one of its members, but also includes many other distributions, like normal, binomial, gamma, and Poisson distributions.

## Definitions – Probability Density Function

1. <u>Explicit Form of the PDF</u>:
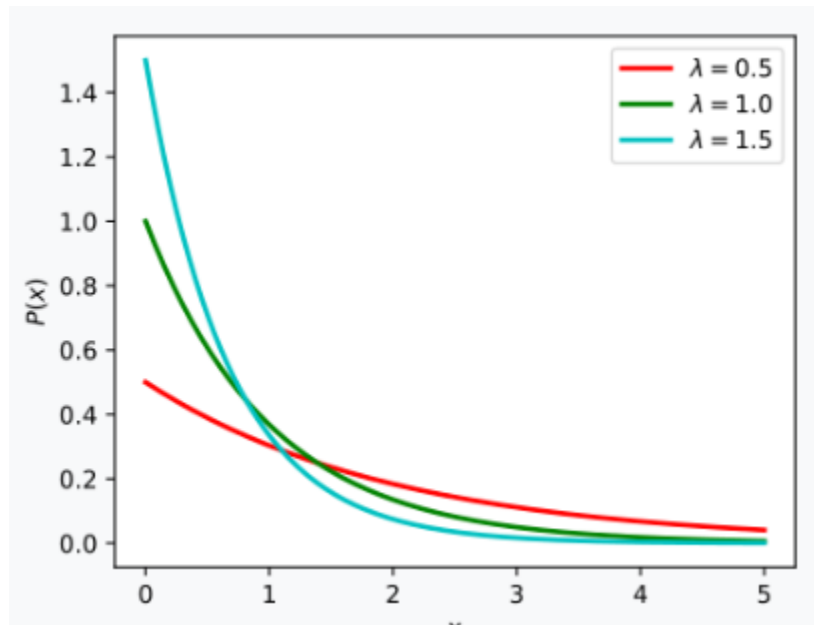
The probability density function pdf of an exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

2.  <u>Parametrization of the Distribution</u>: Here $\lambda > 0$ is the parameter of the distribution, often called the *rate parameter*. This distribution is supported on the interval $[0, \infty)$ If a random variable $X$ has this distribution, one writes

$$X \sim Exp(\lambda)$$

The exponential distribution exhibits infinite divisibility.

## Definition – Cumulative Distribution Function
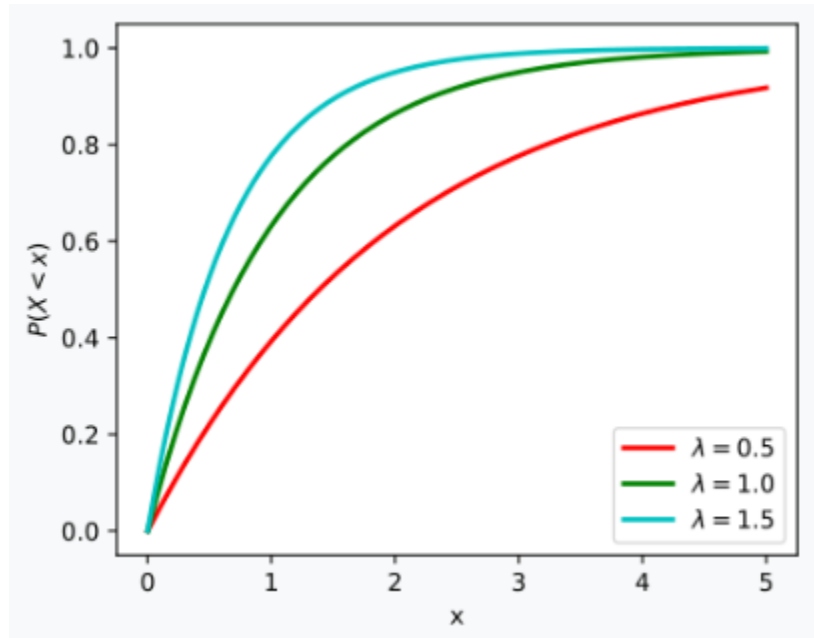
The cumulative distribution function is given by

$$F(x;\ \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

## Definition – Alternative Parameterization

The exponential distribution is sometimes parameterized in terms of the scale parameter

$$\beta = \frac{1}{\lambda}$$

$$f(x;\ \beta) = \begin{cases} \dfrac{1}{\beta} e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$F(x;\ \beta) = \begin{cases} 1 - e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

## Properties Table

1. <u>Parameters</u>:

$$\lambda > 0$$

   the rate or inverse scale parameter.

2. <u>Support</u>:

$$x \in [0, \infty)$$

3. <u>PDF</u>: $\lambda e^{-\lambda x}$

4. <u>CDF</u>: $1 - e^{-\lambda x}$

5. <u>Quantile</u>: $-\frac{\ln(1-p)}{\lambda}$

6. <u>Mean</u>: $\frac{1}{\lambda}$

7. <u>Median</u>: $\frac{\ln 2}{\lambda}$

8. <u>Mode</u>: $0$

9. <u>Variance</u>: $\frac{1}{\lambda^2}$

10. <u>Skewness</u>: $2$

11. <u>Ex. Kurtosis</u>: $6$

12. <u>MGF</u>: $\frac{\lambda}{\lambda - t}$ for

$$t < \lambda$$

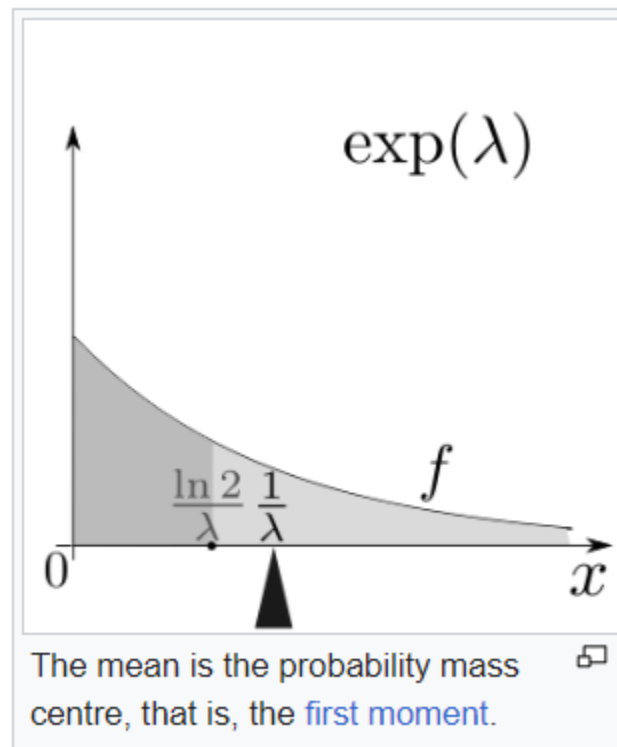13. <u>Fisher Information</u>: $\frac{1}{\lambda^2}$

14. <u>Kullback-Leibler Divergence</u>: $\ln\frac{\lambda_0}{\lambda} + \frac{\lambda_0}{\lambda} - 1$

15. <u>CVaR (Expected Shortfall)</u>: $-\frac{\ln(1-p)+1}{\lambda}$

16. <u>bPOE</u>: $e^{1-\lambda x}$

## Properties – Mean, Variance, Moments, Variance

1. <u>Mean of an Exponential Distribution</u>:



The mean is the probability mass centre, that is, the first moment.

The mean or expected value of an exponentially distributed random variable $X$ with rate parameter $\lambda$ is given by

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

2. <u>Variance and Standard Deviation</u>: The variance of $X$ is given by

$$\mathbb{V}[X] = \frac{1}{\lambda^2}$$

so that the standard deviation is equal to the mean.

3. <u>Moments of an Exponential Distribution</u>: The moments of $X$, for

$$n \in \mathbb{N}$$

are given by

$$\mathbb{E}[X^n] = \frac{n!}{\lambda^n}$$

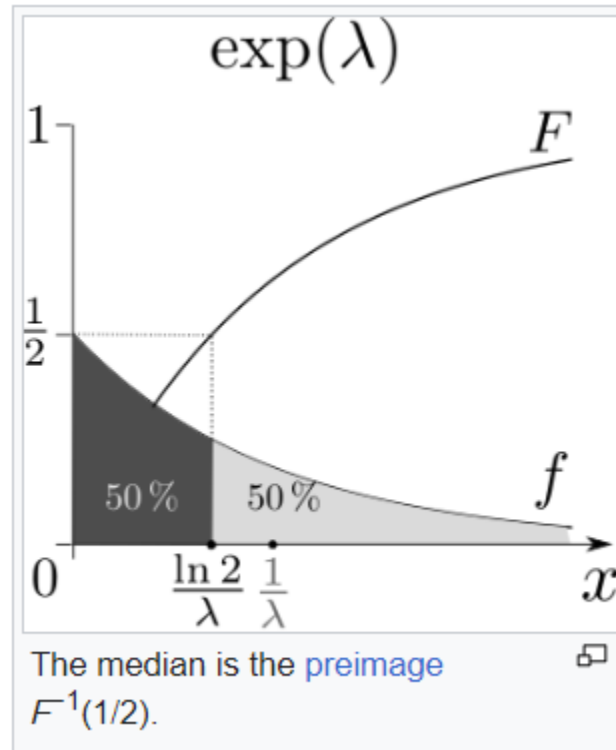4. <u>Central Moments - Use of Subfactorial</u>: The central moments of $X$, for

$$n \in \mathbb{N}$$

are given by

$$\mu_n = \frac{!n}{\lambda^n} = \frac{n!}{\lambda^n} \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

where $!n$ is the subfactorial of $n$.

5. <u>Comparison of Median and Mean</u>:

$$\exp(\lambda)$$

The median is the preimage $F^{-1}(1/2)$.

The median of $X$ is given by

$$\mathbb{M}[X] = \frac{\ln 2}{X} < \mathbb{E}[X]$$

Thus, the absolute difference between the mean and the median is

$$|\mathbb{E}[X] - \mathbb{M}[X]| = \frac{1 - \ln 2}{\lambda} < \frac{1}{\lambda} = \mathbb{S}[X]$$

in accordance with the mean-median inequality.

## Memorylessness Property of Exponential Random Variables

1. Mathematical Statement of Memorylessness Property: An exponentially distributed random variable $T$ obeys the relation

$$\mathbb{P}[T > s + t \mid T > s] = \mathbb{P}[T > t] \; \forall \, s, t \geq 0$$

2. Derivation using Complementary Cumulative Distribution: This can be seen by considering the complementary cumulative distribution function:

$$\mathbb{P}[T > s + t \mid T > s] = \frac{\mathbb{P}[T > s + t]}{\mathbb{P}[T > s]} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}[T > t]$$

3. Intuition behind the Memorylessness Property: When $T$ is interpreted as the waiting time for an event to occur relative to some starting time, the above relation implies that, if $T$ is conditioned on a failure to observe an event over some initial period of time $s$, the distribution of the remaining waiting time is the same as the original unconditional distribution.

4. Practical Illustration of the Property: For example, if an event has not occurred after 30 seconds, the conditional probability that the occurrence will take at least 10 more seconds is equal to the unconditional probability of observing the event more than 10 seconds after the initial time.

5. The Only Memoryless Distribution: The exponential distribution and the geometric distributions are the only distributions that are memoryless.

6. Distribution with Fixed Failure Rate: The exponential distribution is consequently also necessarily the only continuous distribution that has a constant failure rate.

## Quantiles

1. Inverse Cumulative Distribution Function: The quantile function, i.e., the inverse cumulative distribution function, for $Exp\,(\lambda)$ is $F^{-1}(p;\, \lambda) = -\frac{\ln(1-p)}{\lambda} \; 0 \leq p < 1$
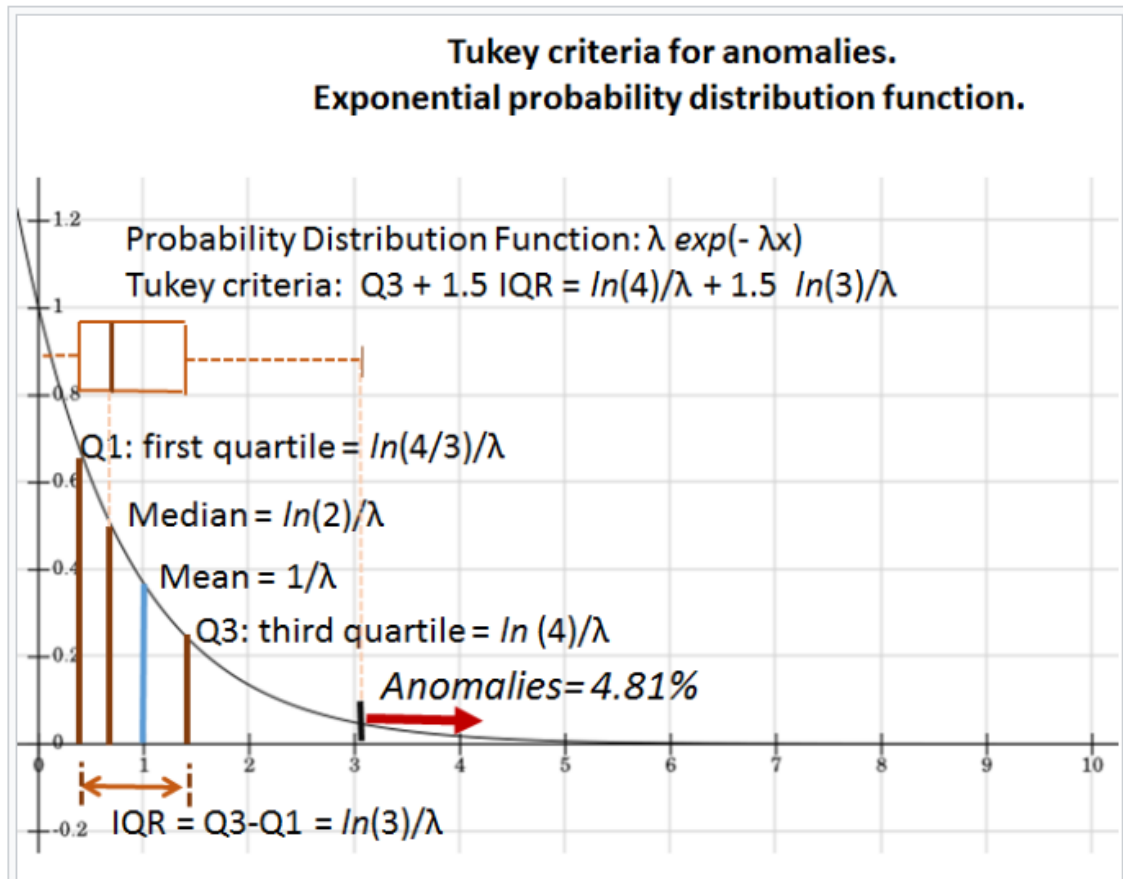
2. <u>First, Second, and Third Quantiles</u>: The quantiles are therefore:

    a. First Quantile - $\dfrac{\ln\frac{4}{3}}{\lambda}$

    b. Second Quantile/median - $\dfrac{\ln 2}{\lambda}$

    c. Third Quantile - $\dfrac{\ln 4}{\lambda}$

3. <u>The Inter-quantile Range</u>: And as a consequence, the inter-quantile range is $\dfrac{\ln 3}{\lambda}$

4. <u>Tukey Criteria for Anomalies</u>:



Tukey criteria for anomalies.
Exponential probability distribution function.

Probability Distribution Function: $\lambda\,exp(-\lambda x)$
Tukey criteria:  Q3 + 1.5 IQR = $In(4)/\lambda$ + 1.5  $In(3)/\lambda$

Q1: first quartile = $In(4/3)/\lambda$

Median = $In(2)/\lambda$

Mean = $1/\lambda$

Q3: third quartile = $In\,(4)/\lambda$

Anomalies = 4.81%

IQR = Q3-Q1 = $In(3)/\lambda$

## Conditional Value at Risk (Expected Shortfall)

1. <u>Definition of CVaR/Expected Shortfall</u>: The conditional value-at-risk – CVaR – also known as the expected shortfall or super-quantile for $Exp\,(\lambda)$ is derived as follows (Norton, Khokhlov, and Uryasev (2019)).

2. <u>Derivation of the Distribution CVaR</u>:

$$\bar{q}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 q_p(X)dp = \frac{1}{1-\alpha}\int_\alpha^1 -\frac{\ln(1-p)}{\lambda}dp = -\frac{1}{\lambda(1-\alpha)}\int_{1-\alpha}^0 -\ln y\,dy$$

$$= -\frac{1}{\lambda(1-\alpha)}\int_0^{1-\alpha}\ln y\,dy$$

$$= -\frac{1}{\lambda(1-\alpha)}[(1-\alpha)\ln(1-\alpha)-(1-\alpha)] = \frac{1-\ln(1-\alpha)}{\lambda}$$

## Buffered Probability of Exceedance bPOE

1. <u>Defining Buffered Probability of Exceedance</u>: The buffered probability of exceedance is one minus the probability level at which the CVaR equals the threshold $x$. It is derived as follows (Norton, Khokhlov, and Uryasev (2019)).

2. <u>Explicit Expression for the bPOE</u>:

$$\bar{p}_x(X) = \{1-\alpha \mid \bar{q}_\alpha(X) = x\} = \left\{1-\alpha \mid \frac{1-\ln(1-\alpha)}{\lambda} = x\right\}$$

$$= \{1-\alpha \mid \ln(1-\alpha) = 1-\lambda x\} = \{1-\alpha \mid e^{\ln(1-\alpha)} = e^{1-\lambda x}\}$$

$$= \{1-\alpha \mid 1-\alpha = e^{1-\lambda x}\} = e^{1-\lambda x}$$

## Kullback-Leibler Convergence

1. Directed Kullback-Leibler Divergence: The directed Kullback-Leibler divergence in nats of $e^\lambda$ – "approximating" distribution – from $e^{\lambda_0}$ – "true" distribution – is given as follows.

2. Explicit Expression for the Divergence:

$$\Delta(\lambda_0 \mid\mid \lambda) = \mathbb{E}_{\lambda_0}\left[\log\frac{\mathbb{P}_{\lambda_0}[x]}{\mathbb{P}_\lambda[x]}\right] = \mathbb{E}_\lambda\left[\log\frac{\lambda_0 e^{\lambda_0 x}}{\lambda e^{\lambda x}}\right] = \log\lambda_0 - \log\lambda - (\lambda_0 - \lambda)\mathbb{E}_{\lambda_0}[x]$$

$$= \log\lambda_0 - \log\lambda + \frac{\lambda}{\lambda_0} - 1$$

## Maximum Entropy Distribution

1. Distribution with Largest Differential Entropy: Among all continuous probability distributions with support $[0, \infty)$ and mean $\mu$ the exponential distribution with

$$\lambda = \frac{1}{\mu}$$

has the largest differential entropy.

2. Maximum Entropy Distribution for $X > 0$:In other words, it is the maximum entropy probability distribution for a random variable $X$ which is greater than or equal to zero and for which $\mathbb{E}[X]$ is fixed (Park and Bera (2009)).

## Distribution of the Minimum of Exponential Random Variables

1. Minimum Distribution is also Exponential: Let $X_1, \cdots, X_n$ be independent exponentially distributed random variables with rate parameters $\lambda_1, \cdots, \lambda_n$. Then $\min(X_1, \cdots, X_n)$ is also exponentially distributed, with parameter

$$\lambda = \lambda_1 + \cdots + \lambda_n$$

2. <u>Derivation using Complementary Cumulative Distribution</u>: This can be seen considering the complementary cumulative distribution function:

$$\mathbb{P}[\min(X_1, \cdots, X_n) > X] = \mathbb{P}[X_1 > x, \cdots, X_n > x] = \prod_{i=1}^{n} \mathbb{P}[X_i > x] = \prod_{i=1}^{n} e^{-x\lambda_i}$$

$$= e^{-x\sum_{i=1}^{n} \lambda_i}$$

3. <u>Distribution of Realized Minimal Index</u>: The index of the variable which achieves the minimum is distributed according to the categorical distribution

$$\mathbb{P}[X_k = \min(X_1, \cdots, X_k)] = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n}$$

4. <u>Derivation of Minimal Index Distribution</u>: A proof can be seen by letting

$$I = \mathop{\arg\min}_{i \in (1, \cdots, n)} \{X_1, \cdots, X_n\}$$

Then

$$\mathbb{P}[I = k] = \int_0^{\infty} \mathbb{P}[X_k = x]\mathbb{P}[\forall_{i \neq k} X_i > x]dx = \int_0^{\infty} \lambda_k e^{-\lambda_k x}\left(\prod_{i=1, i \neq k}^{n} e^{-\lambda_i x}\right)dx$$

$$= \lambda_k \int_0^{\infty} e^{-(\lambda_1 + \cdots + \lambda_n)x}dx = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n}$$

5. <u>Maximum Distribution is not Exponential</u>: Note that $\min(X_1, \cdots, X_n)$ is not exponentially distributed, if $X_1, \cdots, X_n$ do not all have parameter zero.

215

## Joint Moments of i.i.d. Exponential Order Statistics

1. <u>Independent Exponential Variable Order Statistics</u>: Let $X_1, \cdots, X_n$ be $n$ independent and identically distributed exponential random variables with rate parameter $\lambda$. Let $X_{(1)}, \cdots, X_{(n)}$ denote the corresponding order statistics.

2. <u>Expression for Joint Order Statistic</u>: For

$$i < j$$

the joint moment $\mathbb{E}\left[X_{(i)}X_{(j)}\right]$ of the order statistic $X_{(i)}$ and $X_{(j)}$ is given by

$$\mathbb{E}\left[X_{(i)}X_{(j)}\right] = \sum_{k=0}^{j-1} \frac{1}{(n-k)\lambda}\mathbb{E}\left[X_{(i)}\right] + \mathbb{E}\left[X_{(i)}^2\right]$$

$$= \sum_{k=0}^{j-1} \frac{1}{(n-k)\lambda} \sum_{k=0}^{i-1} \frac{1}{(n-k)\lambda} + \sum_{k=0}^{i-1} \frac{1}{[(n-k)\lambda]^2} + \left[\sum_{k=0}^{i-1} \frac{1}{(n-k)\lambda}\right]^2$$

3. <u>Derivation of Joint Order Statistic</u>: This can be seen by invoking the law of total expectation and the memoryless property:

$$\mathbb{E}\left[X_{(i)}X_{(j)}\right] = \int_0^\infty \mathbb{E}\left[X_{(i)}X_{(j)} \mid X_{(i)} = x\right]f_{X_{(i)}}(x)dx = \int_0^\infty x\mathbb{E}\left[X_{(j)} \mid X_{(j)} \geq x\right]f_{X_{(i)}}(x)dx$$

$$= \int_0^\infty x\left(\mathbb{E}\left[X_{(j)}\right] + x\right)f_{X_{(i)}}(x)dx = \sum_{k=0}^{j-1} \frac{1}{(n-k)\lambda}\mathbb{E}\left[X_{(i)}\right] + \mathbb{E}\left[X_{(i)}^2\right]$$

4. <u>Law of Total Expectation</u>: The first equation follows from the law of total expectation.

5. <u>Applying the Order Statistic Ordinal</u>: The second equation exploits the fact that once one conditions on

$$X_{(i)} = x$$

it must follow that

$$X_{(j)} \geq x$$

6. <u>Use of the Memoryless Property</u>: The third equation relies on the memoryless property to replace $\mathbb{E}[X_{(j)} \mid X_{(j)} \geq x]$ with $\mathbb{E}[X_{(j)}] + x$

## Sum of Two Independent Exponential Random Variables

1. <u>Convolution of the Individual Distributions</u>: The probability distribution function pdf of a sum of two independent random variables is the convolution of their individual pdf's.

2. <u>Probability Density of the Sum</u>: If $X_1$ and $X_2$ are independent random variables with respective rate parameters $\lambda_1$ and $\lambda_2$, then the probability density of

$$z = X_1 + X_2$$

is given by

$$f_Z(z) = \int\limits_{-\infty}^{+\infty} f_{X_1}(x_1) f_{X_2}(z - x_1) dx_1 = \int\limits_{0}^{z} \lambda_1 e^{-\lambda_1 x_1} \lambda_2 e^{-\lambda_2(z-x_1)} dx_1$$

$$= \lambda_1 \lambda_2 e^{-\lambda_2 z} \int\limits_{0}^{z} e^{(\lambda_2 - \lambda_1)x} dx_1$$

$$= \begin{cases} \dfrac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \left( e^{-\lambda_1 z} - e^{-\lambda_2 z} \right) & \lambda_1 \neq \lambda_2 \\ \lambda^2 z e^{-\lambda z} & \lambda = \lambda_1 = \lambda_2 \end{cases}$$

3. <u>Entropy of the Probability Density</u>: The entropy of this distribution is available in closed form: assuming

$$\lambda_1 > \lambda_2$$

without loss of generality, then

$$H(z) = 1 + \gamma + \ln\frac{\lambda_1 - \lambda_2}{\lambda_1 \lambda_2} + \Psi\left(\frac{\lambda_1}{\lambda_1 - \lambda_2}\right)$$

where $\gamma$ is the Euler-Mascheroni constant, and $\Psi(\cdot)$ is the digamma function (Eckford and Thomas (2016)).

4. <u>Equal Rate Results in Erlang Distribution</u>: In the case of equal rate parameters, the result is an Erlang distribution with shape 2 and parameter $\lambda$, which in turn is a special case of the gamma distribution.

## Related Distributions

1. <u>Laplace #1</u>: If

$$X \sim Laplace\left(\mu, \beta^{-1}\right)$$

then

$$|X - \mu| \sim Exp\,(\beta)$$

2. <u>Pareto #1</u>: If

$$X \sim Pareto\,(1, \lambda)$$

then

$$\log X \sim Exp\,(\lambda)$$

3. <u>Skew Logistic</u>: If

$$X \sim SkewLogistic\,(0)$$

then

$$\log(1 + e^{-X}) \sim Exp\,(0)$$

4. <u>Uniform</u>: If

$$X_i \sim U\,(0, 1)$$

then

$$\lim_{n \to \infty} n \cdot \min(X_1, \cdots, X_n) \sim Exp\,(1)$$

5. <u>Scaled Beta</u>: The exponential distribution is the limit of a scaled beta distribution:

$$\lim_{n \to \infty} n \cdot Beta\,(1, n) = Exp\,(1)$$

6. <u>Type 3 Pearson</u>: Exponential distribution is a special case of type 3 Pearson distribution.

7. <u>Closure under Scaling</u>:

$$kX \sim Exp\left(\frac{\lambda}{k}\right)$$

closure under scaling by a positive factor.

8. <u>Truncated Exponential</u>:

$$1 + X \sim BenktanderWeibull\,(\lambda, 1)$$

which reduces to a truncated exponential distribution.

9. <u>Pareto #2</u>:

$$ke^X \sim Pareto\,(k, \lambda)$$

10. <u>Beta</u>:

$$e^{-X} \sim Beta\,(X, 1)$$

11. <u>Power Law</u>:

$$\frac{1}{k} e^X \sim PowerLaw\,(k, \lambda)$$

12. <u>Rayleigh</u>:

$$\sqrt{X} \sim Rayleigh\left(\frac{1}{\sqrt{2\lambda}}\right)$$

the Rayleigh distribution.

13. <u>Weibull #1</u>:

$$X \sim Weibull\left(\frac{1}{\lambda}, 1\right)$$

the Weibull distribution.

14. <u>Weibull #2</u>:

$$X^2 \sim Weibull\left(\frac{1}{\lambda^2}, \frac{1}{2}\right)$$

15. <u>Gumbel</u>:

$$\mu - \beta \log(\lambda X) \sim Gumbel\,(\mu, \beta)$$

16. <u>Gumbel #1</u>:

$$\lfloor X \rfloor = Geometric\left(1 - e^{-\lambda}\right)$$

a geometric distribution on $0, 1, 2, 3, \cdots$

17. <u>Gumbel #2</u>:

$$\lceil X \rceil = Geometric\left(1 - e^{-\lambda}\right)$$

a geometric distribution on $1, 2, 3, 4, \cdots$

18. <u>Erlang</u>: If

$$Y \sim Erlang\ (n, \lambda)$$

or

$$Y \sim \Gamma\ \left(n, \frac{1}{\lambda}\right)$$

then

$$\frac{X}{Y} + 1 \sim Pareto\ (1, n)$$

19. <u>Gamma in Shape, Scale Parameterization</u>: If

$$\lambda \sim Gamma\ (k, \theta)$$

then the marginal distribution of $X$ is $Lomax\ \left(k, \frac{1}{\theta}\right)$ i.e., the gamma mixture.

20. <u>Gamma in $k/\theta$ or $\alpha/\beta$ Parameterization</u>: If

$$\lambda_i = \lambda$$

then

$$X_1 + \cdots + X_k = \sum_i X_i \sim Erlang\ (k, \lambda) = Gamma\ \left(k, \frac{1}{\lambda}\right)$$

in the $(k, \theta)$ and $(\alpha, \beta)$ parameterization respectively, with an integer shape parameter $k$ (Ibe (2014)).

21. <u>Laplace #2</u>:

$$X_i - X_j \sim Laplace\left(0, \frac{1}{\lambda}\right)$$

22. <u>Independent $X_i$</u>: If $X_i$ are independent, then

$$\frac{X_i}{X_i + X_j} \sim U(0,1)$$

23. <u>Ratio of Independent $X_i$</u>:

$$z = \frac{\lambda_i X_i}{\lambda_j X_j}$$

has the probability density function

$$f_Z(z) = \frac{1}{(z+1)^2}$$

This can be used to obtain a confidence interval for $\frac{\lambda_i}{\lambda_j}$.

24. <u>Logistic #1</u>: If

$$\lambda = 1$$

$$\mu - \beta \log \frac{e^{-X}}{1 - e^{-X}} \sim Logistic(\mu, \beta)$$

the logistic distribution.

25. <u>Logistic #2</u>: If

$$\lambda = 1$$

$$\mu - \beta \log \frac{X_i}{X_j} \sim Logistic\ (\mu, \beta)$$

26. <u>GEV</u>: If

$$\lambda = 1$$

$$\mu - \sigma \log X \sim GEV\ (\mu, \sigma, 0)$$

27. <u>K-Distribution</u>: If

$$\lambda = 1$$

and

$$Y \sim \Gamma\left(\alpha, \frac{\beta}{\alpha}\right)$$

then

$$\sqrt{XY} \sim K\ (\alpha, \beta)$$

is the K-distribution.

28. <u>Poisson and Geometric</u>: If

$$X \sim Exp\left(\frac{1}{\lambda}\right)$$

and

$$\frac{Y}{X} \sim Poisson(X)$$

then

$$Y \sim Geometric\left(\frac{1}{1+\lambda}\right)$$

i.e., the geometric distribution.

29. Hoyt: The Hoyt distribution can be obtained from exponential distribution and arcsine distribution.

30. Limit of the k-exponential: The exponential distribution is a limit of the K-exponential distribution in the

$$K = 0$$

case.

31. $\kappa$-Generalized Gamma: Exponential distribution is a limit of the $\kappa$-Generalized Gamma distribution in the

$$\alpha = 1$$

and

$$\nu = 1$$

cases:

$$\lim_{(\alpha,\nu)\to(0,1)} p_\kappa(x) = (1+\kappa\nu)(2\kappa)^\nu \frac{\Gamma\left(\frac{1}{2\kappa}+\frac{\nu}{2}\right)}{\Gamma\left(\frac{1}{2\kappa}-\frac{\nu}{2}\right)} \cdot \frac{\alpha\lambda^\nu}{\Gamma(\nu)} x^{\alpha\nu-1} \exp_\kappa(-\lambda x^\alpha)$$

$$= \lambda e^{-\lambda x}$$

32. <u>Hyper-exponential Distribution</u>: Distribution whose density is a weighted sum of exponential densities.

33. <u>Hypo-exponential Distribution</u>: Distribution of a general sum of exponential densities.

34. <u>ex-Gaussian Distribution</u>: Sum of exponential and a normal distribution.

## Statistical Distribution

In the following section, it is supposed that the random variable $X$ is exponentially distributed with rate parameter $\lambda$, and $x_1, \cdots, x_n$ are $n$ independent samples from $X$, with sample mean $\overline{x}$.

## Parameter Estimation

1. <u>Maximum Likelihood Estimator for $\lambda$</u>: The maximum likelihood for $\lambda$ is constructed as follows. The likelihood function for $\lambda$, given an independent and identically distributed sample

$$z = (x_1, \cdots, x_n)$$

drawn from the variable, is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-\lambda n \bar{x}}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. <u>Derivative of Likelihood Function's Logarithm</u>: The derivative of the likelihood function's logarithm is

$$\frac{d}{d\lambda} \ln L(\lambda) = \frac{d}{d\lambda} (n \ln \lambda - \lambda n \bar{x}) = \frac{n}{\lambda} - n\bar{x} \begin{cases} > 0 & 0 < \lambda < \frac{1}{\bar{x}} \\ = 0 & \lambda = \frac{1}{\bar{x}} \\ < 0 & \lambda > \frac{1}{\bar{x}} \end{cases}$$

3. <u>MLE for the Rate Parameter</u>: Consequently, the maximum likelihood estimate for the rate parameter is

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{x}} = \frac{n}{\sum_i x_i}$$

4. <u>Unbiased Nature of the Estimator</u>: This is *not* an unbiased estimator of $\lambda$, although $\bar{x}$ is an unbiased MLE (Johnson and Wichern (2007)) estimator of $\frac{1}{\lambda}$ and the distribution mean.

5. <u>Bias in the MLE Parameter</u>: The bias of $\hat{\lambda}_{MLE}$ is equal to

$$B = \mathbb{E}\big[\hat{\lambda}_{MLE} - \lambda\big] = \frac{\lambda}{n-1}$$

which yields the bias-corrected maximum likelihood estimator

$$\hat{\lambda}^*_{MLE} = \hat{\lambda}_{MLE} - B$$

6. <u>Approximate Minimizer of MSE</u>: An approximate minimizer of mean-squared error can be found, assuming a sample size greater than two, with a correction factor to the MLE:

$$\hat{\lambda} = \left(\frac{n-2}{n}\right)\left(\frac{1}{\bar{x}}\right) = \frac{n-2}{\sum_i x_i}$$

7. <u>Derivative from the Gamma Distribution</u>: This is derived from the mean and the variance of the inverse-gamma distribution $Inv - Gamma\ (n, \lambda)$ (Elfessi and Reineke (2001)).

## Fisher Information

1. <u>Definition of the Fisher Information</u>: The Fisher information, denoted $\mathcal{I}(\lambda)$, for an estimation of the rate parameter $\lambda$ is given by

$$\mathcal{I}(\lambda) = \mathbb{E}\left[\left\{\frac{\partial}{\partial\lambda}\log f(x;\lambda)\right\}^2 \mid \lambda\right] = \int \left\{\frac{\partial}{\partial\lambda}\log f(x;\lambda)\right\}^2 f(x;\lambda)d\lambda$$

2. <u>Fisher Information for Exponential Distribution</u>:

$$\mathcal{I}(\lambda) = \int_0^\infty \left(\frac{\partial}{\partial\lambda}\log \lambda e^{-\lambda x}\right)^2 \lambda e^{-\lambda x}dx = \int_0^\infty \left(\frac{1}{\lambda} - x\right)^2 \lambda e^{-\lambda x}dx = \frac{1}{\lambda^2}$$

3. <u>Intuition behind the Fisher Information</u>: This determines the amount of information each independent sample of an exponential distribution carries about the unknown rate parameter $\lambda$.

## Confidence Intervals

1. <u>Confidence Interval for Rate Parameter</u>: The $100(1 - \alpha)$ confidence interval for the rate parameter of an exponential distribution is given by (Ross (2009)):

$$\frac{2n}{\hat{\lambda}\chi^2_{1-\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n}{\hat{\lambda}\chi^2_{\frac{\alpha}{2},2n}}$$

which is also equal to

$$\frac{2n\bar{x}}{\chi^2_{1-\frac{\alpha}{2},2n}} < \frac{1}{\lambda} < \frac{2n\bar{x}}{\chi^2_{\frac{\alpha}{2},2n}}$$

where $\chi^2_{p,v}$ is the $100(p)$ percentile of the chi-squared distribution with $v$ degrees of freedom, $n$ is the number of observations of inter-arrival times in the sample, and x-bar is the sample average.

2. <u>Approximation for the Confidence Intervals</u>: A simple example to the exact interval endpoints can be derived using a normal approximation to the $\chi^2_{p,v}$ distribution.

3. <u>Expression for the 95% Interval</u>: This approximation gives the following values for a 95% confidence interval:

$$\lambda_{LOWER} = \hat{\lambda}\left(1 - \frac{1.96}{\sqrt{n}}\right)$$

$$\lambda_{UPPER} = \hat{\lambda}\left(1 + \frac{1.96}{\sqrt{n}}\right)$$

4. Sample Size Required for Estimation: This approximation may be acceptable for samples containing at least 15 to 20 elements (Guerriero (2012)).

## Bayesian Inference

1. Conjugate Prior for the Exponential Distribution: The conjugate prior for the exponential distribution is the gamma distribution – of which the exponential distribution is a special case.

2. Parameterizing the Gamma Probability Density: The following parameterization of the gamma probability density is useful:

$$Gamma(\lambda;\ \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$$

3. The Corresponding Posterior Distribution: The posterior distribution $p$ can then be defined in terms of the likelihood function defined above and a gamma prior:

$$p(\lambda) \propto L(\lambda) \propto \Gamma(\lambda;\ \alpha, \beta) = \lambda^n e^{-\lambda n \bar{x}}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta} \propto \lambda^{\alpha+n-1}e^{-\lambda(\beta+n\bar{x})}$$

The posterior property $p(\lambda)$ above has been specified up to a missing normalizing constant.

4. Posterior Under the Gamma Parameterization: Since it has the form of a gamma pdf, this can be filled in, and one obtains

$$p(\lambda) = Gamma(\lambda, \alpha + n, \beta + n\bar{x})$$

5.  Interpretation of the Hyperparameter $\alpha$: Here the hyper-parameter $\alpha$ can be interpreted as the number of prior observations, and $\beta$ as the sum of the prior observations.

6.  Explicit Form for Posterior Mean: The posterior mean here is $\frac{\alpha+n}{\beta+n\bar{x}}$.

## Occurrence and Applications – Occurrence of Events

1.  Occurrence of the Exponential Distribution: The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogenous Poisson process.

2.  Relation to the Geometric Distribution: The exponential distribution may be viewed as a continuous counterpart of the geometric distribution, which describes the number of Bernoulli trials necessary for a *discrete* process to change state. In contract, the exponential distribution describes the times for a continuous process to change state.

3.  Assumption of a Constant Rate: In real-world scenario, the assumption of a constant rate – or probability per unit time – is rarely satisfied. For example, the rate of incoming phone calls differs according to the time of the day.

4.  Regime where Rate is Constant: But if one focuses on a time during which the rate is roughly constant, such as from 2 PM to 4 PM during work days, the exponential distribution can be used as a good approximate model for the time until the next phone call services. Similar caveats apply to the following examples which yield approximately exponentially distributed variables.

5.  Example #1 - Geiger Counter: The time until a radio-active particle decays, or the time between clicks of a Geiger counter.

6.  Example #2 - Telephone Call: The time it takes before the next telephone call.

7.  Example #3 - Reduced Form Credit Risk Model: The time until default – on payment to company debt-holders – in reduced-form credit risk modeling.

8.  Other Instances of Exponential Distribution Occurrence: Exponential variables can also be used to model situations where certain events occur with a constant

probability per unit length, such as the distance between the mutations on a DNA strand, or between road-kills on a given road.

9.  Service Times in Queuing Theory: In queuing theory, the service times of agents in a system – e.g., how long it takes for a bank teller etc. to serve a customer – one often modeled as exponentially distributed variables.

10. Comparison with Poisson Distribution: The arrival of customers, for instance, is also modeled by the Poisson distribution if the arrivals are independent and distributed identically.

11. Relation to the Erlang Distribution: The length of a process that can be thought of as a sequence of several independent tasks follows the Erlang distribution, which is the distribution of the sum of several independent exponentially distributed variables.

12. Use in Reliability Theory/Engineering: Reliability theory and reliability engineering also make extensive use of the exponential distribution.

13. Example Usage in Reliability Theory: Because of the *memoryless* property of this distribution, it is well-suited to model the constant hazard rate portion of the bathtub curve used in reliability theory.

14. Suitability in Applying to Reliability Models: It is also very convenient because it is so easy to add failure rates in a reliability model.

15. Situations where it is Unsuitable: The exponential distribution is, however, not appropriate to model the overall lifetime of organisms or technical devices, because the "failure rates" here are not constant: more failures occur for very young and for very old systems.

16. Application in the Barometric Formula: In physics, if one observes a gas at a fixed temperature and pressure in a uniform gravitational field, the heights of various molecules also follow an approximate exponential distribution, known as the Barometric formula. This is a consequence of the entropy property.

17. Use in Hydrology: In hydrology, the exponential distribution is used to analyze extreme values of such variables as monthly and annual maximum values of daily rainfall and river discharge volumes (Ritzema (1994)).

18. Cumulative Exponential Distribution Fitted:
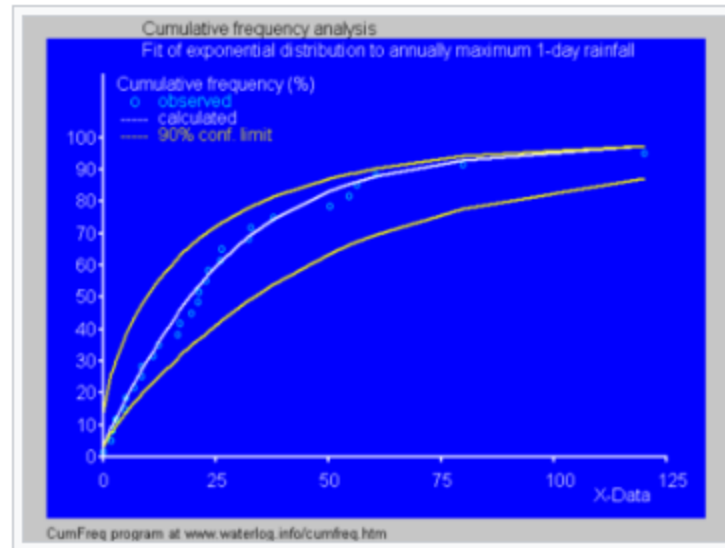
Fitted cumulative exponential to annual maximum 1-day rainfalls.

19. <u>Example: Use in Analyzing One-day Rainfalls</u>: The picture above illustrates an example of fitting the exponential distribution to ranked annually maximum one-day rainfalls showing also the 90% confidence belt on the binomial distribution. The rainfall data are represented by plotting positions as part of the cumulative frequency analysis.

20. <u>Distribution of Surgery Duration</u>: In operating-rooms management, the distribution of surgery duration for a category of surgeries with no typical work-content – like in an emergency room, encompassing all types of surgeries.

## Prediction

1. <u>Predictions from Unknown Exponential Distribution</u>: Having observed a sample of $n$ data points from an unknown exponential distribution a common task is to use these samples to make predictions about future data from the same source.

2. <u>Estimation of $\lambda$ from Data</u>: A common predictive distribution over future samples is the so-called plug-in distribution, formed by plugging a suitable estimate for the rate parameter $\lambda$ into the exponential density function.

3. <u>$\lambda$ from the MLE Principle</u>: A common choice of estimate is the one provided by the principle of maximum likelihood, and using this yields the predictive density over a future sample $x_{n+1}$, conditioned on the observed samples

$$x = (x_1, \cdots, x_n)$$

given by

$$p_{ML}(x_{n+1} \mid x_1, \cdots, x_n) = \frac{1}{\bar{x}} e^{-\frac{x_{n+1}}{\bar{x}}}$$

4. <u>$\lambda$ from a Bayesian Approach</u>: The Bayesian approach provides a predictive distribution which takes into account the uncertainty of the estimated parameter, although this may depend crucially on the choice of prior.

5. <u>Bayesian Approach from Uninformed Prior</u>: A predictive distribution free of the issues of choosing priors that arise under the subjective Bayesian approach is

$$p_{CNML}(x_{n+1} \mid x_1, \cdots, x_n) = \frac{n^{n+1}\bar{x}^n}{(n\bar{x} + x_{n+1})^{n+1}}$$

which can be considered to be one of the following.

6. <u>Frequentist Confidence Distribution</u>: Obtained from the distribution of the pivotal quantity $\frac{x_{n+1}}{x_n}$ (Lawless and Fredette (2005)).

7. <u>Profile Predictive Likelihood</u>: Obtained by eliminating the parameter $\lambda$ from the joint likelihood of $x_{n+1}$ and $\lambda$ by maximization (Bjornstad (1990)).

8. <u>Objective Bayesian Predictive Posterior distribution</u>: Obtained using the non-informative Jeffrey's prior $\frac{1}{\lambda}$

9. <u>Conditional Normalized Maximum Likelihood (CNML)</u>: The CNML predictive distribution, from information theoretic considerations (Schmidt and Makalic (2009)).

10. <u>Accuracy of the Predictive Distribution</u>: The accuracy of a predictive distribution may be measured using the distance or divergence between the true exponential distribution with rate parameter $\lambda_0$ and the predictive distribution based on the sample $x$.

11. <u>Use of Kullback-Liebler Divergence</u>: The Kullback-Leibler divergence is a commonly used parameterization-free measure of the distance between two distributions.

12. <u>KL Divergence for ML/CNML</u>: Letting $\Delta(\lambda_0 \mathbin{||} p)$ denote the Kullback-Leibler divergence between an exponential distribution with rate parameter $\lambda_0$ and a predictive distribution $p$ it can be shown that

$$\mathbb{E}_{\lambda_0}[\Delta(\lambda_0 \mathbin{||} p_{ML})] = \psi(n) + \frac{1}{n-1} - \log n$$

$$\mathbb{E}_{\lambda_0}[\Delta(\lambda_0 \mathbin{||} p_{CNML})] = \psi(n) + \frac{1}{n} - \log n$$

where the expectation is taken with respect to the exponential distribution with the rate parameter

$$\lambda_0 \in (0, \infty)$$

and $\psi(\cdot)$ is the digamma function.

13. <u>Superiority of the CNML Approach</u>: It is clear that the CNML predictive distribution is strictly superior to the maximum likelihood plug-in distribution in terms of the average Kullback-Leibler divergence for all sample size

$$n > 0$$

## Random Variate Generation

1. <u>Generation using Inverse Transform Sampling</u>: A conceptually very simple method for generating exponential variates is based on inverse transform sampling. Given a random variate $U$ drawn from the uniform distribution on the unit interval $(0, 1)$ the variate

$$T = F^{-1}(U)$$

has an exponential distribution, where $F^{-1}$ is the quantile function defined by

$$F^{-1}(p) = -\frac{\ln(1 - p)}{\lambda}$$

2. <u>Simple Scheme for Random Exponential</u>: Moreover, if $U$ is uniform in $(0, 1)$ then so is $1 - U$. This means one can generate exponential variates as follows:

$$T = -\frac{\ln U}{\lambda}$$

3. <u>Alternate Exponential Variate Construction Setting</u>: Other methods for generating exponential variates are discussed by Devroye (1986) and Knuth (1998).
4. <u>Sorted Set of Random Exponentials</u>: Other methods for generating exponential variates are discussed by Devroye (1986) and Knuth (1998).
5. <u>Sorted Set of Random Exponentials</u>: A fast method for generating a set of ready-ordered exponential variates without using a sorting routine is also available (Devroye (1986)).

## References

- Bjornstad, J. F. (1990): Predictive Likelihood: A Review *Statistical Sciences* **5 (2)** 242-254

- Devroye, L. (1986): *Non-uniform Random Variate Generation* **Springer-Verlag** New York, NY

- Eckford, A. W., and P. J. Thomas (2016): Entropy of the Sum of Two Independent, Non-identically Distributed Exponential Random Variables **arXiV**

- Elfessi, A., D. M. Reineke (2001): A Bayesian Look at Classical Estimation: The Exponential Distribution *Journal of Statistical Education* **9 (1)**

- Guerriero, V. (2012): [Power Law Distribution: Method of Multi-scale Inferential Statistics](#)

- Ibe, O. C. (2014): *Fundamentals of Applied Probability and Random Processes 2nd Edition* **Academic Press** Cambridge, MA

- Johnson, R. A., and D. W. Wichern (2017): *Applied Multivariate Statistical Analysis* **Pearson Prentice Hall** Hoboken, NJ

- Knuth, D. E. (1998): *The Art of Computer Programming 3rd Edition* **Addison-Wesley** Boston, MA

- Lawless, J. F., and M. Fredette (2005): Frequentist Predictions Intervals and Predictive Distributions *Biometrika* **92 (3)** 529-542

- Norton, M., V. Khokhlov, and S. Uryasev (2019): Calculating CVaR and bPOE for Common Probability Distributions with Application to Portfolio Optimization and Density Estimation *Annals of Operations Research* **299 (1-2)** 1281-1315

- Park, S. Y., and A. K. Bera (2009): Maximum Entropy Autoregressive Conditional Heteroscedasticity Model *Journal of Econometrics* **150 (2)** 219-230

- Ritzema, H. P. (1994): [Drainage Principles and Applications](#)

- Ross, S. M. (2009): *Introduction to Probability and Statistics for Engineers and Scientists 4th Edition* **Associated Press** New York, NY

- Schmidt, D. F., and D. Makalic (2009): Universal Models for the Exponential Distribution *IEEE Transactions on Information Theory* **55 (7)** 3087-3090

- Wikipedia (2023): [Exponential Distribution](#)

# Hilbert Space

## Overview

1. <u>Origin of the Hilbert Space</u>: *Hilbert spaces* allow generalizing the methods of linear algebra and calculus for finite-dimensional Euclidean vector spaces to spaces that may be infinite-dimensional (Wikipedia (2022)).

2. <u>Dot Product as Completion Metric</u>: A Hilbert space is a vector space equipped with an inner product which defines a distance function for which it is a complete metric.

3. <u>Practical Use in Function Spaces</u>: Hilbert spaces is naturally and frequently in mathematics and physics, typically in function spaces.

4. <u>Applications in Math and Science</u>: Hilbert spaces are indispensable tools in the theory of partial differential equations, quantum mechanisms, Fourier analysis – which include applications to signal processing and heat transfer, and ergodic theory – which forms the mathematical underpinning of thermodynamics.

5. <u>Use Outside of Euclidean Spaces</u>: Apart from the classical Euclidean vector spaces, examples of Hilbert spaces include spaces of square-integrable functions, spaces of sequences, Sobolev spaces consisting of generalized functions, and Hardy spaces of holomorphic functions.

6. <u>Applicability of Euclidean Space Theorems</u>: Geometric intuition plays an important role in many aspects of Hilbert space formulation. Exact analogues of the Pythagorean theorem and the parallelogram law hold in a Hilbert space.

7. <u>Projection onto a Linear Subspace</u>: At a deeper level, perpendicular projection onto a linear subspace plays a significant role in optimization problems and other aspects of the theory.

8. <u>Full Specification of an Element</u>: An element of a Hilbert can be uniquely specified by its coordinates with respect to an orthonormal basis, in analogy with Cartesian coordinates in classical geometry.

9. <u>Finite vs Infinite Hilbert Spaces</u>: When this basis is countably infinite, it allows identifying the Hilbert space with the space of infinite sequences that are square-summable. In the older literature, the latter space is often referred to as *the* Hilbert space.

## Definition and Illustration – Motivating Example: Euclidean Vector Space

1. <u>3D Euclidean Vector Space</u>: One of the most familiar examples of a Hilbert space is the Euclidean vector space consisting of 3D vectors, denoted by $\mathbb{R}^3$, and equipped with the dot product.

2. <u>Dot Product in Cartesian Coordinates</u>: The dot product takes two vectors $x$ and $y$, and produces a real number $x \cdot y$. If $x$ and $y$ are represented in Cartesian coordinates, then the dot product is defined by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3$$

3. <u>Properties of Dot Product Symmetry</u>: The dot product satisfies the following properties. First, it is symmetric in $x$ and $y$:

$$x \cdot y = y \cdot x$$

4. <u>Properties of Dot Product - Linearity</u>: It is linear in its first argument:

$$(ax_1 + bx_2) \cdot y = a(x_1 \cdot y) + b(x_2 \cdot y)$$

for any scalars $a, b$ and vectors $x_1, x_2$, and $y$.

5. <u>Properties of Dot Product – PD</u>: It is positive definite: for all vectors $x$

$$x \cdot x \geq 0$$

with equality if and only if

$$x = 0$$

6. <u>Definition of Real Inner Product</u>: An operation on vectors that, like the dot product, satisfies these three properties is known as a real inner product. A vector space equipped with such an inner product in known as a real inner product space.

7. <u>Finite-Dimensional Inner Product Space</u>: Every finite dimensional inner product space is also a Hilbert space.

8. <u>Pre-Hilbert vs *the*-Hilbert Space</u>: However, some sources call finite-dimensional spaces with these properties pre-Hilbert spaces, reserving the term "Hilbert spaces" for infinite-dimensional spaces.

9. <u>Dot Product in Euclidean Spaces</u>: The basic feature of dot product that connects it with Euclidean geometry is that it is related to both the length – or norm – of a vector, denoted $\|x\|$, and to the angle $\theta$ between two vectors $x$ and $y$ by means of the formula

$$x \cdot y = \|x\| \cdot \|y\| \cos \theta$$

10. <u>Multivariate Calculus in Euclidean Space</u>: Multivariate calculus in Euclidean space relies on the ability to compute limits, and to have useful criteria for concluding that limits exist.

11. <u>Illustration of the Notion of Completeness</u>:
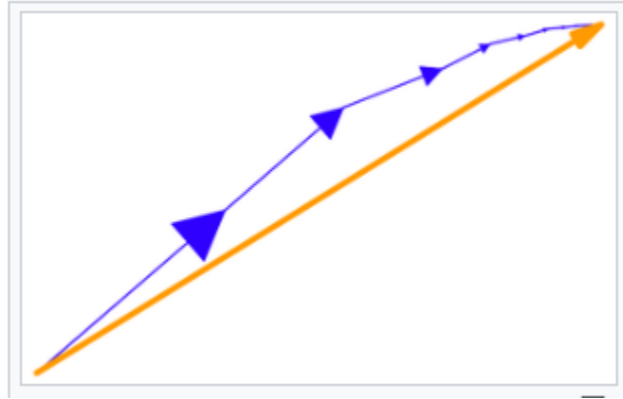
Completeness means that if a particle moves along a broken path – in blue – traveling a finite total distance, then the particle has a well-defined net displacement – shown in orange.

12. <u>Absolute Convergence of a Series</u>: A mathematical series $\sum_{n=0}^{\infty} x_n$ consisting of vectors in $\mathbb{R}^3$ is absolutely convergent provided that the sum of the lengths converges as an ordinary series of real numbers (Marsden (1974))

$$\sum_{k=0}^{\infty} \|x_k\| < \infty$$

13. <u>Convergence of Vectors to a Limit</u>: Just as with a series of scalars, a series of vectors that converges absolutely converges to some limit vector $L$ in the Euclidean space, in the sense that

$$\left\| L - \sum_{k=0}^{N} x_k \right\| \to 0$$

as

$$N \to \infty$$

14. _Completeness_ of the Euclidean Space: This property expresses the completeness of the Euclidean space: that a series that converges absolutely also converges in the ordinary sense.

15. Hilbert Spaces over Complex Numbers: Hilbert spaces are often taken over the complex numbers. The complex plane denoted by $\mathbb{Z}$ is equipped with a notion of magnitude, the complex modulus $|z|$, which is defined as the square root of the product of $z$ wit its complex conjugate:

$$|z|^2 = z\bar{z}$$

16. Modulus of a Complex Number: If $z = x + iy$ is a decomposition of $z$ into its real and imaginary parts, then the modulus is the usual Euclidean 2D length:

$$|z| = \sqrt{x^2 + y^2}$$

17. Inner Product of Complex Numbers: The inner product of a pair of complex numbers $z$ and $w$ is the product of $z$ with the complex conjugate of $w$:

$$\langle z, w \rangle = z\bar{w}$$

18. Real-part of Dot Product: This is complex valued. The real part of $\langle z, w \rangle$ gives the usual 2D Euclidean dot product.

19. Ordered Pair of Complex Numbers: A second example is the space $\mathbb{Z}^2$ whose elements are the pairs of complex numbers

$$z = (z_1, z_2)$$

Then the inner product of $z$ with another such vector

$$w = (w_1, w_2)$$

is given by

$$\langle z, w \rangle = z_1 \overline{w_1} + z_2 \overline{w_2}$$

20. <u>Inner Product is Hermitian Symmetric</u>: The real part of $\langle z, w \rangle$ is then the 2D Euclidean dot product. The inner product is *Hermitian* symmetric, which means that the result of changing $z$ and $w$ is the complex conjugate

$$\langle w, z \rangle = \overline{\langle z, w \rangle}$$

## Definition and Illustration – Definition

1. <u>What is a Hilbert Space</u>? A *Hilbert space H* is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product (Dieudonne (1960), Hewitt and Stromberg (1965), Reed and Simon (1980), Rudin (1987)).

2. <u>Complex Inner Product Space</u>: To say that $H$ is a *complex inner product space* means that $H$ is a complex vector space on which there is an inner product $\langle x, y \rangle$ associating a complex number with each pair of elements $x$, $y$ of $H$ that satisfies the following properties.

3. <u>Inner Product is Conjugate Symmetric</u>: This means that the inner product of a pair of elements id equal to the complex conjugate of the inner product of swapped elements:

$$\langle y, x \rangle = \overline{\langle x, y \rangle}$$

Importantly, this implies that $\langle x, x \rangle$ is a real number.

4. <u>Inner Product Linear in First Argument</u>: For all complex numbers $a$ and $b$

$$\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2 + y \rangle$$

In some conventions, the inner products are linear in their second arguments instead.

5. Positive Definite: The inner product of an element with itself is positive definite:

$$\langle x, x \rangle > 0$$

if

$$x \neq 0$$

$$\langle x, x \rangle = 0$$

if

$$x = 0$$

6. Conjugate Linear in Second Element: It follows from the first and the second properties that a complex inner product is *anti-linear*, also called *conjugate linear*, in its second argument, meaning that

$$\langle x, ay_1 + by_2 \rangle = \bar{a}\langle x, y_1 \rangle + \bar{b}\langle x, y_2 \rangle$$

7. Real Inner Product Space: A *real inner product space* is defined in the same way, except that $H$ is a real-vector space and the inner product takes real values.

8. Dual System with Bilinear Map: Such an inner product will be a bilinear map and $(H, H, \langle ., . \rangle)$ will be a dual system (Schaffer and Wolff (1999)).

9. Norm and Distance in $H$: The norm is the real-values function

$$\|x\| = \sqrt{\langle x, x \rangle}$$

and the distance between two points $x, y$ in $H$ is defined in terms of the norm by

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

**10.** <u>Properties of the Distance Function</u>: That this function is a distance function means firstly that it is symmetric in $x$ and $y$, secondly that the distance between $x$ and itself is zero, and otherwise the distance between $x$ and $y$ must be positive, and lastly that triangle inequality holds, meaning that the length of the leg of one leg of a triangle $xyz$ cannot exceed the sum of the lengths of the other two lengths:

$$d(x, x) \leq d(x, y) + d(y, z)$$

11. <u>Consequence of Cauchy-Schwarz Inequality</u>:
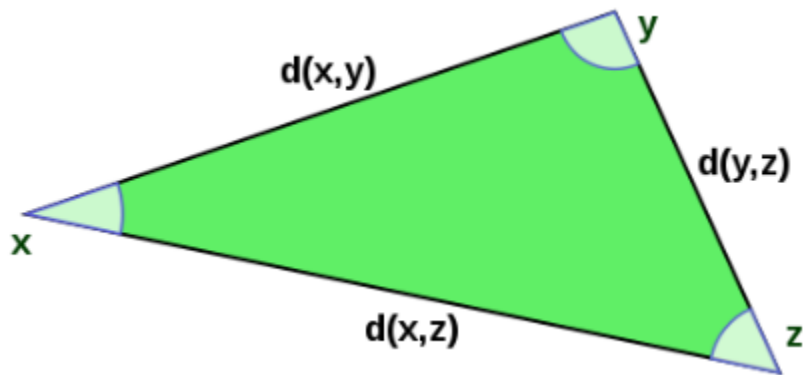


This last property is ultimately a consequence of the more fundamental Cauchy-Schwarz inequality, which asserts

$$|x, y| \leq \|x\|\|y\|$$

with inequality if and only if $x$ and $y$ are linearly dependent.

12. <u>Hausdorff pre-Hilbert Space</u>: With a distance function defined this way, any inner product space is a metric space, and is sometimes known as a *Hausdorff pre-Hilbert space* (Dieudonne (1960)).

13. <u>Pre-Hilbert Space that is Complete</u>: Any pre-Hilbert space that is additionally also a complete space is a Hilbert space.

14. <u>Completeness from the Cauchy Criterion</u>: The *completeness* of $H$ is expressed using a form of Cauchy criterion for sequences in $H$: a pre-Hilbert space $H$ is complete if every Cauchy sequence converges with respect to this norm to an element in the space.

15. <u>Completeness Criterion using Norm Convergence</u>: Completeness can be characterized by the following equivalent condition: if a series of vectors

$$\sum_{k=0}^{\infty} u_k$$

converges absolutely in the sense that

$$\sum_{k=0}^{\infty} \|u_k\| < \infty$$

then the series converges in $H$, in the sense that partial sums converge to an element of $H$.

16. <u>Equivalence between Hilbert and Banach</u>: As a complete normed space, Hilbert spaces are by definition also Banach spaces.

17. <u>As a Topological Vector Banach Spaces</u>: As such, they are topological vector spaces, in which topological notions such as openness and closedness of subsets are well-defined.

18. <u>Closed Linear Subspace of Hilbert Space</u>: Of special importance is the notion of a closed linear subspace of a Hilbert space that, with the inner product induced by

restriction, is also complete, i.e., being a closed set in a complete metric space, and therefore a Hilbert space in its own right.

## Definition and Illustration – Second Example: Sequence Spaces

1. $l_2$ Spaces of Infinite Sequences: The sequence spaces $l_2$ consists of all infinite sequences

$$z = (z_1, z_2, \cdots)$$

of complex numbers such that the series

$$\sum_{n=1}^{\infty} |z_n|^2$$

converges.

2. Inner Product on the Space: The inner product on $l_2$ is defined by

$$\langle z, w \rangle = \sum_{n=1}^{\infty} z_n \overline{w_n}$$

with the latter series converging as a consequence of the Cauchy-Schwarz inequality and the convergence of the previous series.

3. Convergence to within the Space: Completeness of the space holds provided that whenever a series of elements from $l_2$ converges absolutely in norm, it then converges to an element in $l_2$.

4. Consequence of this Completeness: The proof is basic in mathematical analysis, and permits mathematical series of elements of the space to be manipulated with the same

ease as a series of complex numbers, or vectors in finite-dimensional Euclidean space (Dieudonne (1960)).

## Lebesgue Spaces

1. <u>Definition of the Lebesgue Spaces</u>: Lebesgue spaces are function spaces associated with the measure spaces $(X, M, \mu)$, where $X$ is a set, $M$ is a $\sigma$-algebra of subsets of $X$, and $\mu$ is a countably additive measure on $M$.

2. <u>Finiteness of the Lebesgue Integral</u>: Let $L_2(X, \mu)$ be the space of those complex-valued measurable functions on $X$ for which the Lebesgue integral of the square of the absolute value is finite, i.e., for a function $f$ in $L_2(X, \mu)$

$$\int_X |f|^2 d\mu < \infty$$

and where the functions are identified if and only if they differ only on a set of measure zero.

3. <u>Inner Product in Lebesgue Space</u>: The inner product of functions $f$ and $g$ in $L_2(X, \mu)$ is then defined as

$$\langle f, g \rangle = \int_X f(t)\overline{g(t)} d\mu(t)$$

or

$$\langle f, g \rangle = \int_X \overline{f(t)} g(t) d\mu(t)$$

249

where the second form – the conjugation of the first element – is commonly found in theoretical physics literature.

4. <u>Existence of Inner Product Integral</u>: For $f$ and $g$ in $L_2$, the integral exists because of the Cauchy-Schwarz inequality, and defines an inner product on the space. Equipped with this inner product, $L_2$ is in fact complete (Halmos (1957)).

5. <u>Need for the Lebesgue Integral</u>: The Lebesgue integral is essential to ensuring completeness: on domains of real numbers, for instance, not enough functions are Riemann integrable (Hewitt and Stromberg (1965)).

6. <u>Lebesgue Spaces in Natural Settings</u>: The spaces $L_2(\mathbb{R})$ and $L_2([0,1])$ of square-integrable functions with respect to the Lebesgue measure are the real line and the unit intervals, respectively, and are natural domains on which to define the Fourier transform and the Fourier series.

7. <u>Weighted Square-integrable Lebesgue Spaces</u>: In other situations, the measure may be something other than the ordinary Lebesgue measure on the real line. For instance, if $w$ is any positive measurable function, the space of all measure functions $f$ on the interval $[0,1]$ satisfying

$$\int_0^1 |f(t)|^2 w(t)dt < \infty$$

is called the weighted $L_2$ space $L_{2,w}([0,1])$ and $w$ is called the weight function.

8. <u>Weighted Lebesgue Space Inner Product</u>: The inner product is defined by

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g(t)}w(t)dt$$

9. <u>Equivalence of the Weighting Measure</u>: The weighted space $L_{2,w}([0,1])$ is identical to the Hilbert space $L_2([0,1], \mu)$ where the measure $\mu$ of a Lebesgue-measurable set $A$ is defined by

$$\mu(A) = \int_A w(t)\,dt$$

10. <u>Weighted $L_2$ in Orthogonal Polynomials</u>: Weighted $L_2$ spaces like this are frequently used to study orthogonal polynomials, because families of different polynomials are orthogonal to different weighting functions.

## Examples – Sobolev Spaces

1. <u>Representing the Sobolev Hilbert Spaces</u>: Sobolev spaces, denoted $H_S$ or $W_{S,2}$, are Hilbert spaces.

2. <u>Differentiable along with Inner Product</u>: These are a special kind of function space in which differentiation my be performed, but that – unlike other Banach spaces such as Holder spaces – support the structure of an inner product.

3. <u>Use in Solving of PDE's</u>: Because differentiation is permitted, Sobolev spaces are a convenient setting for the theory of partial differential equations (Bers, John, and Schechter (1982)).

4. <u>Direct Methods in Calculus of Variations</u>: They also form the basis for the theory of direct methods in the calculus of variations (Giusti (2003)).

5. <u>$L_2$ Nature of the Sobolev Derivative</u>: For a non-negative integer $s$ and

$$\Omega \in \mathbb{R}^n$$

The Sobolev space $H_S(\Omega)$ consists of $L_2$ functions whose weak derivatives of order up to $s$ are also $L_2$.

6. <u>Inner Product of Sobolev Spaces</u>: The inner product in $H_S(\Omega)$ is

$$\langle f,g \rangle = \int_{\Omega} f(x) \cdot \overline{g(x)} dx + \int_{\Omega} \frac{\partial f(x)}{\partial x} \cdot \frac{\partial \overline{g(x)}}{\partial x} dx + \cdots + + \int_{\Omega} \frac{\partial^s f(x)}{\partial x^s} \cdot \frac{\partial^s \overline{g(x)}}{\partial x^s} dx$$

where the dot indicates the dot product in the Euclidean space of derivatives of each order.

7. *s* Need not be an Integer: Sobolev spaces can also be defined when $s$ is not an integer.

8. Spectral Theory View of Sobolev Spaces: Sobolev spaces are also studied from the viewpoint of spectral theory, relying more specifically on the Hilbert space structure.

9. Representation Using Bessel Potentials: If $\Omega$ is a suitable domain, then one can define the Sobolev spaces $H_S(\Omega)$ as the space of Bessel potentials (Stein (1970)); roughly

$$H_S(\Omega) = \left\{ (1-\Delta)^{-\frac{s}{2}} f \mid f \in L_2(\Omega) \right\}$$

10. Use of Spectral Mapping Theorem: Here $\Delta$ is the Laplacian, and $(1-\Delta)^{-\frac{s}{2}}$ is understood in terms of the spectral mapping theorem.

11. Study of Pseudo-differentiable Operators: Apart from providing a workable definition of Sobolev spaces for non-integer $s$, this definition also has particularly desirable properties under the Fourier transform that make it ideal for the study of pseudo-differential operators.

12. Use in the Hodge Decomposition: Using these methods on a compact Riemannian manifold, one can obtain for instance the Hodge decomposition, which is basis of Hodge theory (Warner (1983)).

## Examples – Hardy Spaces of Holomorphic Functions

1. <u>Holomorphic Elements of Hardy Spaces</u>: The Hardy spaces are functions spaces, arising in complex analysis and harmonic analysis, whose elements are certain holomorphic functions in a complex domain (Duren (1970)).

2. <u>Definition of the Hardy Spaces</u>: Let $U$ be the unit disc in the complex plane. Then the Hardy space $H_2(U)$ is defined as the space of holomorphic functions $f$ on $U$ such that the means

$$M_r(f) = \frac{1}{2\pi} \int\limits_0^{2\pi} |f(re^{i\theta})|^2 d\theta$$

remain bounded for

$$r < 1$$

3. <u>Norm on this Hardy Space</u>: This is defined by

$$\|f\|_2 = \lim_{r \to 1} \sqrt{M_r(f)}$$

4. <u>Comparison with the Fourier Series</u>: Hardy spaces in the disc are related to the Fourier series. A function is in $H_2(U)$ if and only if

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

where

$$\sum_{n=0}^{\infty} |a_n|^2 < \infty$$

253

5. L$_2$ Functions on the Circle: Thus $H_2(U)$ consists of those functions that are $L_2$ on the circle, and whose negative frequency Fourier coefficients vanish.

## Bergman Spaces

1. <u>What are the Bergman Functions</u>? The Bergman spaces are another family of Hilbert spaces of holomorphic functions (Krantz (2002)).
2. <u>Square Integrability over Lebesgue Measure</u>: Let $D$ be a bounded open set in the complex plane – or a higher-dimensional complex space – and let $L_{2,n}(D)$ be the space of holomorphic functions $f$ in $D$ that are also in $L_2(D)$ in the sense that

$$\|f\|^2 = \int_D |f(z)|^2 d\mu(z) < \infty$$

   where the integral is taken with respect to the Lebesgue measure in $D$.
3. <u>$L_{2,n}(D)$ as a Self-contained Hilbert Subspace</u>: Clearly, $L_{2,n}(D)$ is a subspace of $L_2(D)$; in fact, it is a closed subspace, and so a Hilbert space in its own right. This is a consequence of the estimate, valid on compact subsets $K$ of $D$, that

$$\sup_{z \in K} |f(z)| \leq C_K \|f\|_2$$

   which in turn follows from Cauchy's integral formula.
4. <u>Holomorphic Nature of Cauchy Limit</u>: Thus, convergence of a sequence of holomorphic functions in $L_2(D)$ implies also compact convergence, and so the limit function is also holomorphic.
5. <u>Linear Functional Continuous on $L_{2,n}(D)$</u>: Another consequence of this inequality is that the linear functional that evaluates a function $f$ at a point $D$ is actually continuous on $L_{2,n}(D)$.

6. <u>Using the Reisz Representation Theorem</u>: The Reisz representation theorem implies that the evaluation functional can be represented as an element of $L_{2,n}(D)$.

7. <u>Implied Reproducing Bergman Kernel Function</u>: Thus, for every $z \in D$ there is a function

$$\eta_z \in L_{2,n}(D)$$

such that

$$f(z) = \int_D f(\zeta)\overline{\eta_z(\zeta)}d\mu(\zeta)$$

for all

$$f \in L_{2,n}(D)$$

8. <u>Bergman Kernel of $D$</u>: The integrand

$$K(\zeta,z) = \overline{\eta_z(\zeta)}$$

is called the Bergman kernel of $D$.

9. <u>Reproducing Property satisfied by the Kernel</u>:

$$f(z) = \int_D f(\zeta)K(\zeta,z)d\mu(\zeta)$$

10. <u>Bergman Space - Example of RKHS</u>: The Bergman space is an example of a reproducing kernel Hilbert space, which is a Hilbert space of functions along with a kernel $K(\zeta,z)$ that verifies a reproducing property analogous to the one above.

11. <u>Szego Kernel - Hardy Space Reproducing Kernel</u>: The Hardy space $H_2(D)$ also admits a reproducing kernel, known as the Szego kernel (Krantz (2002)).

12. <u>Reproducing Kernels in other Areas</u>: Reproducing kernels are common in other areas of mathematics as well. For instance, in harmonic analysis, the Poisson kernel is a reproducing kernel for the Hilbert space of square-integrable harmonic functions in the unit ball. That the latter is a Hilbert space at all is a consequence of the mean value theorem for harmonic functions.

## Applications

1. <u>Value behind Hilbert Space Formulation</u>: Many of the applications of Hilbert spaces exploit the fact that Hilbert spaces support generalizations of simple geometric concepts like projection and change of basis from their usual finite dimension setting.

2. <u>Spectral Theory of Linear Operators</u>: In particular, the spectral theory of continuous self-adjoint linear operators on a Hilbert space generalizes the usual spectral decomposition of a matrix, and this plays a major role in the application of the theory to other areas of mathematics and physics.

## Application – Sturm-Liouville Theorem

1. <u>Use in Ordinary Differential Equations</u>: In the theory of ordinary differential equations, spectral methods on a suitable Hilbert space are used to study the behavior of eigenvalues and eigenfunctions of differential equations.

2. <u>Example - The Sturm-Liouville Theorem</u>: For example, the Sturm-Liouville problem arises in the study of harmonics of waves in a violin string or a drum, and is a central problem in ordinary differential equations (Young (1988)).

3. <u>The ODE and Boundary Conditions</u>: The problem is a differential equation of the form

$$-\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + q(x)y = \lambda w(x)y$$

for an unknown function $y$ on an interval $[a, b]$ satisfying the general homogenous Robin boundary conditions

$$\alpha y(a) + \alpha' y'(a) = 0$$

and

$$\beta y(b) + \beta' y'(b) = 0$$

4. <u>Existence of Solutions to the ODE</u>: The functions $p$, $q$, and $w$ are given in advance, and the problem is to find the function $y$ and the constants $\lambda$ for which the equation has a solution.

5. <u>Eigenvalues of Sturm Liouville Equation</u>: The problem only has solutions for certain values of $\lambda$, called the eigenvalues of the system, and this is a consequence of the spectral theorem for compact operators applied to the integral operator defined by the Green's function for the system.

6. <u>Kernel Associated with the System</u>: Furthermore, another consequence of this general result is that the eigenvalues $\lambda$ of the system can be arranged in an increasing sequence tending to infinity. The eigenvalues of the Fredholm kernel are $\frac{1}{\lambda}$, which tend to zero.

7. <u>Illustration of the Oscillating Modes</u>:
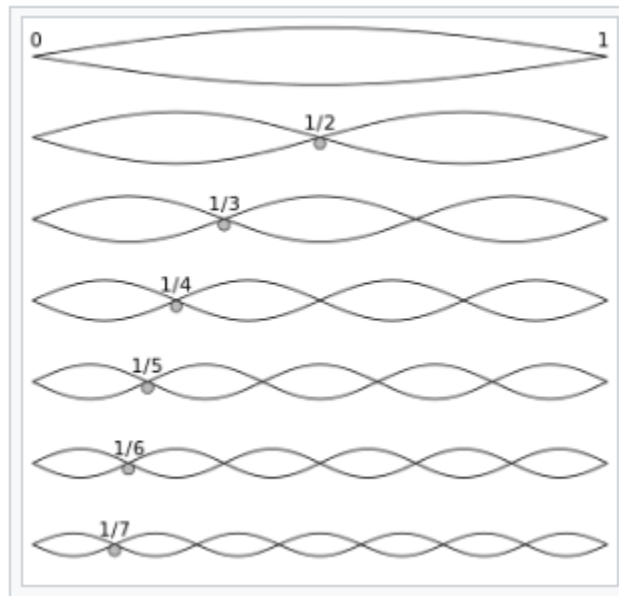
Picture illustrates the overtones of a vibrating string. These are eigenfunctions of an associated Sturm-Liouville problem. The eigenvalues $1, \frac{1}{2}, \frac{1}{3}, \cdots$ form the musical harmonic series.

## Partial Differential Equations

1. <u>Use in Study of PDE</u>: Hilbert spaces form a basic tool in the study of partial differential equations (Bers, John, and Schechter (1981)).
2. <u>Generalized Weak Solutions of PDE's</u>: For many classes of partial differential equations, such as linear elliptic equations, it is possible to consider a generalized solution – known as a weak solution – by enlarging the class of functions.
3. <u>Weak Formulations using Sobolev Functions</u>: Many weak formulations involve the class of Sobolev functions, which is a Hilbert space.
4. <u>Benefits of the Weak Formulation</u>: A suitable weak formulation reduces to a geometrical problem the analytic problem of finding a solution, or often what is more important, showing that a solution exists and is unique for given boundary data.

5. <u>Applying the Lax-Milgram Theorem</u>: For linear elliptic equations, one geometrical result that ensures the unique solvability for a large class of problems is the Lax-Milgram theorem.

6. <u>Basis of the Galerkin Method</u>: This strategy forms the rudiment of the Galerkin method – a finite element method – for numerical solution of partial differential equations. More details on the finite element methods from this point of view can be found in Brenner and Scott (2005).

7. <u>Poisson Equations with Dirichlet Boundary Conditions</u>: A typical example is the Poisson equation

$$\nabla u = g$$

with Dirichlet boundary conditions in a bounded domain $\Omega$ in $\mathbb{R}^2$.

8. <u>Application of the Weak Formulation</u>: The weak formulation consists of funding a function $u$ such that, for all continuously differentiable functions $v$ in $\Omega$ vanishing in the boundary:

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega gv$$

9. <u>Recast Using the Hilbert Space</u>: This can be re-cast in terms of the Hilbert space $H_0^1(\Omega)$ consisting of functions $u$ such that $u$, along with its weak partial derivatives, are square-integrable on $\Omega$, and vanish on the boundary.

10. <u>Recast Using Linear/Bilinear Functionals</u>: The question then reduces to finding $u$ in this space such that for all $v$ in this space

$$a(u, v) = b(v)$$

where $a$ is a continuous bilinear form, and $b$ is a continuous linear functional, given by

$$a(u, v) = \int\limits_{\Omega} \nabla u \cdot \nabla v$$

$$b(v) = \int\limits_{\Omega} gv$$

11. <u>Coercive Nature of Bilinear Forms</u>: Since the Poisson equation is elliptic, it follows from Poincare's inequality that the bilinear form $a$ is coercive.

12. <u>Existence and Uniqueness of Solutions</u>: The Lax-Milgram theorem then ensures the existence and the uniqueness of solutions of these equations.

13. <u>Hilbert Spaces and Lax-Milgram</u>: Hilbert spaces allow for many elliptic partial differential equations to be formulated in a similar way, and the Lax-Milgram theorem is then a basic tool in their analysis.

14. <u>Use in Parabolic/Hyperbolic PDE's</u>: With suitable modifications, similar techniques can be applied to parabolic partial differential equations and certain hyperbolic partial differential equations.

## Ergodic Theory

1. <u>Chaotic Systems Long Term Dynamics</u>: The field of ergodic theory is the study of the long-term behavior of chaotic dynamical systems.

2. <u>Thermodynamics Application of Chaos Theory</u>: The prototypical case of a field that ergodic theory applies to is thermodynamics, in which, though the microscopic state of a system is extremely complicated – it is impossible to understand the ensemble of individual collisions between particles of matter – the average behavior over sufficiently long time-intervals is tractable.

3. Laws of Thermodynamics: The average behavior. laws of thermodynamics are assertions about such behavior.

4. The Zeroth Law of Thermodynamics: In particular, one formulation of the zeroth law of thermodynamics asserts that over sufficiently long timescales, the only functionally independent measurement that one can make of a thermodynamic system in equilibrium is its total energy, in the form of temperature.

5. Bunimovich Stadium Billiard Ball Trajectory:
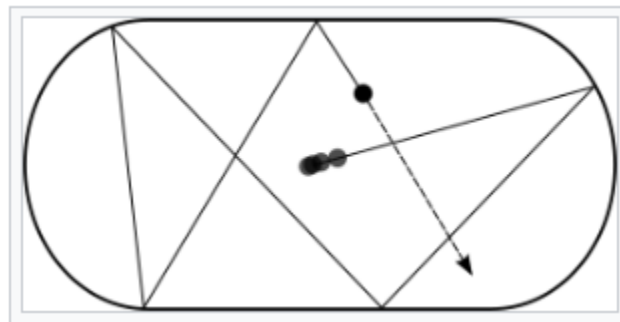


The path of a billiard ball in the Bunimovich stadium is described by an ergodic dynamical system.

6. Conserved Quantities in Phase Space: An ergodic dynamical system is one for which, apart from the energy – measured by the Hamiltonian – there are no other functionally independent conserved quantities in the phase space.

7. Energy Space and its Evolution: More explicitly, suppose that the energy $E$ is fixed, and let $\Omega_E$ be the subset of the phase space consisting of all the states of energy $E$ – an energy surface – and let $T_t$ denote the evolution operator on the phase space.

8. Ergodic Nature of Dynamical Systems: The dynamical system is ergodic if there are no continuous non-constant functions on $\Omega_E$ such that

$$f(T_t w) = f(w)$$

for all $w$ in $\Omega_E$ and all time $t$.

9. <u>Implication of Liouville's Theorem</u>: Liouville's theorem implies then that there exists a measure $\mu$ on the energy surface that is invariant under the time translation.

10. <u>Time Evolution Operator</u>: As a result, the time translation is a unitary transformation of the Hilbert space $L_2(\Omega_E, \mu)$ consisting of square-integrable functions on the energy surface $\Omega_E$ with respect to the inner product

$$\langle f, g \rangle_{L_2(\Omega_E, \mu)} = \int_E f \bar{g} d\mu$$

11. <u>von Neumann Mean Ergodic Theorem</u>: von Neumann (1932) states the following.

12. <u>Orthogonal Projection of Unitary Operator</u>: If $U_t$ is a strongly continuous one-parameter semi-group of unitary operators on a Hilbert space $H$, and $P$ is the orthogonal projection onto the space of common fixed points of $U_t$

$$\{x \in H \mid U_t x = x, \forall t > 0\}$$

then

$$Px = \lim_{T \to \infty} \frac{1}{T} \int_0^T U_t x \, dt$$

13. <u>Consequence of the Ergodic Theorem</u>: For an ergodic system, the fixed set of time evolution consists only of the constant functions, so the ergodic theorem implies the following (Reed and Simon (1980)): for any function

$$f \in L_2(\Omega_E, \mu)$$

$$L_2 = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(T_t w) d\mu = \int_{\Omega_E} f(y) d\mu(y)$$

14. <u>Long Time Average of an Observable</u>: That is, the long-term average of an observable $f$ is equal to its expectation value over an energy surface.

## Fourier Analysis

1. <u>Goal of Fourier Analysis</u>: One of the basic goals of Fourier analysis is to decompose into a – possibly infinite – linear combination of given basis functions: the associates Fourier series.

2. <u>Decomposition into a Fourier Series</u>: The classical Fourier series associated with a function $f$ defined on the internet $[0, 1]$ in a series of the form

$$\sum_{n=-\infty}^{+\infty} a_n e^{2\pi i n \theta}$$

where

$$a_n = \int_0^1 f(\theta) e^{-2\pi i n \theta} \, d\theta$$

3. <u>Fourier Decomposition of Sawtooth Wave</u>:
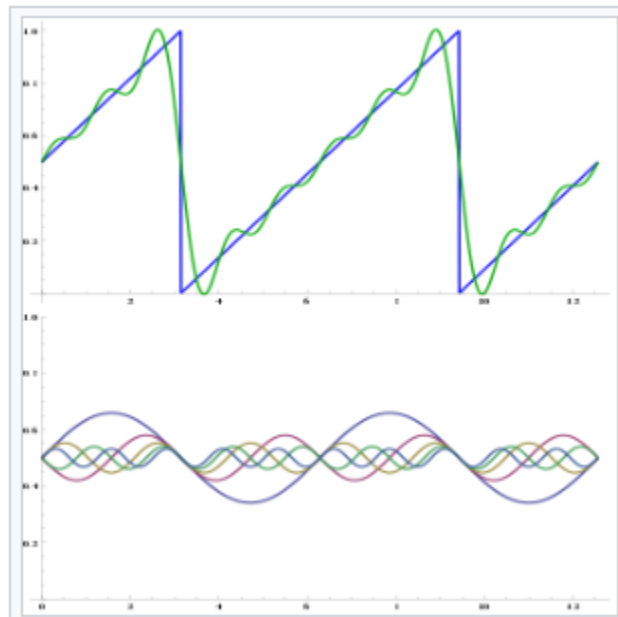
Superposition of sinusoidal wave basis functions – bottom – to form a sawtooth wave – top.

4. <u>Fourier Terms of a Sawtooth Function</u>: The example of adding to the first terms of a Fourier series for a sawtooth function is shown in the picture above.

5. <u>Wavelengths of the Basis Functions</u>: The basis functions are since waves with wavelength $\frac{\lambda}{n}$ – for integer $n$ – shorter than the wavelength $\lambda$ of the sawtooth itself – except for

$$n = 1$$

the *fundamental* wave.

6. <u>Nodes of the Basis Functions</u>: All the basis functions have nodes at the nodes of the sawtooth, but all but the fundamental have additional nodes.

7. <u>Departure from the Original Sawtooth</u>: The oscillation of the summed terms about the sawtooth is called the Gibbs phenomenon.

8. <u>Convergence of the Fourier Series</u>: A significant problem is classical Fourier series asks in what sense the Fourier series converges, if at all, to the function.

9. <u>Hilbert Space Based Convergence Estimation</u>: Hilbert space methods provide one possible answer to this question (Rudin (1987), Folland (2009)).

10. <u>Hilbert Space Orthogonal Basis Functions</u>: The functions

$$e_n(\theta) = e^{2\pi in\theta}$$

from an orthogonal basis of the Hilbert space $L_2([0,1])$

11. <u>Square-integrable Functions Series Expression</u>: Consequently, any square-integrable function can be expressed as a series

$$f(\theta) = \sum_n a_n e_n(\theta)$$

$$a_n = \langle f, e_n \rangle$$

and, moreover, this series converges in the Hilbert space sense, i.e., in the $L_2$ mean sense.

12. <u>Hilbert Space based Orthonormal Abstraction</u>: The problem can also be studied from the abstract point of view; every Hilbert space has an orthonormal basis, and every element of the Hilbert space can be written in a unique way as a sum of multiplies of these basis elements.

13. <u>Coefficients of the Orthonormal Basis</u>: The coefficients appearing on these basis elements are sometimes known abstractly as the Fourier coefficients of the elements of the space (Halmos (1957)).

14. <u>Choice of Different Orthonormal Basis</u>: The abstraction is especially useful when it is more natural to use the basis functions for a space such as $L_2([0,1])$

15. <u>Examples of Orthonormal Basis Functions</u>: In many circumstances, it is desirable not to decompose a function into trigonometric functions, but rather into orthogonal polynomials or wavelets for instance (Bachman, Narici, and Beckenstein (2000)), and in higher dimension into spherical harmonics (Stein and Weiss (1971)).

16. <u>Illustration of Spherical Harmonic Basis</u>:
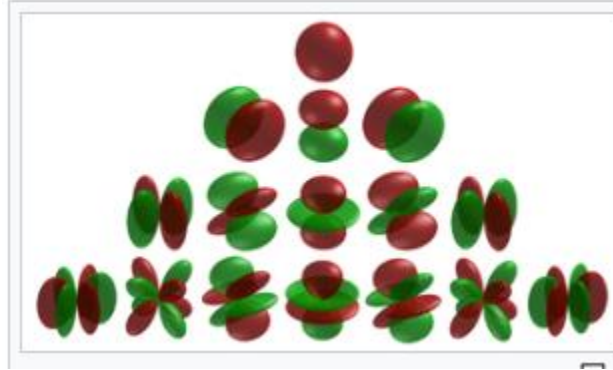


Spherical harmonics, an orthonormal basis for the Hilbert space of square-integrable functions on the sphere, shown graphed along the radial direction.

17. <u>Explicit Form of Series Sum</u>: For instance, if $e_n$ are any orthonormal basis functions of $L_2[0,1]$, then a given function in $L_2[0,1]$ can be approximated as a finite linear combination (Lanczos (1988))

$$f(x) \approx f_n(x) = a_1 e_1(x) + \cdots + a_n e_n(x)$$

18. <u>Choice of the Loading Coefficients</u>: The coefficients $\{a_j\}$ are selected to make the magnitude of the difference $\|f - f_n\|^2$ as small as possible.

19. <u>Loading Coefficients as Dot Products</u>: Geometrically, the best approximation is the orthogonal projection of $f$ onto the subspace consisting of all linear combinations of the $\{e_j\}$, and can be calculated by Lanczos (1988):

$$a_j = \int_0^1 \overline{e_j(x)} f(x) dx$$

20. <u>Minimization of the Approximation Error</u>: That this formula minimizes the difference $\|f - f_n\|^2$ is a consequence of Bessel's inequality and Parseval's formula.

21. <u>Decomposition Using Differential Operator Eigenfunctions</u>: In various applications to physical problems, a function can be decomposed into physically meaningful eigenfunctions of a differential operator, typically the Laplace operator: this forms the foundation for the spectral study of functions, in reference to the spectrum of the differential operator (Courant and Hilbert (1953), Reed and Simon (1975)).

22. <u>Hearing the Shapes of Drums</u>: A concrete physical application involves the problem of hearing the shape of a drum: given the fundamental modes of vibration that a drumhead is capable of producing, can one infer the shape of drum itself (Kac (1966))?

23. <u>Dirichlet Eigenvalues of Laplace Equation</u>: The mathematical formulation of this equation involves the Dirichlet eigenvalues of Laplace equation in the plane, that represent the fundamental modes of vibration in direct analogy with the integers that represent the fundamental modes of vibration of the violin string.

24. <u>Spectral Theory in Fourier Transform</u>: Spectral theory also underlies certain aspects of the Fourier transform of a function.

25. <u>Fourier Analysis vs. Fourier Transform</u>: Whereas Fourier analysis decomposes a function defined on a compact set into the discrete spectrum of the Laplacian – which corresponds to the vibrations of a violin string or drum – the Fourier transform of a function is the decomposition of a function on all of Euclidean space into its components in the continuous spectrum of the Laplacian.

26. <u>Isometry Inherent in Fourier Transform</u>: The Fourier transformation is also geometrical, in a sense made precise by the Plancherel theorem, that asserts that it is an isometry of one Hilbert spaces – the "time domain" – with another – the "frequency domain".

27. <u>Impact of the Fourier Isometry</u>: The isometry property of the Fourier transformation is a recurring theme in abstract harmonic analysis, since it reflects the conservation of energy for the continuous Fourier Transform, as evidenced by the Plancherel theorem for spherical functions occurring in non-commutative harmonic analysis.

**Applications – Quantum Mechanics**

1. <u>Rigorous Formulation of Quantum Mechanics</u>: In the mathematically rigorous formulation of quantum mechanics, developed by von Neumann (1955), the possible states – more precisely, the pure states – of a quantum mechanical system are represented by unit vectors – called *state vectors* – residing in a complex separable Hilbert space known as the state space, well defined up to a complex number of norm 1, the phase factor.

2. <u>States as Hilbert Space Projectivizations</u>: In other words, the possibly states are points in the projectivizations of a Hilbert space, usually called a complex projective space.

3. <u>Nature of the State Spaces</u>: The exact nature of the Hilbert space is dependent on the system; for example, the position and the momentum states for a single non-relativistic spin zero particle is the space of all square-integrable functions; for the states for the spin of a single proton are unit elements of a two-dimensional complex Hilbert space of spinors.

4. <u>Observables as Self-adjoint Operators</u>: Each observable is represented by a self-adjoint operator acting on the state space.

5. <u>Eigenstates and Eigenvalues of Observables</u>: Each eigenstate of an observable corresponds to the eigenvector of an operator, and the associated eigenvalue corresponds to the value of the observable in that eigenstate.

6. <u>Electron Orbitals in Hydrogen Atom</u>:
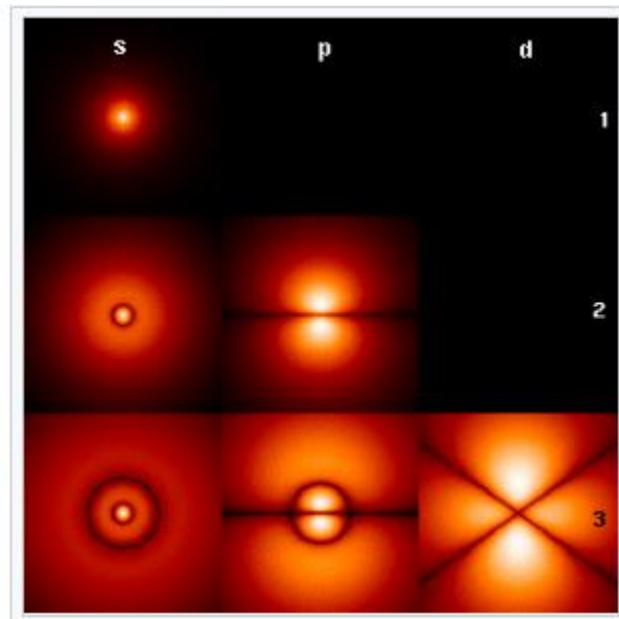
The orbitals of an electron in a hydrogen atom are eigenfunctions of the energy.

7. <u>Inner Product between State Vectors</u>: The inner product between two state vectors is a complex number known as a probability amplitude.

8. <u>Target State Collapse Probability</u>: During an ideal measurement of a quantum mechanical system, the probability that a system collapses from an initial state to a particular eigenstate is given by the square of the absolute value of the probability amplitudes between the initial and the final states (Rieffel and Polak (2011)).

9. <u>Eigenvalue of the Operator</u>: The possible results of a measurement are the eigenvalues of the operator – which explains the choice of self-adjoint operators, for all the eigenvalues must be real.

10. <u>Probability of Observing a State</u>: The probability distribution of an observable being in a given state can be found by computing the spectral decomposition of the corresponding operator (Peres (1993)).

11. <u>Statistical Mixtures of Pure States</u>: For a general system, states are not typically pure, but instead are represented as statistical mixtures of pure states, or mixed states, given by density matrices; self-adjoint operators of trace one on a Hilbert space (Peres (1993)).

12. <u>Single Measurement System-wide Impact</u>: Moreover, for general quantum mechanical systems, the effects of a single measurement can influence other parts of a system in a manner that is described by a positive operator valued measure.

13. <u>States/Observables in General Systems</u>: Thus, the structure of both the states and the observables in the general theory is considerably more complicated than the idealization for pure states (Nielsen and Chuang (2000)).

## Applications – Color Perception

1. <u>Physical Colors from Spectral Combinations</u>: Any true physical color can be represented by a combination of pure spectral colors.

2. <u>Hilbert Space for Physical Colors</u>: As physical colors can be composed of any number of spectral colors, the space of physical colors may aptly be represented by a Hilbert space over spectral colors.

3. <u>Human Color Perception</u>: Humans have three types of cone cells for color perception, so the perceivable colors can be represented by 3-dimensional Euclidean space.

4. <u>Mapping Physical to Human Perceived</u>: The many-to-one mapping from the Hilbert space of physical colors to the Euclidean space of human perceivable colors explains why many distinct physical colors may be perceived by humans to be identical, e.g., pure yellow light versus a mix of red and green light, i.e., meta-merism.

## Properties - Pythagorean Identity

1. <u>Orthogonal Vectors in Hilbert Space</u>: Two vectors $u$ and $v$ in a Hilbert space are orthogonal when

$$\langle u, v \rangle = 0$$

The notation for this is

$$u \perp v$$

2. <u>Element wise Orthogonality in Sets</u>: More generally, when $S$ is a subset of $H$, the notation

$$u \perp S$$

means that $u$ is orthogonal to every element from $S$.

3. <u>Norm of $u$ Plus $v$</u>: When $u$ and $v$ are orthogonal, one has

$$\|u + v\|^2 = \langle u + v, u + v \rangle = \langle u, u \rangle + 2\, Re\langle u, v \rangle + \langle v, v \rangle = \|u\|^2 + \|v\|^2$$

4. <u>Extension to $n$ Orthogonal Vectors</u>: By induction on $n$, this is extended to any family $u_1, \cdots, u_n$ of $n$ orthogonal vectors

$$\|u_1 + \cdots + u_n\|^2 = \|u_1\|^2 + \cdots + \|u_n\|^2$$

5. <u>Extending Pythagorean Identity to Series</u>: Whereas Pythagorean identity as stated is valid in any inner space, completeness is required for the extension of the Pythagorean identity to series.

6. <u>Convergence of Orthogonal Vector Series</u>: A series $\sum u_k$ of *orthogonal* vectors converges in $H$ if and only if the series of squares of norms converges, and

$$\left\| \sum_{k=0}^{\infty} u_k \right\|^2 = \sum_{k=0}^{\infty} \|u_k\|^2$$

7. <u>Importance of the Order of Sum</u>: Furthermore, the sum of a series of orthogonal vectors is independent of the order in which it is taken.

## Properties – Parallelogram Identity and Polarization

1. Parallelogram Identity in Hilbert Spaces: By definition, every Hilbert space is also a Banach space. Furthermore, in every Hilbert space the following parallelogram identity holds:

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2)$$

2. Criteria for Banach being Hilbert: Conversely, every Banach space in which the parallelogram identity holds is a Hilbert space, and the inner product is uniquely determined by the norm by the polarization identity (Young (1988)).

3. Geometrical Illustration of Polarization Identity:
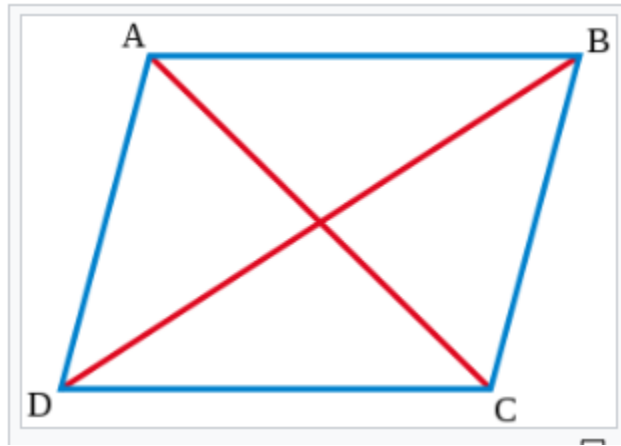


Geometrically, the parallelogram identity asserts that

$$AC^2 + BD^2 = 2(AB^2 + AD^2)$$

In words, the sum of the squares of the diagonals is twice the sum of the squares of any two adjacent sides.

4. <u>Polarization Identity in Real Spaces</u>: For real Hilbert spaces, the polarization identity is

$$\langle u, v \rangle = \frac{1}{4}[\|u + v\|^2 + \|u - v\|^2]$$

5. <u>Polarization Identity in Complex Spaces</u>: For complex Hilbert spaces, it is

$$\langle u, v \rangle = \frac{1}{4}[\|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2]$$

6. <u>Hilbert as Uniformly Complex Banach</u>: The parallelogram law implies that any Hilbert space is a uniformly convex Banach space (Clarkson (1936)).

## Best Approximation

1. <u>Statement of Hilbert Projection Theorem</u>: This subsection employs the Hilbert projection theorem. If $C$ is a non-empty closed convex subset of a Hilbert space $H$ and $x$ a point in $H$, there exists a unique point

$$y \in C$$

that minimizes the distance between $x$ and points in $C$ (Rudin (1987)).

2. <u>Translated Convex Set Minimal Norm</u>: This is equivalent to saying that there is a point with minimal norm in the translated convex set

$$D = C - x$$

3. <u>Proof of the Norm Convergence</u>: The proof consists in showing that every minimizing sequence

$$(d_n) \subset D$$

is Cauchy – using the parallelogram identity, hence converges - using completeness, to a point in $D$ that has minimal norm. More generally, this holds in any uniformly convex Banach space (Dunford and Schwartz (1958)).

4. <u>Application to a Closed Subspace</u>: When this result is applied to a closed subspace $F$ of $H$, it can be shown that the point

$$y \in F$$

closest to $x$ is characterized as

$$y \in F$$

$$x - y \perp F$$

(Rudin (1987)).

5. <u>Orthogonal Projection of $x$ on $F$</u>: This point $y$ is the *orthogonal projection* of $x$ onto $F$, and the mapping

$$P_F : x \rightarrow y$$

is linear – see the section on orthogonal components and projections.

6. <u>Basis of Least Squares</u>: This result is especially significant in applied mathematics, especially analysis, where it forms the basis of least square methods (Blanchet and Charbit (2014)).

7. <u>Dense Vector Subspaces Theorem</u>: In particular, when $F$ is not equal to $H$, one can find a non-zero vector $v$ orthogonal to $F$, i.e., select

$$x \notin F$$

and

$$v = x - y$$

A very useful criterion is obtained by applying this observation to the closed subspace $F$ generated by a subset $S$ of $H$: A subset $S$ of $H$ spans a dense vector subspace if – and only if – the vector 0 is the sole

$$v \in H$$

orthogonal to $S$.

## Duality

1. <u>Definition of the Dual Space</u>: The dual space $H^*$ is the space of all continuous linear functions from the space $H$ into the base field.
2. <u>Norm of the Dual Space</u>: It carries a natural norm, defined by

$$\|\varphi\| = \sup_{\|x\| \in 1, x \in H} |\varphi(x)|$$

3. <u>Dual Space as Inner Product Space</u>: This norm satisfies the parallelogram law, and so the dual space is also an inner product space where this inner product can be defined in terms of this dual norm by using the polarization identity.
4. <u>Dual Space also a Hilbert</u>: The dual space is also complete so it is a Hilbert space in its own right.
5. <u>Inner Product on Dual Space</u>: If

$$e_. = (e_i)_{i \in I}$$

is a complete orthonormal basis for $H$ then the inner product on the dual space of any two

$$f, g \in H^*$$

is

$$\langle f, g \rangle_{H^*} = \sum_{i \in I} f(e_i) \overline{g(e_\iota)}$$

where all but countably many of the terms in this series are zero.

6. <u>Riesz Representation of Dual Space</u>: The Riesz representation theorem affords a convenient description of the dual space. To every element $u$ of $H$ there is a unique element $\varphi_u$ of $H^*$ defined

$$\varphi_u(x) = \langle x, u \rangle$$

where, moreover

$$\|\varphi_u\| = \|u\|$$

7. <u>Riesz Representation as Isometric Antilinear Isomorphism</u>: The Riesz representation theorem states that the map from $H$ to $H^*$ defined by

$$u \mapsto \varphi_u$$

is surjective, which makes this map an isometric antilinear isomorphism.

8. <u>Element Correspondence between $H^*$ and $H$</u>: So, to every element $\varphi$ of the dual $H^*$ there exists one and only one $u_\varphi$ in $H$ such that

$$\langle x, u_\varphi \rangle = \varphi(x)$$

for all

$$x \in H$$

9. <u>Dual Space Inner Product</u>: The inner product on the dual space $H^*$ satisfies

$$\langle \varphi, \psi \rangle = \langle u_\psi, u_\varphi \rangle$$

10. <u>$\varphi$ Linearity vs. $u_\varphi$ Anti-linearity</u>: The reversal of order on the right-hand side restores linearity in $\varphi$ from the anti-linearity of $u_\varphi$.

11. <u>Isomorphic Nature of Hilbert Spaces</u>: In the real case, the anti-linear isomorphism from $H$ to its dual is actually an isomorphism, and so real Hilbert are naturally isomorphic to their own deals.

12. <u>Representing the $u_\varphi$ Vector</u>: The representing $u_\varphi$ is obtained in the following way: when

$$\varphi \neq 0$$

the kernel

$$F = Kernel\,(\varphi)$$

is a closed vector subspace of $H$, both equal to $H$, hence there exists a non-zero vector $v$ orthogonal to $F$.

13. <u>$u$ as Multiple of $v$</u>: The $u$ is a suitable multiple $\lambda v$ of $v$. The requirement that

277

$$\varphi(v) = \langle v, u \rangle$$

yields

$$u = \frac{\overline{\varphi(v)}v}{\langle v, v \rangle}$$

14. Bra-ket Notation of Physics: This correspondence

$$\varphi \leftrightarrow u$$

exploited by the bra-ket notation popular in physics. It is common in physics to assume that the inner product, denoted $\langle x \mid y \rangle$, is linear on the right

$$\langle x \mid y \rangle = \langle y, x \rangle$$

15. Linear Functional Acting on State Vector: The result $\langle x \mid y \rangle$ can also be seen as the action of the linear functional $\langle x \mid$ - the *bra* – on the vector $\mid y \rangle$ - the *ket*.

16. Criteria Underlying Riesz Representation Theorem: The Riesz representation theorem relies fundamentally not just on the presence of the inner product, but also on the completeness of the space. In fact, the theorem implies that the topological dual of any inner product space can be identified with its completion.

17. Reflexive Nature of Hilbert Space: An immediate consequence of the Riesz representation theorem is also that a Hilbert space $H$ is reflexive, meaning that the natural map from $H$ into its double dual space is an isomorphism.

## Weakly Convergent Sequences

1. <u>Weakly Convergent Sequence in $H$</u>:  In a Hilbert space $H$, the sequence $\{x_n\}$ is weakly convergent to a vector

$$x \in H$$

when

$$\lim_n \langle x_n, v \rangle = \langle x, v \rangle$$

for every

$$v \in H$$

2. <u>Weakly Convergent Sequences are Bounded</u>: For example, any orthonormal sequence $\{f_n\}$ converges weakly to $0$, as a consequence of Bessel's inequality. Every weakly convergent sequence $\{x_n\}$ is bounded by the uniform boundedness principle.

3. <u>Weakly Convergent Subsequences from Bounded Sequence</u>: Conversely, every bounded sequence in a Hilbert space admits weakly convergent subsequences – Alaoglu's theorem.

4. <u>Use in Minimization Problems</u>: This fact may be used to prove minimization results for continuous convex functionals, in the same way that the Bolzano-Weierstrass theorem is used for continuous functions on $\mathbb{R}^d$.

5. <u>Convex Continuous Function Minimum - Statement</u>: Among several variants, one simple statement is as follows: If

$$f : H \to \mathbb{R}$$

is a convex continuous function such that $f(x)$ tends to $+\infty$ when $\|x\|$ tends to $\infty$, then $f$ admits a minimum at some point

$$x_0 \in H$$

6. Use in Calculus of Variations: This fact – and its various generalizations – are fundamental for direct methods in the calculus of variations.
7. Closed Bounded Subsets are Weakly Compact: Minimization results for convex functionals are also a direct consequence of the slightly more abstract fact that closed bounded convex subsets in a Hilbert space $H$ are weakly compact, since $H$ is reflexive.
8. Special Case of Eberlein-Smulian Theorem: The existence of weakly convergent subsequences is a special case of the Eberlein-Smulian theorem.

## Banach Space Properties

1. Open Mapping Theorem: Any general property of Banach spaces continues to hold for Hilbert spaces. The open mapping theorem states that a continuous surjective linear transformation from one Banach space to another is an open mapping meaning that it sends open sets to open sets.
2. Bounded Linear Theorem: A corollary is the bounded inverse theorem, that a continuous and bijective linear function from one Banach space to another is an isomorphism, i.e., a continuous linear map whose inverse is also continuous.
3. Proof of Bounded Inverse Theorem: This theorem is considerably simpler to prove in the case of Hilbert spaces than in general Banach spaces (Halmos (1982)).
4. Equivalence to Closed Graph Theorem: The open mapping theorem is equivalent to the closed graph theorem, which asserts that a linear function from one Banach space to another is continuous if and only if its graph is a closed set (Rudin (1973)). In the case of Hilbert spaces, this is basic in the study of unbounded operators.
5. Geometrical Hahn-Banach Theorem: This theorem asserts that a closed convex set can be separated from any point outside it by means of a hyperplane of the Hilbert space.

6. <u>Consequence of the Best Approximation Property</u>: This is an immediate consequence of the best approximation property; if $y$ is the element of a closed convex set $F$ closes to $x$, then the separating hyperplane is the plane perpendicular to the segment $xy$ passing through its midpoint (Treves (1967)).

## Bounded Operators on Hilbert Spaces

1. <u>Definition of Bounded Linear Operators</u>: The continuous linear operators

$$A : H_1 \rightarrow H_2$$

from a Hilbert space $H_1$ to a second Hilbert space $H_2$ are *bounded* in the sense that they map bounded sets to bounded sets. Conversely, if an operator is bounded, then it is continuous.

2. <u>Norm of Bounded Linear Operators</u>: The space of such bounded linear operators has a norm, the operator norm given by

$$\|A\| = \sup_{\|x\| = 1} \|Ax\|$$

3. <u>Composition of Bounded Linear Operators</u>: The sum and the composite of two bounded linear operators is again bounded and linear.

4. <u>Composition of $x \in H_1$ and $y \in H_2$</u>: For $y$ in $H_2$, the map that sends

$$x \in H_1$$

to $\langle Ax, y \rangle$ is linear and continuous, and according to the Riesz representation theorem can therefore be represented in the form

$$\langle x, A^*y \rangle = \langle Ax, y \rangle$$

for some vector $A^*y$ in $H_1$.

5. <u>Adjoint of Riesz Representation Operator</u>: This defines another bounded linear operator

$$A^* : H_2 \to H_1$$

the adjoint of $A$. The adjoint satisfies

$$A^{**} = A$$

6. <u>Intuition behind Reisz Adjoint Operator</u>: When the Reisz representation theorem is used to identify each Hilbert space with its continuous dual space, the adjoint of $A$ can be shown to be identical to the transpose

$$A^T : H_2^* \to H_1^*$$

of $A$, which, by definition, sends

$$\psi \in H_2^*$$

to the functional

$$\psi \circ A \in H_1^*$$

7. <u>Definition of $C^*$ - Operator Algebra</u>: The set $B(H)$ of all bounded linear operators on $H$ – meaning

$$H \to H$$

together with the addition and the composition operations, the norm and the adjoint operation, is a $C^*$ −algebra, which is a type of operator algebra.

8. <u>Elements of Bounded Operator Set</u>: An element $A$ of $B(H)$ is called 'self-adjoint' or 'Hermitian' if

$$A^* = A$$

9. <u>Elements that are Positive/Non-negative</u>: If $A$ is Hermitian and

$$\langle Ax, x \rangle \geq 0$$

for every $x$, then $A$ is called 'non-negative', written

$$A \geq 0$$

if equality holds only when

$$x = 0$$

then $A$ is called 'positive'.

10. <u>Partial Order among Self-adjoint Operators</u>: The set of self-adjoint operators admits a partial order, in which

$$A \geq B$$

if

$$A - B \geq 0$$

11. <u>Non negative and Positive Elements</u>: If $A$ has the form $B^*B$ for some $B$, then $A$ is non-negative; if $B$ is invertible, then $A$ is positive.

12. <u>Square-root of Self-adjoint Operators</u>: A converse is also true in the sense that, for a non-negative operator $A$, there exists a unique non-negative square root $B$ such that

$$A = B^2 = B^*B$$

13. <u>Self-adjoint Operators are Real</u>: In a sense made precise by the spectral theorem, self-adjoint operators can usefully be thought of operators that are "real".

14. <u>Normal Bounded Linear Operators</u>: An element $A$ of $B(H)$ is called normal if

$$A^*A = AA^*$$

Normal operators decompose into the sum of a self-adjoint operator and an imaginary multiple of a self-adjoint operator

$$A = \frac{A + A^*}{2} + i\frac{A - A^*}{2i}$$

that commute with each other. Normal operators can also usefully be thought of in terms of their real and imaginary parts.

15. <u>Defining Unitary Bounded Linear Operators</u>: An element $U$ of $B(H)$ is called unitary if $U$ is invertible and its inverse is given by $U^*$. This can also be expressed by requiring that $U$ be into and

$$\langle Ux, Uy \rangle = \langle x, y \rangle$$

for all

$$x, y \in H$$

The unitary operators form a group under composition, which is the isometry group of $H$.

16. <u>Defining Compact Bounded Linear Operators</u>: An element of $B(H)$ is compact if it sends bounded sets to relatively compact sets.

17. <u>Definition of Bounded, Compact Operators</u>: Equivalently, a bounded operator $T$ is compact if, for any bounded sequence $\{x_k\}$, the sequence $\{Tx_k\}$ has a convergent subsequence.

18. <u>Hilbert-Schmidt - Compact Integral Operators</u>: Many integral operators are compact, and in fact define a special class of operators known as Hilbert-Schmidt operators that are especially important in the study of integral equations.

19. <u>Distinction from a Fredholm Operator</u>: Fredholm operators differ from compact operator by a multiple of the identity, and are equivalently characterized as operators with a finite dimensional kernel and co-kernel.

20. <u>Index of the Fredholm Operator</u>: The index of a Fredholm operator $T$ is defined by

$$index\ T = \dim(Kernel\ T) - \dim(coKernel\ T)$$

21. <u>Homotopy Invariance of Operator Index</u>: The index is homotopy invariant, and plays a deep role in differential geometry via the Atiyah-Stinger index theorem.

## Unbounded Operators on Hilbert Space

1. <u>Unbounded, Tractable Hilbert Space Operators</u>: Unbounded tractable are also Hilbert spaces, and have important applications to quantum mechanics (Reed and Simon (1980), Prugovecki (1981), Folland (1989)).

2. <u>Definition of the Unbounded Operator</u>: An unbounded operator $T$ on a Hilbert space $H$ is defined as a linear operator $D(T)$ whose domain is a linear subspace of $H$.

3. <u>Domain of Densely Defined Operator</u>: Often the domain $D(T)$ is a dense subspace of $H$, in which case $T$ is known as a densely defined operator.

4. <u>Densely Defined Unbounded Operator Adjoint</u>: The adjoint of a densely defined unbounded operator is defined in essentially the same manner as for bounded operator.

5. <u>Self-bounded Unbounded Operator in QM</u>: Self-adjoint unbounded operators play the role of the *observables* in the mathematical formulation of quantum mechanics.

6. <u>Examples of Self-adjoint Unbounded Operators</u>: Prugovecki (1981) provided the following examples on the Hilbert space $L^2(\mathbb{R})$.

7. <u>Extension of the Differential Operator</u>: The suitable extension

$$(Af)(x) = -i\frac{d}{dx}f(x)$$

where $i$ is the imaginary unit and $f$ is a differential function of compact support.

8. <u>The Multiplication by $x$ Factor</u>:

$$(Bf)(x) = xf(x)$$

9. <u>Momentum and Position Operators</u>: The differential and the multiplication-by-$x$ operators correspond to the momentum and the position observables, respectively.

10. <u>Domain of Position/Momentum Operators</u>: Neither that neither position nor momentum is defined on all of $H$, since in the case of the momentum operator the derivative need not exist, and in the case of the position operator the product function need not be square integrable.

11. <u>Domain of Position/Momentum Operators</u>: In both cases, the set of possible arguments form dense subspace of $L^2(\mathbb{R})$.

## Constructions – Direct Sums

1. Orthogonal, External, Direct Hilbert Sum: Two Hilbert spaces $H_1$ and $H_2$ can be combined into another Hilbert space, called the orthogonal direct sum (Dunford and Schwartz (1958)), and denoted

$$H_1 \oplus H_2$$

consisting of the set of all ordered pairs $(x_1, x_2)$ where

$$x_i \in H_i$$

$$i = 1, 2$$

and inner product defined by

$$\langle (x_1, x_2), (y_1, y_2) \rangle_{H_1 \oplus H_2} = \langle x_1, y_1 \rangle_{H_1} + \langle x_2, y_2 \rangle_{H_2}$$

2. Cartesian Product across Component Spaces: More generally, if $H_i$ is a family of Hilbert spaces indexed by

$$i \in I$$

then the direct sum of the $H_i$, denoted

$$\bigoplus_{i \in I} H_i$$

consists of the set of all indexed families

$$x = (x_i \in H_i \mid i \in I) \in \prod_{i \in I} H_i$$

in the Cartesian product of the $H_i$ such that

$$\sum_{i \in I} \|x_i\|^2 < \infty$$

3. <u>Defining the Cartesian Inner Product</u>: The inner product is defined by

$$\langle x, y \rangle = \sum_{i \in I} \langle x_i, y_i \rangle_{H_i}$$

4. <u>Orthogonal, Closed Individual $H_i$ Subspaces</u>: Each of the $H_i$ is included as a closed subspace in the direct sum of all $H_i$. Moreover, the $H_i$ are pairwise orthogonal.

5. <u>Orthogonal Internal Direct Hilbert Sum</u>: Conversely, if there is a system of closed subspaces $V_i$

$$i \in I$$

in a Hilbert space $H$ that are pairwise orthogonal and whose union is dense in $H$, then $H$ is canonically isomorphic to the direct sum of $V_i$. In this case $H$ is called the internal direct sum of $V_i$.

6. <u>Orthogonal Projections of Direct Sums</u>: A direct sum – internal or external – is also equipped with a family of orthogonal projections $E_i$ on the $i^{\text{th}}$ direct summand $H_i$. These projections are bounded, self-adjoint, idempotent operators that satisfy the orthogonality condition

$$E_i E_j = 0$$

$$i \neq j$$

7. <u>Spectral Theorem for Self-adjoint Operators</u>: The spectral theorem for compact self-adjoint operators on a Hilbert space $H$ states that $H$ splits into an orthogonal direct sum of the eigenspaces of an operator, and also gives an explicit decomposition of the operator as a sum of projections onto the eigenspaces.

8. <u>Use in Quantum Mechanics</u>: The direct sum of the Hilbert space also appears in quantum mechanics as the Fock space of a system containing a variable number of particles, where each Hilbert space in the direct sum corresponds to an additional degree of freedom for the quantum mechanical systems.

9. <u>Implication of Peter-Weyl Theorem</u>: In representation theory, the Peter-Weyl theorem guarantees that any unitary representation of a compact group on a Hilbert space splits as a direct sum of finite-dimensional representations.

## Constructions – Tensor Products

1. <u>Ordinary Tensor Inner Product – Definition</u>: If

$$x_1, y_1 \in H$$

and

$$x_2, y_2 \in H$$

then one defines the inner product on the ordinary tensor product as follows. On simple tensors, let

$$\langle x_1 \otimes x_2, y_1 \otimes y_2 \rangle = \langle x_1, y_1 \rangle \otimes \langle x_2, y_2 \rangle$$

2. <u>Extending Definition to Hilbert Spaces</u>: This formula then extends sesqui-linearity to an inner product on

$$H_1 \otimes H_2$$

3. <u>Definition of Hilbertian Tensor Product</u>: The Hilbertian tensor product of $H_1$ and $H_2$, sometimes denoted

$$H_1 \widehat{\otimes} H_2$$

is the Hilbert space is obtained by completing the

$$H_1 \otimes H_2$$

for the metric associated with this tensor product.

4. <u>Example - Hilbert Space of $L_2([0,1])$</u>: An example is provided by the Hilbert space $L_2([0,1])$. The Hilbertian tensor product of two copies of $L_2([0,1])$ is isometrically and linearly isomorphic to the space $L_2([0,1]^2)$ of square-integrable functions on the square $[0,1]^2$.

5. <u>Function Isomorphism in Hilbert Space</u>: This isomorphism sends a simple tensor

$$f_1 \otimes f_2$$

to the function

$$(s,t) \longmapsto f_1(s)f_2(t)$$

on the square.

6. <u>Corresponding Explicit Tensor Space Mapping</u>: This example is typical in the following sense (Kadison and Ringrose (1983)). Associated with every simple tensor product

$$x_1 \otimes x_2$$

is the rank one operator from $H_1^*$ to $H_2$ that maps a given

$$x^* \in H_1^*$$

as

$$x^* \longmapsto x^*(x_1)x_2$$

7. <u>Extension to Finite Rank Operators</u>: This mapping defined on the simple tensors extends to a linear identification between

$$H_1 \otimes H_2$$

and the space of finite rank operators from $H_1^*$ to $H_2$.

8. <u>Implied Hilbertian Tensor Product Space</u>: This extends to a linear isometry of the Hilbertian tensor product

$$H_1 \widehat{\otimes} H_2$$

with the Hilbert space $HS(H_1^*, H_2)$ of Hilbert-Schmidt operators from $H_1^*$ to $H_2$.

## Orthonormal Bases

1. <u>Linear Algebra to Hilbert Spaces</u>: The notion of an orthonormal basis from algebra generalizes over to the case of Hilbert spaces (Dunford and Schwartz (1958)).

2. <u>Criteria Satisfied by Orthonormal Basis</u>: In a Hilbert space $H$, an orthonormal basis is a family $\{e_k\}_{k \in B}$ of elements of $H$ satisfying the following conditions.

3. Orthogonality: Every two elements of $B$ are orthogonal;

$$\langle e_k, e_j \rangle = 0$$

for all

$$k, j \in B$$

with

$$k \neq j$$

4. Normalization: Every element of the family has the norm 1:

$$\|e_k\| = 1$$

for all

$$k \in B$$

5. Completeness: The linear span of the family $e_k$

$$k \in B$$

is dense in $H$.

6. Definition of Orthonormal System/Sequence: A system of vectors satisfying the first two conditions called an orthonormal basis or an orthonormal set – or an orthonormal sequence if $B$ is countable. Such a system is always linearly independent.

7. Completeness of an Orthogonal System: Completeness of an orthogonal system of vectors of a Hilbert space can be equivalently re-stated as: if

$$\langle v, e_k \rangle = 0$$

for all

$$k \in B$$

and some

$$v \in H$$

then

$$v = 0$$

8. <u>Rationale behind the Completeness Criterion</u>: This is related to the fact that the only vector orthogonal to a dense linear subspace is the zero vector, for if $S$ is any orthonormal set and $v$ is orthogonal to $S$, then $v$ is orthogonal to the closure of the linear span of $S$, which is the whole space.

9. <u>Example of Orthonormal Basis #1</u>: The set $\{(1,0,0),(0,1,0),(0,0,1)\}$ forms an orthonormal basis of $\mathbb{R}^3$ with the dot product.

10. <u>Example of Orthonormal Basis #2</u>: The sequence

$$\{f_n : n \in \mathbb{Z}\}$$

with

$$f_n(x) = e^{2\pi i n x}$$

forms an orthonormal basis of the complex $L_2([0,1])$.

11. <u>Orthonormal Basis is Infinite Dimensions</u>: In the infinite-dimensional case, an orthonormal basis will not be a basis in the sense of linear algebra; to distinguish the two, the latter basis is also called a Hamel basis.

12. <u>Unique Nature of Basis Decomposition</u>: That the span of the basis vectors is dense implies that every vector in the space can be written as the sum of an infinite series, and the orthogonality implies that this decomposition is unique.

## Sequence Spaces

1. <u>$l_2$ Space of Square-summable Sequences</u>: The space $l_2$ of square-summable sequences of complex numbers is the set of infinite sequences $(c_1, c_2, c_3, \cdots)$ of real or complex numbers such that

$$|c_1|^2 + |c_2|^2 + |c_3|^2 + \cdots < \infty$$

2. <u>Space of the Orthonormal Basis</u>:

$$e_1 = (1, 0, 0, \cdots)$$

$$e_2 = (0, 1, 0, \cdots)$$

$$e_3 = (0, 0, 1, \cdots)$$

$$\vdots$$

3. <u>Infinite-dimensional Generalization of $l_2^n$ Space</u>: This space is the infinite-dimensional generalization of the $l_2^n$ space of finite-dimensional vectors.

4. <u>Closed/Bounded, but not Compact</u>: The above example is usually the first example used to show that in infinite-dimensional spaces, a set that is closed and bounded is not necessarily sequentially compact – as is the case in all *finite* dimensional spaces.

5. <u>Example - Set of Orthonormal Vectors</u>: Indeed, the set of orthonormal vectors above shows that it is an infinite sequence of vectors in the unit ball, i.e., the ball of point with norm less than or equal to one.

6. <u>Bounded Set with No Convergence</u>: This set is clearly bounded and closed; yet, no sub-sequence of these vectors converges to anything and consequently the unit ball in $l_2$ is not compact.

7. <u>Intuition behind such Compact Sets</u>: Intuitively, this is because "there is always another coordinate direction" into which the next elements of the sequence can evade.

8. <u>Generalization of the $l_2$ Sequence</u>: One can generalize the space $l_2$ in many ways. For example, if $B$ is any – infinite – set, then one can form a Hilbert space of sequences with set $B$, defined by

$$l_2(B) = \left\{ x : B \xrightarrow{x} C \mid \sum_{b \in B} |x(b)|^2 < \infty \right\}$$

9. <u>Norm over the Index Set</u>: The summation over $B$ is here defined by

$$\sum_{b \in B} |x(b)|^2 = \sup_{b \in B} \sum_{n=1}^{N} |x(b_n)|^2$$

10. <u>Countably Non-zero Index Set Items</u>: It follows that, for this sum to be finite, every element of $l_2(B)$ has only countably many non-zero items.

11. <u>Inner Product over Base Set Space</u>: This space becomes a Hilbert space with the inner product

$$\langle x, y \rangle = \sum_{b \in B} x(b)\overline{y(b)}$$

for all

$$x, y \in l_2(B)$$

12. <u>Countably many Non-zero Items</u>: Here the sum also has only countably many non-zero items, and is unconditionally convergent by the Cauchy-Schwarz inequality.

13. <u>Orthonormal Basis of $l_2(B)$</u>: An orthonormal basis of $l_2(B)$ is indexed by the set $B$ given by

$$e_b(b') = \begin{cases} 1 & if \ b = b' \\ 0 & otherwise \end{cases}$$

## Bessel's Inequality and Parseval's Formula

1. <u>Decomposition using Dot Product Loadings</u>: Let $f_1, \cdots, f_n$ be a finite orthonormal system in $H$. For an arbitrary vector

$$x \in H$$

let

$$y = \sum_{j=1}^{n} \langle x, f_j \rangle f_j$$

2. <u>Pythagoras Applied on $x - y \ / \ y$</u>: Then

$$\langle x, f_k \rangle = \langle y, f_k \rangle$$

296

for every

$$k = 1, \cdots, n$$

It follows that $x - y$ is orthogonal to each $f_k$, hence $x - y$ is orthogonal to $y$. Using the Pythagorean identity twice, it follows that

$$\|x\|^2 = \|x - y\|^2 + \|y\|^2 \geq \|y\|^2 = \sum_{j=1}^{n} \left|\langle y, f_j \rangle\right|^2$$

3.  Statement of the Bessel's Inequality: Let $\{f_i\}$

$$i \in I$$

be an arbitrary orthonormal system in $H$. Applying the preceding to every finite subset $J$ of $I$ gives Bessel's inequality:

$$\sum_{i \in I} |\langle x, f_i \rangle|^2 \leq \|x\|^2$$

$$x \in H$$

according to the definition – of the sum of an arbitrary family of non-negative real numbers.

4.  Finite vs Infinite Index Sets: For the case Bessel's inequality for the finite index sets, Holmes (1957) contains details. Other treatments contain details for infinite index sets.

5.  Geometric Intuition behind Bessel's Inequality: Geometrically, Bessel's inequality implies that the orthogonal projection of $x$ onto the linear subspace spanned by the $f_i$ has norm that does not exceed that of $x$.

6. <u>Bessel's Inequality in Euclidean Dimensions</u>: In two dimensions, this is the assertion that the length of the leg of a right triangle may not exceed the length of the hypotenuse.

7. <u>Motivation behind the Parseval's Identity</u>: Bessel's inequality is a stepping stone to the stringer result called Parseval's identity, which governs the case when Bessel's inequality is actually an equality.

8. <u>Orthonormal Decomposition using Hilbert Basis</u>: By definition, is $\{e_k\}_{k \in B}$ is an orthonormal basis of $H$, then every element of $x$ of $H$ may be written as

$$x = \sum_{k \in B} \langle x, e_k \rangle e_k$$

9. <u>Countably Non-zero Index Set Entities</u>: Even if $B$ is uncountable, Bessel's inequality guarantees that the expression is well-defined and consists only of countably many non-zero terms.

10. <u>Fourier Expansion and Coefficients</u>: This sum is called the Fourier expansion of $x$, and the individual coefficients $\langle x, e_k \rangle$ are the Fourier coefficients of $x$.

11. <u>Statement of the Parseval's Identity</u>: Parseval's identity then asserts that

$$\|x\|^2 = \sum_{k \in B} |\langle x, e_k \rangle|^2$$

12. <u>Orthonormal Set where Parseval's Identity Holds</u>: Conversely, if $\{e_k\}$ is an orthonormal set such that Parseval's identity holds for every $x$, then $\{e_k\}$ is an orthonormal basis.


## Hilbert Dimension

1. Cardinality of the Hilbert Space: As a consequence of Zorn's lemma, *every* Hilbert space admits an orthonormal basis; furthermore, any two orthonormal basses of the same space have the same cardinality, called the Hilbert dimension of the space.

2. Finite vs Infinite Hilbert Dimensions: Many authors, such as Dunford and Schwartz (1958), refer to this just as the dimension. Unless the Hilbert space is finite dimensional, this is not the same thing as its dimension as a linear space – the cardinality of a Hamel basis.

3. Orthonormal Basis Indexed by $B$: For instance, since $l_2(B)$ has an orthonormal basis indexed by $B$, its Hilbert dimension is the cardinality of $B$ – which may be a finite integer, or a countable or uncountable cardinal number.

4. Dot Product over Hilbert Space: As a consequence of Parseval's identity, of $\{e_k\}_{k \in B}$ is an orthonormal basis of $H$, then the map

$$\Phi : H \longrightarrow l_2(B)$$

defined by

$$\Phi(x) = \langle x, e_k \rangle_{k \in B}$$

is an isometric isomorphism of Hilbert spaces; it is a bijective linear mapping such that

$$\langle \Phi(x), \Phi(y) \rangle_{l_2(B)} = \langle x, y \rangle_H$$

for all

$$x, y \in H$$

5. Isometric Isomorphism between $H$ and $B$: The cardinal number of $B$ is the Hilbert dimension of $H$. Thus, every Hilbert space is isometrically isomorphic to a sequence space $l_2(B)$ for some set of $B$.

## Separable Spaces

1. Definition of Separable Hilbert Spaces: By definition, a Hilbert space is separable provided it contains a dense countable subset.
2. Countable Orthonormal Hilbert Space Basis: Along with Zorn's lemma, this means a Hilbert space is separable if and only if it admits a countable orthonormal basis.
3. Infinite-dimensional Separable Hilbert Spaces: All infinite-dimensional separable Hilbert spaces are therefore isometrically isomorphic to $l_2$.
4. Separable Hilbert Spaces - Past Definitions: In the past, Hilbert spaces were often required to be separable as part of the definition (Prugovecki (1981)).
5. Isometric and Separable Hilbert Spaces: Most spaces used in physics are separable, and since these are all isomorphic to each other, one often refers to any infinite-dimensional separable Hilbert as "*the* Hilbert space" or just Hilbert space".
6. von Neumann's Countable Hilbert Space: von Neumann (1955) defines a Hilbert space via a countable Hilbert basis which amounts to an isometric isomorphism with $l_2$. The conversion still persists in most rigorous treatments of quantum mechanics (Sobrino (1996)).
7. Wightman Axioms of Quantum Field Theory: Even in quantum field theory, most of the Hilbert spaces are in fact separable, as stipulated by the Wightman axioms.
8. Non-separable Hilbert Spaces in Quantum Field Theory: However, it is sometime argued that non-separable Hilbert spaces are also important in quantum field theory, roughly because the systems in the theory possess an infinite number of degrees of freedom and any infinite Hilbert tensor product – of spaces of dimension greater than one – is non-separable (Streater and Wightman (1964)).

9. Example - The Bosonic Field: For instance, a bosonic field can be naturally thought of as an element of a tensor product whose factors represent harmonic oscillators at each point of space.

10. Separable Subspace that is Observable: From this perspective, the natural state space of a boson might seem to be a non-separable space (Streater and Wightman (1964)). However, it is only a small separable subspace of the full tensor product that can contain physically meaningful fields – on which the observables can be defined.

11. Second Non-separable Hilbert Space Example: Another non-separable Hilbert space models the state of an infinite collection of particles in an unbounded region of space.

12. Non countability of the Basis: An orthonormal basis of the basis is indexed by the density of the particles, a continuous parameter, and since the set of possible densities is uncountable, the basis is uncountable (Streater and Wightman (1964)).

## Orthogonal Complements and Projections

1. Given Set Orthogonal to Subset: If $S$ is a subset of Hilbert space $H$, the set of vectors orthogonal to $S$ is defined

$$\{S^\perp = x \in H : \langle x, s \rangle = 0 \ \forall \ s \in S\}$$

2. $S^\perp$ as a Closed Subspace: The set $S^\perp$ is a closed subspace of $H$ – this can be proved easily using the linearity and the continuity of the inner product – and so forms itself a Hilbert space.

3. $V^\perp$ as an Orthogonal Component: If $V$ is a closed subspace of $H$, then $V^\perp$ is called the *orthogonal complement* of $V$.

4. Decomposition into $V / V^\perp$ Components: In fact, every

$$x \in H$$

can then be written uniquely as

$$x = v + w$$

with

$$v \in V$$

and

$$w \in V^\perp$$

5. $\underline{V/V^\perp}$ - Internal Direct Sum: Therefore, $H$ is the internal Hilbert direct sum of $V$ and $V^\perp$.

6. Orthogonal Projection onto Hilbert Space: The linear operator

$$P_V : H \longrightarrow H$$

that maps $x$ to $v$ is called the *orthogonal projection* onto $V$.

7. Closed Subspace vs Self-adjoint Operators: There is a natural one-to-one correspondence between the set of all closed subspaces of $H$ and the set of all bounded self-adjoint operators $P$ such that

$$P^2 = P$$

8. Orthogonal Projection Theorem Statement #1: The orthogonal projection $P_V$ is a self-adjoint linear operator on $H$ of norm less than or equal to 1 with the property

$$P_V^2 = P_V$$

9. Orthogonal Projection Theorem Statement #2: Moreover, any self-adjoint linear operator $E$ such that

$$E^2 = E$$

is of the form

$$P_V^2 = P_V$$

10. Orthogonal Projection Theorem Statement #3: For every $x$ in $H$, $P_V(x)$ is the unique element $v$ of $V$ that minimizes the distance $\|x - v\|$.

11. Geometrical Intuition behind Projection Operator: This provides the geometrical interpretation of $P_V(x)$: it is the best approximation to $x$ by elements of $V$ (Young (1988)).

12. Mutually Orthogonal Projection Subspaces: Projections $P_U$ and $P_V$ are called mutually orthogonal if

$$P_U P_V = 0$$

This is equivalent to $U$ and $V$ being orthogonal as subspaces of $H$.

13. Sum of Projections being Orthogonal: The sum of the two projections $P_U$ and $P_V$ is a projection only if $U$ and $V$ are orthogonal to each other, and in that case

$$P_U + P_V = P_{U+V}$$

14. Orthogonality of the Composition Projection: The composite $P_U P_V$ is generally not a projection: in fact, the composite is a projection if and only if the two projections commute, and in that case

$$P_U P_V = P_{U \cap V}$$

15. <u>Inclusion Mapping - Adjoint of Projection</u>: By describing the co-domain to the Hilbert space $V$, the orthogonal projection $P_V$ gives rise to a projection mapping

$$\pi : H \to V$$

It is the adjoint of the inclusion mapping

$$i : V \to H$$

meaning that

$$\langle ix, y \rangle_H = \langle x, \pi y \rangle_V$$

for all

$$x \in V$$

and

$$y \in H$$

16. <u>Operator Norm of Orthogonal Projection</u>: The operator norm of the orthogonal projection $P_V$ onto a non-zero closed subspace $V$ is equal to 1:

$$\|P_V\| = \underset{x \in H, x \neq 0}{\sup} \frac{\|P_V x\|}{\|x\|} = 1$$

17. <u>Key Characteristic of Closed Subspace</u>: Every closed subspace $V$ of a Hilbert space is therefore the image of an operator $P$ of an operator $P$ of norm one such that

$$P^2 = P$$

18. <u>Kakutani's Characterization of Hilbert Spaces</u>: The property below of possessing appropriate projection operators characterize Hilbert spaces (Kakutani (1939)).

19. <u>When is Banach Projection a Hilbert Space</u>? A Banach space of dimension higher than 2 is – isometrically – a Hilbert space if and only if, for every closed subspace $V$, there is an operator $P_V$ of norm one whose image is $V$ such that

$$P_V^2 = P_V$$

20. <u>Lindenstrauss and Tzafriri Hilbert Space Characteristics</u>: While the above result characterizes the metric structure of a Hilbert space, the structure of a Hilbert space as a topological vector space can itself be characterized in terms on the presence of complementary subspaces below (Lindenstrauss and Tzafriri (1971)).

21. <u>Topological Banach Spaces as Hilbert</u>: A Banach space $X$ is topologically and linearly isomorphic to a Hilbert space if and only if, to every closed subspace $V$, there is a closed subspace $W$ such that $X$ is equal to the internal direct sum

$$V \oplus W$$

22. <u>Monotone Property of Orthogonal Complement</u>: The orthogonal complement satisfies some more elementary results. It is a monotone function in the sense that if

$$U \subset V$$

then

$$V^\perp \subseteq U^\perp$$

with the equality holding of and only if $V$ is contained in the closure of $U$. This result is a special case of the Hahn-Banach theorem.

23. <u>Closure of Orthogonal Component Subspace</u>: The closure of a subspace can be completely characterized in terms of the orthogonal complement: if $V$ is a subspace of $H$, then the closure of $V$ is equal to $V^{\perp\perp}$

24. <u>Orthogonal Complement as Galois Connection</u>: The orthogonal complement is thus a Galois connection on the partial order of the subspaces of a Hilbert space.

25. <u>Orthogonal Complement Space as Intersection</u>: In general, the orthogonal complement of a sum of subspaces is the intersection of the orthogonal complements (Halmos (1957)):

$$\left(\sum_i V_i\right)^{\perp} = \bigcap_i V_i^{\perp}$$

26. <u>Case when $V_i$ are Closed</u>: If the $V_i$ are in addition closed, then

$$V_i^+ = \left(\bigcap_i V_i\right)^{\perp}$$

## Spectral Theory

1. <u>Spectral Theory for Self-adjoint Operators</u>: There is a well-developed spectral theory for self-adjoint operator in a Hilbert space, that is roughly analogous to the study of symmetric matrices over the reals or self-adjoint matrices over complex numbers.

2. <u>Literature Treatment of Spectral Theory</u>: A general account of spectral theory in Hilbert spaces can be found in Riesz and Sz-Nagy (1990). A more sophisticated account in the language of $C^*$- algebras is in Rudin (1973).

3. <u>Diagonalization of Self-adjoint Operators</u>: In the same sense, one can obtain a "diagonalization" of a self-adjoint operator as a suitable sum – actually an integral – of orthogonal projection operators.

4. <u>Spectrum of a Self-adjoint Operator</u>: The spectrum of an operator $T$, denoted $\sigma(T)$, is the set of complex numbers $\lambda$ such that $T - \lambda$ lacks a continuous inverse.

5. <u>Compact Set in the Complex Plane</u>: If $T$ is bounded, then the spectrum is always a compact set in the complex plane, and lies inside the disc

$$\|z\| \leq \|T\|$$

6. <u>Range of the Spectrum</u>: If $T$ is elf-adjoint, then the spectrum is real. In fact, it is contained in the interval $[m, M]$ where

$$m = \inf_{\|x\| = 1} \langle Tx, x \rangle$$

$$M = \sup_{\|x\| = 1} \langle Tx, x \rangle$$

Moreover, $m$ and $M$ are both actually contained within the spectrum.

7. <u>Eigenspaces of an Operator $T$</u>: The eigenspaces of an operator $T$ are given by

$$H_\lambda = Kernel(T - \lambda)$$

8. <u>Elements of the Operator Spectrum</u>: Unlike with finite matrices, not every element of the spectrum $T$ must be an eigenvalue: the linear operator $T - \lambda$ may only lack an inverse because it is not surjective.

9. <u>Spectral Values of the Operator</u>: Elements of the spectrum of an operator in the general sense are known as *spectral values*.

10. <u>Subtleties of the Spectral Decomposition</u>: Since spectral values need not be eigenvalues, the spectral decomposition is often mor subtle in finite dimensions.

11. <u>Spectral Theorem for Compact Operators</u>: However, the spectral theorem of a self-adjoint operator $T$ takes a particularly simple form if, in addition, $T$ is assumed to be a compact operator. The spectral theorem for compact self-adjoint is given below.

12. <u>Countably Many Spectral Values</u>: A compact self-adjoint operator $T$ has only countably – or finitely – many spectral values. The spectrum of $T$ has no limit point in the complex plane except possibly zero.

13. <u>Decomposing the Eigenspaces of the Operator</u>: The eigenspaces of $T$ decompose $H$ into an orthogonal direct sum

$$H = \bigoplus_{\lambda \in \sigma(T)} H_\lambda$$

14. <u>Projection onto the Orthogonal Eigenspace</u>: Moreover, if $E_\lambda$ denotes the orthogonal projection onto the eigenspace $H_\lambda$, then

$$T = \sum_{\lambda \in \sigma(T)} \lambda E_\lambda$$

where the sum converges with to the norm on $B(H)$.

15. <u>Literature on the Spectral Theory</u>: Treatments include Riesz and Sz-Nagy (1990). The result was already known to Schmidt (1908) in the case of operators arising from integral kernels.

16. <u>Use in the Theory of Integral Equations</u>: This theorem plays a fundamental role in the theory of integral equations, as many integral operators are compact, in particular those that arise from Hilbert-Schmidt operators.

17. <u>Operator-valued Riemann-Stieltjes Integral</u>: The general spectral theorem for self-adjoint operators involves a kind of operator-valued Riemann-Stieltjes integral, rather than an infinite summation (Riesz and Sz-Nagy (1990)).

18. <u>Element of the Spectral Family</u>: The *spectral family* associated with $T$ links each real number $\lambda$ to an operator $E_\lambda$, which is the projection onto the null space of the operator $(T - \lambda)^+$, where the positive part of a self-adjoint operator is defined by

$$A^+ = \frac{1}{2}\left[\sqrt{A^2} + A\right]$$

19. <u>Monotone Increasing Nature of $E_\lambda$</u>: The operators $E_\lambda$ are monotone increasing relative to the partial order defined – on self-adjoint operators; the eigenvalues correspond precisely to the jump discontinuities.

20. <u>Statement of the Spectral Theorem</u>: The spectral theorem asserts that

$$T = \int_{\mathbb{R}}^{\mathbb{R} } \lambda dE_\lambda$$

21. <u>Bounding of the Integral Norm</u>: The integral is understood as a Riemann-Stieltjes integral, convergent with respect to the norm on $B(H)$.

22. <u>Ordinary Scalar-valued Integral Representation</u>: In particular, one has the ordinary scalar-valued integral representation

$$\langle Tx, y \rangle = \int_{\mathbb{R}}^{\mathbb{R}} \lambda d\langle E_\lambda x, y \rangle$$

23. <u>Spectral Decomposition for Normal Operators</u>: A somewhat similar spectral decomposition holds for normal operators, although because the spectrum may now contain non-real complex numbers, the operator-valued Stieltjes measure $\Delta E_\lambda$ must instead be replaced by a resolution of the identity.

24. <u>Application to Spectral Mapping Theorem</u>: A major application of spectral methods is the spectral mapping theorem, which allows one to apply to a self-adjoint operator $T$ any continuous complex function on the spectrum of $T$ by forming the integral

$$f(T) = \int\limits_{\sigma(T)}^{\sigma(T)} f(\lambda)dE_\lambda$$

25. <u>Continuous Operator Functional Calculus</u>: The resulting continuous functional calculus has applications in particular to pseudo-differential operators (Shubin (1987)).

26. <u>Case of *Unbounded* Self-adjoint Operators</u>: The spectral theory of *unbounded* self-adjoint operators is only marginally more difficult than that for unbounded operators.

27. <u>Spectrum of an Unbounded Operator</u>: The spectrum of an unbounded operator is defined in precisely the same way as for bounded operators: $\lambda$ is a spectral value if the resolvent operator

$$R_\lambda = (T - \lambda)^{-1}$$

fails to be a well-defined continuous operator.

28. <u>Guarantee that Spectrum is Real</u>: The self-adjointedness of $T$ still guarantees that the spectrum is real. Thus, the essential idea of working with unbounded operators is to look instead at the resolvent $R_\lambda$ where $\lambda$ is non-real.

29. <u>*Bounded* Normal Operator</u>: This is a *bounded* normal operator, admits a spectral representation that can be transferred to a spectral representation of $T$ itself.

30. <u>Spectrum from an Associated Resolvent</u>: A similar strategy is used, for instance, to study the spectrum of the Laplace operator: rather than address the operator directly, one instead looks as an associated resolvent such as Riesz potential or Bessel potential.

31. <u>Precise Version of the Spectral Theorem</u>: The following statement on the spectral theorem below comes from Rudin (1973).

32. <u>Self-adjoint Operator Spectral Theorem</u>: Given a densely defined self-adjoint operator $T$ on a Hilbert space $H$, there corresponds a unique resolution of the identity $E$ on the Borel sets of $\mathbb{R}$, such that

$$\langle Tx, y \rangle = \int\limits_{\mathbb{R}}^{\mathbb{R}} \lambda dE_{x,y}(\lambda)$$

for all

$$x \in D(T)$$

and

$$y \in H$$

33. <u>Space of Eigenvalue Spectrum</u>: The spectral measure $E$ is concentrated on the spectrum of $T$.
34. <u>Application to Unbounded Normal Operator</u>: There is also a version of the spectral theorem that applies to unbounded operators.

## References

- Bachman, G., L. Narici, and E. Beckenstein (2000): *Fourier and Wavelet Analysis* **Springer-Verlag** New York, NY
- Bers, L., F. John, and M. Schechter (1981): *Partial Differential Equations* **American Mathematical Society** Providence, RI
- Blanchet, G., and M. Charbit (2014): *Digital Signal and Image Processing using MATLAB* **Wiley** Hoboken, NJ
- Brenner, S., and R. L. Scott (2005): *The Mathematical Theory of Finite Element Methods 2$^{nd}$ Edition* **Springer** New York, NY

- Clarkson, J. A. (1936): Uniformly Convex Spaces *Transactions of American Mathematical Society* **40 (3)** 396-414

- Courant, R., and D. Hilbert (1953): *Methods of Mathematical Physics* **Wiley Interscience** Hoboken, NJ

- Dieudonne, J. (1960): *Foundations of Modern Analysis* **Academic Press** Cambridge, MA

- Dunford, N., and J. T. Schwartz (1958): *Linear Operators* **Wiley-Interscience** Hoboken, NJ

- Duren, P. (1970): *Theory of $H^P$-Spaces* **Academic Press** New York, NY

- Folland, G. B. (1989): *Harmonic Analysis in Phase Space* **Princeton University Press** Princeton, NJ

- Folland, G. B. (2009): *Fourier Analysis and its Application* **American Mathematical Society** Providence, RI

- Giusti, E. (2003): *Direct Methods in the Calculus of Variations* **World Scientific Publishing** Singapore

- Halmos, P. (1957): *Introduction to Hilbert Space and the Theory of Spectral Multiplicity* **Chelsea Publishing Company** New York, NY

- Halmos, P. (1957): *A Hilbert Space Problem Book* **Springer Verlag** New York, NY

- Hewitt, E., K. Stromberg (1965): *Real and Abstract Analysis* **Springer-Verlag** New York, NY

- Kac, M. (1966): Can on hear the Shape of the Drum? *American Mathematical Monthly* **73 (4)** 1-23

- Kadison, R. V., and J. R. Ringrose (1983): *Fundamentals of the Theory of Operator Algebras, Volume I Elementary Theory* **Academic Press** New York

- Kakutani, S. (1939): Some Characterizations of Euclidean Spaces *Japanese Journal of Mathematics* **16** 93-97

- Krantz, S. G. (2002): *Function Theory of Several Complex Variables* **American Mathematical Society** Providence, RI

- Lanczos, C. (1988): *Applied Analysis* **Dover Publications** Mineola, NY

- Lindenstrauss, J., and L. Tzafriri (1971): On the Complemented Subspace Problem *Israeli Journal of Mathematics* **9 (2)** 263-269

- Marsden, J. E. (1974): *Elementary Classical Analysis* **W. H. Freeman** New York, NY

- Nielsen, M. A., and I. L. Chuang (2000): *Quantum Computation and Quantum Information 1$^{st}$ Edition* **Cambridge University Press** Cambridge, UK

- Peres, A. (1993): *Quantum Theory: Concepts and Methods* **Kluwer** Amsterdam, Netherlands

- Prugovecki, E. (1981): *Quantum Mechanics in Hilbert Space 2$^{nd}$ Edition* **Dover** Mineola, NY

- Reed, M., and B. Simon (1975): *Fourier Analysis, Self-Adjointedness, Methods of Modern Mathematical Analysis* **Academic Press** New York, NY

- Reed, M., and B. Simon (1980): *Functional Analysis, Methods of Modern Mathematical Analysis* **Academic Press** New York, NY

- Rieffel, E. G., and W. H. Polak (2011): *Quantum Computing: A Gentle Introduction* **MIT Press** Cambridge, MA

- Riesz, F., and B. Sz-Nagy (1990): *Functional Analysis* **Dover** Mineola, NY

- Rudin, W. (1973): *Functional Analysis* **McGraw-Hill** New York, NY

- Rudin, W. (1987): *Real and Complex Analysis* **McGraw-Hill** New York, NY

- Schaffer, H. H., and W. P. Wolff (1999): *Topological Vector Spaces GTM Volume 8 2$^{nd}$ Edition* **Springer** New York, NY

- Schmidt, E. (1908): Über die Auflösung Linearer Gleichungen mit Unendlich Vielen Unbekannten *Rendiconti del Circolo Matematico di Palermo* **25** 63-77

- Shubin B. A. (1987): *Pseudo-differential Operators and Spectral Theory* **Springer** New York, NY

- Sobrino, L. (1996): *Elements of Non-relativistic Quantum Mechanics* **World Scientific Publishing** River Edge, NJ

- Stein, E. (1970): *Singular Integrals and Differentiability Properties of Functions* **Princeton University Press** Princeton, NJ

- Stein, E., and G. Weiss (1971): *Introduction to Fourier Analysis on Euclidean Spaces* **Princeton University Press** Princeton, NJ

- Streater, R., and A. Wightman (1964): *Pact, Spin, and Statistics and All That* **Addison Wesley** Boston, MA

- Treves, F. (1967): *Topological Vector Spaces, Distributions, and Kernels* **Academic Press** Cambridge, MA

- von Neumann, J. (1932): Physical Applications of the Ergodic Hypothesis *Proceedings of the National Academy of Sciences* **18 (3)** 263-266

- von Neumann, J. (1955): *Mathematical Foundations of Quantum Mechanics* **Princeton University Press** Princeton, NJ

- Warner, F. (1983): *Foundations of Differentiable Manifolds and Lie Groups* **Springer-Verlag** New York, NY

- Wikipedia (2022): [Hilbert Space](#)

- Young, N. (1988): *An Introduction to Hilbert Space* **Cambridge University Press** Cambridge, UK

# Positive-definite Kernel

## Overview

1. <u>Definition of a Positive-definite Kernel</u>: A *positive definite kernel* is a generalization of a positive-definite function or a positive-definite matrix (Wikipedia (2022)).
2. <u>Solutions to the Integral Operator Equations</u>: It was first introduced in the context of solving integral operator equations. Since then, positive-definite functions and their various analogues and generalizations have arisen in diverse parts of mathematics.
3. <u>Applications to the Positive-definite Kernel</u>: They occur naturally in Fourier analysis, probability theory, operator theory, complex function theory, moment problems, integral equations, boundary-value problems for partial differential equations, machine learning, embedding problem, information theory, and other areas.

## Definition

1. <u>Defining Symmetric Positive-definite Function</u>: Let $\mathcal{X}$ be a non-empty set, sometimes referred to as the index set. A symmetric function

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is called a positive-definite – p. d. – kernel on $\mathcal{X}$ if

$$\sum_{i=i}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

implies holds for any

$$x_1, \cdots, x_n \in \mathcal{X}$$

given

$$n \in \mathbb{N}$$

$$c_1, \cdots, c_n \in \mathbb{R}$$

2. <u>Positive Definite vs Semi-definite Kernels</u>: In probability theory, a distinction is sometimes made between positive-definite kernels, for which the equality in

$$\sum_{i=i}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

implies

$$c_i = 0 \ \forall \ i$$

and positive semi-definite – p. s. d. – kernels, which do not impose this condition. Note that this is equivalent to requiring that any finite matrix constructed by pairwise evaluation

$$K_{ij} = K(x_i, x_j)$$

has either entirely positive – p. d. – or non-negative – p. s. d. – eigenvalues.

3. <u>Chapter Focus – Real-valued Functions</u>: In mathematical literature, kernels are usually complex-valued functions, but this chapter assumes real-valued functions, which is the common practice in applications of positive definite kernels.

## Definition – Some General Properties

1. <u>Family of Positive Definite Kernels</u>: For a family of positive definite kernels $(K_i)_{i \in \mathbb{N}}$

$$K_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

the following hold.

2. <u>Canonical Sum</u>: The canonical sum

$$\sum_{i=1}^{n} \lambda_i K_i$$

is positive definite, given

$$\lambda_1, \cdots, \lambda_n \geq 0$$

3. <u>Product</u>: The product $K_1^{a_1} \cdots K_n^{a_n}$ is positive definite, given

$$a_1, \cdots, a_n \in \mathbb{N}$$

4. <u>Limit</u>: The limit

$$K = \lim_{n \to \infty} K_n$$

is positive definite if the limit exists.

5. <u>Positive-definite Kernels on Set Sequences</u>: If

$$(\mathcal{X}_i)_{i=1}^n$$

is a sequence of sets, and

$$(K_i)_{i=1}^n$$

a sequence of positive definite kernels

$$K_i : \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$$

then both

$$K\big((x_1, \cdots, x_n), (y_1, \cdots, y_n)\big) = \prod_{i=1}^n K_i(x_i, y_i)$$

and

$$K\big((x_1, \cdots, x_n), (y_1, \cdots, y_n)\big) = \sum_{i=1}^n K_i(x_i, y_i)$$

are positive definite kernels on

$$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$$

6. <u>Restriction of the Positive definite Kernel to a Subdomain</u>: Let

$$\mathcal{X}_0 \subset \mathcal{X}$$

Then the restriction of $K_0$ of $K$ to $\mathcal{X}_0 \times \mathcal{X}_0$ is also a positive definite kernel.

# Definition – Examples of Positive Definite Kernels

1. Examples of $\mathbb{R}^d$ Positive-definite Kernels: Common examples of positive-definite kernels defined on Euclidean space $\mathbb{R}^d$ include the following.

2. Linear Kernel:

$$K(x, y) = x^T y$$

$$x, y \in \mathbb{R}^d$$

3. Polynomial Kernel:

$$K(x, y) = (x^T y + r)^n$$

$$x, y \in \mathbb{R}^d$$

$$r \geq 0$$

$$n \geq 1$$

4. Gaussian Kernel - RBF:

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

$$x, y \in \mathbb{R}^d$$

$$\sigma > 0$$

5.  Laplacian Kernel:

$$K(x, y) = e^{-\alpha \|x - y\|}$$

$$x, y \in \mathbb{R}^d$$

$$\alpha > 0$$

6.  Abel Kernel:

$$K(x, y) = e^{-\alpha |x - y|}$$

$$x, y \in \mathbb{R}$$

$$\alpha > 0$$

7.  Kernel Generating Sobolev Spaces:

$$K(x, y) = \|x - y\|_2^{k - \frac{d}{2}} B_{k - \frac{d}{2}}(\|x - y\|_2)$$

where $B_v$ is the Bessel function of the third kind.

8.  Kernel Generating Paley-Wiener Space:

$$K(x, y) = sinc(\alpha[x - y])$$

$$x, y \in \mathbb{R}$$

$$\alpha > 0$$

9. Kernels from Hilbertian Inner Products: If $H$ is a Hilbert space, then its corresponding inner product

$$(\cdot,\cdot)_H : H \times H \to \mathbb{R}$$

is a positive definite kernel. Indeed, one has

$$\sum_{i,j=1}^{n} c_i c_j (x_i, x_j)_H = \left( \sum_{i=1}^{n} c_i x_i, \sum_{j=1}^{n} c_j x_j \right)_H = \left\| \sum_{i=1}^{n} c_i x_i \right\|^2 \geq 0$$

10. Kernels Defined on $\mathbb{R}_+^d$ and Histograms: Histograms are frequently encountered I applications of real-life problems. Most observations are usually available under the form of non-negative vectors of counts, which, if normalized, yield histograms of frequencies.

11. Distance Metrics are Kernel Parameters: It has been shown (Hein and Bousquet (2005)) that the following family of squared metrics, respectively Jensen divergence, the $\chi$-square, total variation, and two variations of the Hellinger distance.

12. Jensen Divergence:

$$\psi_{JD} = H \left( \frac{\theta + \theta'}{2} \right) - \frac{H(\theta) + H(\theta')}{2}$$

13. $\chi$-square:

$$\psi_{\chi^2} = \sum_i \frac{(\theta_i - \theta_i')^2}{\theta_i + \theta_i'}$$

$$\psi_{TV} = \sum_i |\theta_i - \theta_i'|$$

14. <u>Hellinger Distance Variants</u>:

$$\psi_{H_1} = \sum_i \left| \sqrt{\theta_i} - \sqrt{\theta_i'} \right|$$

$$\psi_{H_2} = \sum_i \left| \sqrt{\theta_i} - \sqrt{\theta_i'} \right|^2$$

15. <u>Kernels Generated over Distance Metrics</u>: The squared metrics can be used to define positive definite kernels using the formula

$$K(\theta, \theta') = e^{-\alpha(\theta - \theta')}$$

$$\alpha > 0$$

## Connection with Reproducing Kernel Hilbert Space and Feature Maps

1. <u>Positive-definite Kernels and Hilbert Spaces</u>: Positive-definite kernels provide a framework that encompasses some basic Hilbert space constructions. The following section presents a tight relationship between positive-definite kernels and two mathematical objects, namely reproducing Hilbert spaces and feature maps.

2. <u>Axiomatic Basis behind the Formulation</u>: Let $X$ be a set, $H$ a Hilbert space of functions

$$f : X \to \mathbb{R}$$

and

$$(\cdot, \cdot)_H : H \times H \to \mathbb{R}$$

322

be the corresponding inner product. For any

$$x \in X$$

the evaluation functional

$$e_x : H \to \mathbb{R}$$

is defined by

$$f \mapsto e_x(f) = f(x)$$

3. <u>Reproducing Kernel Hilbert Space Definition</u>: A reproducing Kernel Hilbert Space RHKS is defined as follows. $H$ is called a reproducing kernel Hilbert space if the evaluation functionals are continuous.

4. <u>Reproducing Kernel Associated with RKHS</u>: Every RKHS has a special function associated with it, namely the reproducing kernel.

5. <u>Reproducing Property of the PD Kernel</u>: Reproducing kernel is a function

$$K : X \times X \longrightarrow \mathbb{R}$$

such that
   a.

$$K_x(\cdot) \in H \ \forall \ x \in X$$

   and
   b.

$$(f, K_x) = f(x)$$

for all

$$f \in H$$

and

$$x \in X$$

The latter property is called the reproducing property.

6. <u>Equivalence between RKHS and Reproducing Kernel</u>: The following theorem shows the equivalence: Every reproducing kernel induces a unique RKHS, and every RKHS has a unique reproducing kernel.

7. <u>Connection between PD Kernels and RKHS</u>: The connection between positive definite kernels and RKHS is given by the following theorem: Every reproducing kernel is positive-definite, and every positive-definite kernel defines a unique RKHS, of which it is the unique reproducing kernel. Thus, given a positive-definite kernel $K$, it is possible to build an associated RKHS with $K$ as a reproducing kernel.

8. <u>PD Kernels and Feature Maps</u>: As stated earlier, positive definite kernels can be constructed from inner products. This fact can be used to connect positive definite kernels with another interesting object that arises in machine learning applications, namely the feature map.

9. <u>Implied Kernel Feature Map</u>: Let $F$ be a Hilbert space, and $(\cdot, \cdot)_F$ the corresponding inner product. Any map

$$\Phi : X \to F$$

is called a feature map. In this case $F$ is called a feature map. It is easy to see that every feature map defines a unique positive definite kernel by

$$K(x, y) = \big(\Phi(x), \Phi(y)\big)_F$$

10. <u>Positive Definiteness of the Kernel</u>: Indeed, positive definiteness of $K$ follows from the PD property of the inner product.

11. <u>RKHS Associated with Feature Maps</u>: On the other hand, every p. d. kernel, and its corresponding RKHS, have many associated feature maps. For example, let

$$F = H$$

and

$$\Phi(x) = K_x$$

for all

$$x \in X$$

Then

$$\big(\Phi(x), \Phi(y)\big)_F = \big(K_x, K_y\big)_H = K(x, y)$$

by the reproducing property.

12. <u>Constructing the Corresponding RKHS</u>: This suggests a new look at positive definite kernels as inner products in appropriate Hilbert spaces, or in other words, positive definite kernels can be viewed as similarity maps which quantify effectively how similar two points $x$ and $y$ are through the value $K(x, y)$. Moreover, through the equivalence of positive definite kernels and its corresponding RKHS, every feature map can be used to construct a RKHS.

## Kernels and Distances

1. <u>Kernel Methods vs. Nearest Neighbors</u>: Kernel methods are often compared to distance-based methods such as nearest neighbors. This section discusses the parallels between their two respective ingredients, namely kernels $K$ and distance $d$.

2. <u>Definition of the Distance Metric</u>: Here, by a distance function between each pair of elements of some set $X$, one means a metric defined on that set, i.e., any non-negative valued function $d$ on $\mathcal{X} \times \mathcal{X}$ which satisfies

   a.

$$d(x, y) \geq 0$$

   and

$$d(x, y) = 0$$

   if and only if

$$x = y$$

   b.

$$d(x, y) = d(y, z)$$

   c.

$$d(x, z) \leq d(x, y) + d(y, z)$$

3. <u>Motivation behind Negative Definite Kernel</u>: One link between distances and p. d. kernels is given by a particular kind of kernel, called a negative definite kernel, and is defined as follows.

4. <u>Defining Negative Definite Kernel</u>: A symmetric function

$$\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is called a negative definite – n. d. – kernel on $\mathcal{X}$ if

$$\sum_{i,j=1}^{n} c_i c_j \psi(x_i, x_j) \leq 0$$

holds for any

$$n \in \mathbb{N}$$

$$x_1, \cdots, x_n \in \mathcal{X}$$

and

$$c_1, \cdots, c_n \in \mathbb{R}$$

such that

$$\sum_{i=1}^{n} c_i = 0$$

5. <u>Parallel between Negative-definite Kernels and Distances</u>: This parallel is in the following: whenever a negative-definite kernel vanishes on the set

$$\{(x, x) : x \in \mathcal{X}\}$$

and is zero only on this set, then its square root is a distance for $\mathcal{X}$.

6. <u>Definition of the Hilbertian Distance</u>: At the same time each distance does not correspond necessarily to a negative definite kernel. This is true only for Hilbertian distances, where the distance $d$ is called Hilbertian if one can embed the metric space $(\mathcal{X}, d)$ isometrically into some Hilbert space.

7. <u>Defining Infinitely Divisible Kernels</u>: On the other hand, negative-definite kernels can be identified with a subfamily of positive-definite kernels known as infinitely divisible kernels. The non-negative kernel $K$ is said to be infinitely divisible if for every

$$n \in \mathbb{N}$$

there exists a positive-definite kernel $K_n$ such that

$$K = (K_n)^n$$

8. <u>Definition of the Distance Pseudo-metric</u>: Another link is that a positive definite kernel induces a pseudo-metric, where the first constraint on the distance function is loosened to allow

$$d(x, y) = 0$$

for

$$x \neq y$$

9. <u>Definition of the Distance Function</u>: Given a positive-definite kernel $K$, one can define a distance function as

$$d(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}$$

## Applications to Kernels in Machine Learning

1. <u>Importance of Positive-definite Kernels</u>: Positive-definite kernels, through their equivalence with reproducing kernel Hilbert spaces, are particularly important in the field of statistical learning theory because of the celebrated representer theorem which states that every minimizer function in an RKHS can be written as a linear combination of the kernel function evaluated at the training points.

2. <u>Simplification of Empirical Risk Minimization</u>: This is a practically useful result as it effectively simplifies the empirical risk minimization problem from an infinite dimensional to a finite dimensional optimization problem.

## Applying Kernels in Probabilistic Models

1. <u>Non-deterministic Recovery Problems</u>: Assume that one wants to find the response $f(x)$ of an unknown model function at a new point $x$ of a set $\mathcal{X}$, provided that one has a sample of input-response pairs

$$(x_i, f_i) = \left(x_i, f(x_i)\right)$$

given by observation or experiment.

2. <u>Stochastic Responses to Predictors</u>: The response $f_i$ at $x_i$ is not a fixed function of $x_i$ but rather a realization of a real-valued random variable $Z(x_i)$. The goal is to get information about the function $\mathbb{E}[Z(x_i)]$ which replaces $f$ in the deterministic setting.

3. <u>Proximity of the Stochastic Responses</u>: For two elements

$$x, y \in \mathcal{X}$$

the random variables $Z(x)$ and $Z(y)$ will not be uncorrelated, because if $x$ is too close to $y$ the random experiments described by $Z(x)$ and $Z(y)$ will often show similar behavior.

4. <u>Covariance between the Stochastic Responses</u>: This is defined by a covariance kernel

$$K(x, y) = \mathbb{E}[Z(x) \cdot Z(y)]$$

Such a kernel exists and is positive-definite under weak additional assumptions.

5. <u>Interpolation using the Covariance Kernel</u>: A good estimate for $Z(x)$ can be obtained by using kernel interpolation with the covariance kernel, ignoring the probabilistic background completely.

6. <u>Incorporation of a Stochastic Noise</u>: Assuming that a noise variable $\epsilon(x)$, with a zero mean and variance $\sigma^2$, is added to $x$, such that the noise is independent for different $x$ and independent of $Z$ there, then the problem of finding a good estimate for $f$ is identical to the one above, but with a modified kernel given by

$$K(x, y) = \mathbb{E}[Z(x) \cdot Z(y)] + \sigma^2 \delta_{xy}$$

7. <u>Density Estimation by Kernels</u>: The problem is to recover the density $f$ of a multivariate distribution over a domain $\mathcal{X}$, from a large sample

$$x_1, \cdots, x_n \in \mathcal{X}$$

including repetitions. When sampling points lie dense, the true identity must take large values.

8. <u>Histogram-based Sample Density Estimation</u>: A simple density estimate is possible by counting the number of samples in each cell of a grid, and plotting the resulting histogram, which yields a piecewise constant density estimate.

9. Kernel-based Sample Density Estimate: A better estimate can be obtained by using a non-negative translation invariant kernel $K$, with total integral equal to one, and define

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

as a smooth estimate.

## Application to Numerical Solution of Partial Differential Equations

1. Numerical Solution of PDEs: One of the greatest application areas of the so-called meshfree methods is in the numerical solution of PDEs.
2. MLPG, RKPM, and SPH: Some of the popular meshfree methods are closely related to positive-definite kernels, such as meshless local Petrov Galerkin MLPG, reproducing kernel particle method RKPM, and smoothed particle hydrodynamics SPH. These methods use radial basis kernel for collocation (Schaback and Wendland (2006)).

## Other Applications

1. The Response Surface Methodology: In literature on computer experiments (Haaland and Qian (2012)) and other engineering experiments one increasingly encounters models on positive definite kernels, RBFs, or kriging. One such topic response surface methodology.
2. Estimation using Implicit Surface Models: Other types of applications that boil down of fitting data are rapid prototyping and computer graphics. Here one often uses implicit surface models to approximate or interpolate point cloud data.

3. <u>Applications of Positive Definite Kernels</u>: Applications in various other branches of mathematics are in multivariate integration, multivariate optimization, and in numerical analysis and scientific computing, where one studies fast, accurate, and adaptive algorithms ideally implemented in high-performance computing environments (Gumerov and Duraiswami (2005)).

## Reference

- Gumerov, N. A., and R. Duraiswami (2005): Fast Radial Basis Function Interpolation via Pre-conditioned Krylov Iteration *SIAM Journal on Scientific Computing* **29 (5)** 1876-1899
- Haaland, B., P. Z. G. Qian (2012): Accurate Emulators for Large-scale Computer Experiments **arXiV**
- Hein, M., and O. Bousquet (2005): [Hilbertian Metrics and Positive Definite Kernels on Probability Measures](#)
- Schaback, R., and H. Wendland (2006): Kernel Techniques: From Machine Learning to Meshless Methods *Acta Numerica* **15** 543-639
- Wikipedia (2022): [Positive-definite Kernel](#)

# Reproducing Kernel Hilbert Space

## Overview

1. Reproducing Kernel Hilbert Space – Definition: A *reproducing kernel Hilbert Space* (RKHS) is a Hilbert space of functions in which point evaluation is a continuous linear functional.

2. Closeness in Terms of Norm: Roughly speaking, this means that if two functions $f$ and $g$ in the RKHS are close in norm, i.e., $\|f - g\|$ is small, then $f$ and $g$ are also point-wise close, i.e., $|f(x) - g(x)|$ is small for all $x$. The converse need not be true (Wikipedia (2022)).

3. Building Hilbert Spaces of Functions: It is not entirely straightforward to construct a Hilbert space of functions which is not an RKHS (Alpay and Mills (2003)). Some examples, however, have been found (Pasternak-Winiarski (1992), Zynda (2020)).

4. Equivalence Classes of Hilbert Spaces: Note that $L^2$ spaces are not Hilbert spaces of functions – and hence not RKHS's – but rather Hilbert spaces of equivalence classes of functions; for example, the functions $f$ and $g$ defined by

$$f(x) = 0$$

and

$$g(x) = \mathbb{I}_\mathbb{Q}$$

are equivalent in $L^2$.

5. Space of Band limited Functions: However, there are RKHS's in which the norm is an $L^2$-norm, such as the band-limited functions – as shown in an example below.

6.  <u>Reproducing Kernels used in RKHS</u>: An RKHS is associated with a kernel that reproduces every function in the space in the sense that for every $x$ in the set on which the functions are defined, "evaluation at $x$" can be performed by taking an inner product with a function defined by the kernel.

7.  <u>Existence of a Reproducing Kernel</u>: Such a *reproducing kernel* exits if and only if every evaluation functional is continuous.

8.  <u>Use in Harmonic/Biharmonic Functions</u>: The reproducing kernel was first introduced in the 1907 work of Stanislaw Zaremba concerning boundary value problems for harmonic and biharmonic functions.

9.  <u>Use in Integral Equation Theory</u>: James Mercer simultaneously examined functions which satisfy the reproducing property in the theory of integral equations.

10. <u>Usage in Other Areas</u>: These spaces have wide applications, including complex analysis, harmonic analysis, and quantum mechanics.

11. <u>Use in Statistical Learning Theory</u>: Reproducing kernel Hilbert spaces are particularly important in the field of statistical learning theory because of the celebrated representer theorem which states that every function on an RKHS that minimizes an empirical risk functional can be written as a linear combination of the kernel function evaluated at the training points.

12. <u>Reduction to Finite Dimensional Problem</u>: This is a practically useful result as it effectively simplifies the empirical risk minimization problem from an infinite dimensional to a finite dimensional optimization problem.

13. <u>Real-valued Hilbert Spaces Treatment</u>: For ease of understanding, this chapter provides the framework for real-values Hilbert spaces. The theory can be easily extended to spaces of complex-valued functions and hence include many important examples of reproducing kernel Hilbert spaces that are spaces of analytic functions (Paulsen (2009)).

## Definition

1. Pointwise Additive/Multiplicative Hilbert Spaces: Let $X$ be an arbitrary set and $H$ a Hilbert space of real-valued functions on $X$, equipped with pointwise addition and pointwise scalar multiplication.

2. Evaluation Functional over Hilbert Space: The evaluation functional over the Hilbert spaces of functions $H$ is a linear functional that evaluates each function at a point $x$

$$L_x : f \longmapsto f(x) \ \forall \ f \in H$$

3. Formal Definition of the RKHS: $H$ is said to be a *reproducing kernel Hilbert space* if, for all $x$ in $X$, $L_x$ is continuous at every $f$ in $H$ or, equivalently, if $L_x$ is a bounded operator on $H$, i.e., there exists some

$$M_x > 0$$

such that

$$|L_x(f)| := |f(x)| \leq M_x \|f\|_H \ \forall \ f \in H$$

4. Bounded in $X$; Unbounded Overall: Although

$$M_x < \infty$$

is assumed for all

$$x \in X$$

it might still be the case that

$$\underset{x}{\sup} \ M_x = \infty$$

5. <u>Weak Nature of RKHS Definition</u>: While the property

$$|L_x(f)| := |f(x)| \le M_x \|f\|_H \ \forall \ f \in H$$

is the weakest condition that ensures both existence of an inner product and the evaluation of every function in $H$ at every point in the domain, it does not tend itself to easy application in practice.

6. <u>Alternative Representation of Reproducing Kernel</u>: A more intuitive property guarantees that the evaluation functional can be represented by taking the inner product $f$ with a function $K_x$ in $H$. This function is the so-called *reproducing kernel* for the Hilbert space $H$ from which the RKHS takes its name.

7. <u>Application of Riesz Representation Theorem</u>: More formally, the Riesz representation theorem implies that for all $x$ in $X$ there exists a unique element $K_x$ of $H$ with the reproducing

$$f(x) = L_x(f) = \langle f, K_x \rangle_H \ \forall \ f \in H$$

8. <u>Dot-Product of $K$ with itself</u>: Since $K_x$ is itself a function defined on $X$ with values in the field $\mathbb{R}$ - or $\mathbb{C}$ in the case of complex Hilbert spaces and as $K_x$ is in $H$ we have that

$$K_x(y) = L_y(K_x) = \langle K_x, K_y \rangle_H$$

where

$$K_y \in H$$

is the element in $H$ associated with $L_y$.

9. <u>Reproducing Kernel of Hilbert Spaces</u>: This allows one to define the reproducing kernel of $H$ as a function

$$K : X \times X \to \mathbb{R}$$

by

$$K(x, y) = \langle K_x, K_y \rangle_H$$

10. <u>Property of the Reproducing Kernel</u>: From tis definition it is easy to see that

$$K : X \times X \to \mathbb{R}$$

- or $\mathbb{C}$ in the complex case – is both symmetric, respectively conjugate symmetric, and positive definite, i.e.

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n} c_i \langle K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j} \rangle_H = \langle \sum_{i=1}^{n} c_i K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j} \rangle_H$$

$$= \left\| \sum_{j=1}^{n} c_j K_{x_j} \right\|_H^2 \geq 0$$

for every

$$n \in \mathbb{N}$$

$$x_1, \cdots, x_n \in X$$

and

$$c_1, \cdots, c_n \in \mathbb{R}$$

11. The Converse - Moore-Aronszajn Theorem: The Moore-Aronszajn theorem – as is shown below – is a sort of converse to this: if a function $K$ satisfies these conditions, then there is a Hilbert space of functions on $X$ for which it is a reproducing kernel.

## Example

1. Space of Band-limited Continuous Function: The space of band-limited continuous functions $H$ is a RKHS, as shown below.

2. Set of Continuous Square-integrable Functions: Formally, fix some cutoff frequency

$$0 < a < \infty$$

and define the Hilbert space

$$H = \{f \in C(\mathbb{R}) \mid supp\ (F) \subset [-a, a]\}$$

where $C(\mathbb{R})$ is the set of continuous square integrable functions, and

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-i\omega t} dt$$

is the Fourier transform of $f$.

3. Inner Product of Hilbert Space: As the inner product of this Hilbert space, one uses

$$\langle f, g \rangle_{L^2} = \int_{-\infty}^{+\infty} f(x) \cdot \overline{g(x)}\ dx$$

4. Using the Fourier Inversion Theorem: From the Fourier Inversion theorem, one has

$$f(x) = \frac{1}{2\pi} \int\limits_{-a}^{+a} F(\omega)\, e^{i\omega x} d\omega$$

5.  $f(x)$ in Terms of $\|f\|_{L^2}$: It then follows by the Cauchy-Schwarz inequality and Plancherel's theorem that, for all $x$

$$|f(x)| \leq \sqrt{\frac{1}{2\pi} \int\limits_{-a}^{+a} 2a|F(\omega)|^2 d\omega} = \frac{1}{\pi}\sqrt{\frac{a}{2} \int\limits_{-a}^{+a} 2a|F(\omega)|^2 d\omega} = \sqrt{\frac{a}{\pi}}\,\|f\|_{L^2}$$

6.  Proof that $f(x)$ is Bounded: This inequality shows that the evaluation functional is bounded, proving that $H$ is indeed a RKHS.

7.  Kernel Function for Laplace Transform: The kernel function $K_x$ in this case is given by

$$K_x(y) = \frac{a}{\pi}\,sinc\left(\frac{a}{\pi}[y-x]\right) = \frac{\sin(\pi[y-x])}{\pi[y-x]}$$

8.  Fourier Transform of $K_x(y)$: To see this, one first notes that the Fourier transform of $K_x(y)$ defined above is given by

$$\int\limits_{-\infty}^{+\infty} K_x(y)e^{-i\omega y} dy = \begin{cases} e^{-i\omega x} & if\ \omega \in [-a, +a] \\ 0 & otherwise \end{cases}$$

which is a consequence of the time-shifting property of the Fourier transform.

9.  $f(x)$ – Kernel Dot Product: Consequently, using the Plancherel's theorem, one has

$$\langle f, K_x \rangle_{L^2} = \int_{-\infty}^{+\infty} f(y) \cdot \overline{K_x(y)} dy = \frac{1}{2\pi} \int_{-a}^{+a} F(\omega) \cdot e^{i\omega x} dx = f(x)$$

Thus, one obtains the reproducing property of the kernel.

10. <u>Band-limited Version of Dirac Delta</u>: Note that $K_x$ in this case is the "band-limited version" of the Dirac delta function, and that $K_x(y)$ converges to $\delta(y - x)$ in the weak sense as the cutoff frequency a tends to infinity.

## Moore-Aronszajn Theorem

1. <u>RKHS from a PSD Kernel</u>: It was seen above how a reproducing kernel Hilbert space defines a reproducing kernel function that is both symmetric and positive-definite.

2. <u>Objective of Moore-Aronszajn Theorem</u>: The Moor-Aronszajn theorem goes in the other direction; it states that every symmetric, positive-definite kernel defines a unique reproducing kernel Hilbert space.

3. <u>Statement of Moore-Aronszajn Theorem</u>: Suppose $K$ is a symmetric, positive-definite kernel on a set $X$. Then there is a unique Hilbert space of functions on $X$ for which $K$ is a reproducing kernel.

4. <u>Axiomatization behind Moore-Aronszajn Theorem</u>: For all $x$ in $X$, define

$$K_x = K(x, \cdot)$$

Let $H_0$ be the linear span of

$$\{K_x : x \in X\}$$

5. <u>Defining Inner Product on $H_0$</u>: Define an inner product on $H_0$ by

$$\langle \sum_{j=1}^{n} b_j K_{y_j}, \sum_{i=1}^{m} a_i K_{x_i} \rangle_{H_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K(y_j, x_i)$$

which implies that

$$K(x,y) = \langle K_x, K_y \rangle_{H_0}$$

6. <u>Symmetry of the Inner Product</u>: The symmetry of this inner product follows from the symmetry of $K$ and the non-degeneracy follows from the fact that $K$ is positive definite.

7. <u>Completion of Inner Product Space</u>: Let $H$ be the completion of $H_0$ with respect to this inner product. Then $H$ consists of the functions of the form

$$f(x) = \sum_{i=1}^{\infty} a_i K_{x_i}(x)$$

where

$$\lim_{n \to \infty} \sup_{p \geq 0} \left\| \sum_{i=n}^{n+p} a_i K(x_i) \right\| = 0$$

8. <u>Reproducing Property of Complete Space</u>: Now, the reproducing property

$$f(x) = L_x(f) = \langle f, K_x \rangle_H \ \forall \ f \in H$$

may be verified.

$$\langle f, K_x \rangle_H = \sum_{i=1}^{\infty} a_i \langle K_{x_i}, K_x \rangle_{H_0} = \sum_{i=1}^{\infty} a_i K(x_i, x) = f(x)$$

9. <u>Uniqueness of Hilbert Space #1</u>: To prove uniqueness, let $G$ be another Hilbert space of functions for which $K$ is a reproducing kernel. For every $x$ and $y$ in $X$

$$f(x) = L_x(f) = \langle f, K_x \rangle_H \; \forall f \in H$$

implies that

$$\langle K_x, K_y \rangle_H = K(x,y) = \langle K_x, K_y \rangle_G$$

10. <u>Uniqueness of Hilbert Space #2</u>: By linearity

$$\langle \cdot, \cdot \rangle_H = \langle \cdot, \cdot \rangle_G$$

on the span of

$$\{K_x : x \in X\}$$

Then

$$H \subset G$$

because $G$ is complete and contains $H_0$ and hence contains its completion.

11. <u>Uniqueness of Hilbert Space #3</u>: Now, one needs to prove that every element of $G$ is in $H$.

12. <u>Uniqueness of Hilbert Space #4</u>: Let $f$ be an element of $G$. Since $H$ is a closed subspace of $G$, one can write

$$f = f_H + f_{H_\perp}$$

where

$$f_H \in H$$

and

$$f_{H_\perp} \in H_\perp$$

13. <u>Uniqueness of Hilbert Space #5</u>: Now, if

$$x \in X$$

then, since $K$ is a reproducing kernel of $G$ and $H$:

$$f(x) = \langle K_x, f \rangle_G = \langle K_x, f_H \rangle_G + \langle K_x, f_{H_\perp} \rangle_G = \langle K_x, f_H \rangle_H = f_H(x)$$

where one has used the fact that $K_x$ belongs to $H$ so that its inner product with $f_{H_\perp}$ in $G$ is zero.

14. <u>Uniqueness of Hilbert Space #6</u>: This shows that

$$f = f_H$$

in $G$ and concludes the proof.

## Integral Operators and Mercer's Theorem

1. <u>Integral Operator using Mercer's Theorem</u>: One may characterize a symmetric positive definite kernel $K$ via the integral operator using Mercer's theorem and obtain an additional view of the RKHS.

2. <u>Compact, Borel Space with PSD Kernel</u>: Let $X$ be a compact space equipped with a strictly positive finite Borel measure $\mu$ and

$$K : X \times X \to \mathbb{R}$$

a continuous, symmetric, and positive-definite function.

3. <u>Defining the Integral Operator</u>: Define the integral operator

$$T_K : L_2(X) \to L_2(X)$$

as

$$[T_K f](\cdot) = \int_X^X K(\cdot, t) f(t) d\mu(t)$$

where $L_2(X)$ is the space of square integrable functions with respect to $\mu$.

4. <u>Statement of the Mercer's Theorem</u>: Mercer's theorem states that the spectral decomposition of the integral operator $T_K$ of $K$ yields a series representation of $K$ in terms of the eigenvalues and the eigenfunctions of $T_K$.

5. <u>Implication - $K$ is a Reproducing Kernel</u>: This then implies that $K$ is a reproducing kernel so that the corresponding RKHS can be defined in terms of these eigenvalues and eigenfunctions.

6. <u>$T_K$ – Compact, Positive, Self-adjoint, Continuous</u>: Under these assumptions, $T_K$ is a compact, continuous, self-adjoint, and positive operator.

7. <u>Spectral Theorem for Self-adjoint Operators</u>: The spectral theorem for self-adjoint operators implies that there is at most a countable decreasing sequence

$$(\sigma_i)_i \geq 0$$

such that

$$\lim_{i \to \infty} \sigma_i = 0$$

and

$$T_K \varphi_i(x) = \sigma_i \varphi_i(x)$$

where the $\{\varphi_i\}$ form an orthonormal basis of $L_2(X)$.

8. <u>Choosing Eigenvectors as Continuous Functions</u>: By positivity of $T_K$

$$\sigma_i > 0$$

for all $i$. One can also show that $T_K$ maps continuously into the space of continuous functions $C(X)$ and therefore one may choose continuous functions as the eigenvectors, that is

$$\varphi_i \in C(X)$$

for all $i$.

9. <u>Expressing $K$ using Eigenvalues/Eigenvectors</u>: Then by Mercer's theorem $K$ may be written in terms of the eigenvalues and continuous eigenfunctions as

$$K(x, y) = \sum_{j=1}^{\infty} \sigma_j \varphi_j(x) \varphi_j(y)$$

for all

$$x, y \in X$$

such that

$$\lim_{n \to \infty} \sup_{u, v} \left| K(u, v) - \sum_{j=1}^{\infty} \sigma_j \varphi_j(u) \varphi_j(v) \right| = 0$$

This above series is referred to as a Mercer kernel or Mercer representation of $K$.

10. <u>RKHS of the Reproducing Kernel</u>: Furthermore, it can be shown that the RKHS $H$ of $K$ is given by

$$H = \left\{ f \in L_2(X) \mid \sum_{i=1}^{\infty} \frac{\langle f, \varphi_i \rangle_{L_2}^2}{\sigma_i} < \infty \right\}$$

where the inner product of $H$ is given by

$$\langle f, g \rangle_H = \sum_{i=1}^{\infty} \frac{\langle f, \varphi_i \rangle_{L_2} \langle g, \varphi_i \rangle_{L_2}}{\sigma_i}$$

11. <u>Usage of the Mercer Kernel</u>: This representation of RKHS has applications in probability and statistics, for example, to the Karhunen-Loeve representation of stochastic processes and kernel PCA.

## Feature Maps

1. <u>Definition of a Feature Map</u>: A *feature map* is the map

$$\varphi : X \to F$$

where $F$ is a Hilbert space which will be called the feature space.

2. <u>RKHS Representation using Feature Maps</u>: The first sections presented the connection between bounded/continuous evaluation functions, positive definite functions, and integral operators, and this section provides another representation of the RKHS in terms of feature maps.

3. <u>Setting the Feature Map Kernel</u>: One first notes that every feature map defines a kernel via

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_F$$

4. <u>Resulting Symmetric and PSD Kernel</u>: Clearly $K$ is symmetric and the positive-definiteness follows from the properties of inner product in $F$. Conversely, every positive definite function and the corresponding reproducing kernel Hilbert space has infinitely many associated feature maps such that the above $K(x, y)$ holds.

5. <u>A Trivial Feature Map</u>: For example, one can trivially set

$$F = H$$

and

$$\varphi(x) = K_x$$

for all

$$x \in X$$

Then

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_F$$

is satisfied by the reproducing property.

6. <u>Feature Map from Integral Operators</u>: Another example of a feature map relates to previous section regarding integral operators by taking

$$F = l^2$$

and

$$\varphi(x) = \left( \sqrt{\sigma_i} \varphi_i(x) \right)_i$$

7. <u>Connection between Kernels and Feature Maps</u>: This connection between kernels and feature maps provides a new way to understand positive definite functions and hence reproducing kernels as inner products in $H$. Moreover, every feature map can naturally define a RKHS by means of the definite of a positive definite function.

8. <u>Function Spaces of Feature Maps</u>: Lastly, feature maps allow the construction of function spaces that reveal another perspective on the RKHS. Consider the linear space

$$H_\varphi = \{ f : X \to \mathbb{R} \mid \exists\, \omega \in F, f(x) = \langle \omega, \varphi(x) \rangle_F, \forall\, x \in F \}$$

9. <u>Norm of this Function Space</u>: One can define a norm on $H_\varphi$ by

$$\|f\|_\varphi = inf\{ \|\omega\|_F :\ \omega \in F, f(x) = \langle \omega, \varphi(x) \rangle_F, \forall\, x \in F \}$$

10. <u>RKHS Kernel of Function Spaces</u>: It can be shown that $H_\varphi$ is a RKHS with kernel defined

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_F$$

11. Inner Products on Feature Spaces: This representation that the elements of the RKHS are inner products of elements in the feature space and can accordingly be seen as hyperplanes. This view of the RKHS is related to the kernel trick in machine learning.

## Properties

1. Kernel in the $\mathbb{C}^p$ Space: Let

$$(X_i)_{i=1}^p$$

be a sequence of sets and

$$(K_i)_{i=1}^p$$

be a collection of positive definite functions on

$$(X_i)_{i=1}^p$$

It then follows that

$$K(\{x_1, \cdots, x_p\}, \{y_1, \cdots, y_p\}) = K_1(x_1, y_1) \cdots K_p(x_p, y_p)$$

is a kernel on

$$X = X_1 \times \cdots \times X_p$$

2. Subspace Restriction of a Kernel: Let

$$X_0 \subset X$$

Then the restriction of $K$ to $X_0 \times X_0$ is also a reproducing kernel.

3. <u>Case of a Normalized Kernel</u>: Consider a normalized kernel $K$ such that

$$K(x, x) = 1$$

for all

$$x \in X$$

4. <u>Distance Pseudo-metric on $X$</u>: Define a pseudo-metric on $X$ as

$$d_K(x, y) = \left\| K_x - K_y \right\|_H^2 = 2[1 - K(x, y)] \ \forall \ x \in X$$

By Cauchy-Schwartz inequality

$$K(x, y)^2 \le K(x, x)K(y, y) = 1 \ \forall \ x, y \in X$$

5. <u>$K$ as a Similarity Metric</u>: This inequality allows one to view $K$ as a measure of similarity between inputs. If

$$x, y \in X$$

are similar then $K(x, y)$ will be closer to 1 while if

$$x, y \in X$$

are dissimilar then $K(x, y)$ will be closer to 0.

6. <u>Closure of the Hilbert Space</u>: The closure of the span of

$$\{K_x \mid x \in X\}$$

coincides with $H$.

## Common Examples – Bilinear Kernels

$$K(x, y) = \langle x, y \rangle$$

The RKHS $H$ corresponding to this kernel is the dual space consisting of functions

$$f(x) = \langle x, \beta \rangle$$

satisfying

$$\|f\|_H^2 = \|\beta\|^2$$

## Common Examples – Polynomial Kernels

$$K(x, y) = (\alpha \langle x, y \rangle + 1)^d$$

$$\alpha \in \mathbb{R}^d$$

$$d \in \mathbb{N}$$

## Common Examples – Radial Basis Function

These are another common class of kernels that satisfy

$$K(x, y) = K(\|x - y\|)$$

## Common Examples – Radial Basis Function – Gaussian or Squared Exponential Kernel

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

$$\sigma > 0$$

## Common Examples – Radial Basis Function – Laplacian Kernel

$$K(x, y) = e^{-\frac{\|x-y\|}{\sigma}}$$

$$\sigma > 0$$

The squared norm of a function $f$ in the RKHS $H$ with this kernel is (Berlinet and Thomas (2004)):

$$\|f\|_H^2 = \int f^2(x)dx + \int f'(x)^2 dx$$

## Bergman Kernels

1. <u>Setup behind the Bergman Kernel</u>: Let $X$ be finite and $H$ consist of all complex values functions on $X$. Then an element of $H$ can be represented as an array of complex numbers.

2. <u>As a Complex Identity Function</u>: If the usual inner product is used, then $K_x$ is the function whose value is 1 at $x$ and 0 everywhere else, and $K(x, y)$ can be thought of as an identity matrix since

$$K(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

In this case, $H$ is isomorphic in $\mathbb{C}^n$.

3. <u>Kernel on a Unit Disc</u>: The case of

$$X = \mathbb{D}$$

where $\mathbb{D}$ denotes the unit disc, is more sophisticated. Here, the Bergman space $H^2(\mathbb{D})$ is the space of square-integrable holomorphic functions on $\mathbb{D}$.

4. <u>The Corresponding Reproducing Kernel</u>: It can be shown that the corresponding reproducing kernel for $H^2(\mathbb{D})$ is

$$K(x, y) = \frac{1}{\pi} \frac{1}{(1 - x\bar{y})^2}$$

5. <u>Kernel for Band-limited Functions</u>: Lastly, the space of band-limited functions in $L^2(\mathbb{R})$ with a bandwidth $2a$ is a RKHS with a reproducing kernel

$$K(x, y) = \frac{\sin(a[x - y])}{\pi[x - y]}$$

**Extension to Vector-valued Functions**

1. <u>Motivation behind Vector-valued RKHS</u>: This section extends the definition of RKHS to spaces of vector-valued functions as this is particularly important in multi-task learning and manifold regularization.
2. <u>Kernel as a PSD Matrix</u>: The main difference is that the reproducing kernel $\Gamma$ is a symmetric function that is now a positive semi-definite *matrix* for every $x, y$ in $X$.
3. <u>Formal Definition of the Kernel</u>: More formally, one defines a vector-valued kernel vvRKHS as a Hilbert space of functions

$$f : X \to \mathbb{R}^T$$

such that for all

$$c \in \mathbb{R}^T$$

and

$$x \in X$$

$$\Gamma_x c(y) = \Gamma(x, y)c \in H$$

for

$$y \in X$$

and

$$\langle f, \Gamma_x c \rangle_H = f(x)^T c$$

4.  <u>Contrast to Scalar-valued Case</u>: This second property parallels the reproducing property for scalar-valued case.

5.  <u>Equivalence with other Scalar Kernels</u>: One notes that this definition can also be connected to integral operators, bounded evaluation functions, and feature maps as seen earlier for the scalar valued RKHS.

6.  <u>vvRKHS as a Vector-valued Space</u>: One can alternatively define vvRKHS as a vector-valued Hilbert space with a bounded evaluation functional and show that this implies the existence of a unique reproducing kernel by the Riesz representation theorem.

7.  <u>Mercer's Theorem in Vector-valued Setting</u>: Mercer's theorem can also be extended to address the vector-valued setting and one can therefore obtain a feature-map view of RKHS.

8.  <u>Closure of the Vector-valued Span</u>: Lastly, it can also be shown that the closure of the span

$$\{\Gamma_x c : x \in X, c \in \mathbb{R}^T\}$$

coincides with $H$, another property similar to the scalar-valued case.

9.  <u>Isometric Isomorphism with Scalar RKHS</u>: One can gain intuition for the vvRKHS by taking a component-wise perspective on these spaces. In particular, one finds that every vvRKHS is isometrically isomorphic to a scalar-valued RKHS on a particular input space.

10. <u>Reproducing Scalar-valued Kernel</u>: Let

$$\Lambda = \{1, \cdots, T\}$$

Consider the space $X \times \Lambda$ and the corresponding reproducing kernel

$$\gamma : X \times \Lambda \times X \times \Lambda \to \mathbb{R}$$

11. <u>Corresponding Scalar-valued RKHS</u>: As noted above, the RKHS associated with this reproducing kernel is given by the closure of the span of

$$\{\gamma_{(x,t)} : x \in X, t \in \Lambda\}$$

where

$$\gamma_{(x,t)}(y, s) = \gamma\big((x,t),(y,s)\big)$$

for every set of pairs

$$(x,t),(y,s) \in X \times \Lambda$$

12. <u>Connecting Vector-to-scalar-valued Kernel</u>: The connection to scalar-valued RKHS can then be made by the fact that every matrix-valued kernel can be identified with a kernel of

$$\gamma : X \times \Lambda \times X \times \Lambda \to \mathbb{R}$$

via

$$\Gamma(x,y)_{(t,s)} = \gamma\big((x,t),(y,s)\big)$$

13. <u>Extension to Matrix-valued Kernels</u>: Moreover, every kernel with the form of

$$\gamma : X \times \Lambda \times X \times \Lambda \to \mathbb{R}$$

defines a matrix-valued kernel with the above expression. Now, letting the map

$$D : H_\Gamma \to H_\gamma$$

356

be defined as

$$(Df)(x,t) = \langle f(x), e_t \rangle_{\mathbb{R}^T}$$

where $e_t$ is the $t^{th}$ component of the canonical basis for $\mathbb{R}^T$, one can show that $D$ is bijective and is an isometry between $H_\Gamma$ and $H_\gamma$.

14. Complexities with Isometrized Scalar Reduction: While this view of vvRKHS can be useful in multi-task learning, this isometry does not reduce the vector-valued case to that of the scalar-valued case. In fact, this isometry procedure can make both the scalar-valued kernel and the input space too difficult to work with in practice as properties of the original kernels are often lost (Alvarez, Rosasco, and Lawrence (2011), Zhang, Xu, and Zhang (2012), De Vito, Umanita, and Villa (2013)).

15. The Class of Separable Kernels: An important class of matrix-valued reproducing kernels are *separable* kernels which can be factorized as the product of a scalar-valued kernel and a $T$-dimensional symmetric positive semi-definite matrix.

16. Form of Separable Kernels: In light of the previous discussions, these kernels are of the form

$$\gamma\big((x,t),(y,s)\big) = K(x,y)K_T(t,s)$$

for all $x, y$ in $X$ and $t, s$ in $T$.

17. Kernel Encoding Input/Output Dependencies: As the scalar valued kernel encodes the dependencies between the inputs, one can observe that the matrix-valued kernel encodes the dependencies among both the inputs and the outputs.

18. Challenges with Kernels in Function Spaces: As a final remark, the above treatment can be further extended to spaces of functions with values in function spaces but obtaining kernels for these spaces is a more difficult task.

# Connection between RKHS and ReLU Function

1. <u>Definition of the ReLU Function</u>: The ReLU function is commonly defined as

$$f(x) = \max(0, x)$$

and is a mainstay of the architecture of neural networks where it is used as an activation function.

2. <u>ReLU-like Functions using RKHS</u>: One can construct a ReLU-like non-linear function using the theory of reproducing kernel Hilbert spaces.

3. <u>ReLU-Type Function Construction</u>: This section derives this construction and shows how it implies the representation power of neural networks with ReLU activations.

4. <u>Premise - Hilbert Spaces of Continuous Functions</u>: This section works with the Hilbert space

$$\mathcal{H} = L_2^1(0)[0, \infty)$$

of absolutely continuous functions with

$$f(0) = 0$$

and square integrable, i.e., $L_2$, derivative. It has the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \int_0^\infty f'(x)g'(x)dx$$

5. <u>Dense Subspace for the Kernel</u>: To construct the reproducing kernel it suffices to consider a dense subspace, so let

$$f \in \mathbb{C}^1[0, \infty)$$

358

and

$$f(0) = 0$$

6. <u>Explicit Form for the Kernel</u>: The Fundamental Theorem of Calculus then gives

$$f(y) = \int_0^y f'(x)dx = \int_0^1 G(x,y)f'(x)dx = \langle K_y, f \rangle$$

where

$$G(x,y) = \begin{cases} 1 & x < y \\ 0 & otherwise \end{cases}$$

and

$$K'_y(x) = G(x,y)$$

$$K_y(0) = 0$$

i.e.,

$$K(x,y) = K_y(x) = \int_0^x G(z,y)dz = \begin{cases} x & 0 \le x < y \\ 0 & otherwise \end{cases} = \min(x,y)$$

This implies

$$K_y = K(\cdot, y)$$

reproduces $f$.

7. <u>Kernel Form at $y \to \infty$</u>: By taking the limit

$$y \to \infty$$

one obtains the ReLU function

$$K_\infty = \begin{cases} x & if \ x \geq 0 \\ 0 & otherwise \end{cases} = ReLU(x)$$

8. <u>Optimality of ReLU in Neural Networks</u>: Using this formulation, one can apply the representer theorem to the RKHS, letting one prove the optimality of using ReLU activations in neural network settings.

## References

- Alpay, D., and T. M. Mills (2003): A Family of Hilbert Spaces which are not Reproducing Kernel Hilbert Spaces *Journal of Analysis and Applications* **1 (2)** 107-111

- Alvarez, M., L. Rosasco, and N. Lawrence (2011): Kernels for Vector-valued Functions: A Review **arXiV**

- Berlinet, A., and C. Thomas (2004): *Reproducing Kernel Hilbert Spaces in Probability and Statistics* **Kluwer Academic Publishers** Amsterdam, Netherlands

- De Vito, E., D. Umanita, and S. Villa (2013): An Extension of Mercer Theorem to Vector-valued Measurable Kernels **arXiV**

- Pasternak-Winiarski, Z. (1992): On Weights which admit the Reproducing Kernel of Bergman Type *International Journal of Mathematics and Mathematical Sciences* **15 (1)** 1-14

- Paulsen, V. I. (2009): An Introduction to the Theory of Reproducing Kernel Hilbert Spaces

- Wikipedia (2022): Reproducing Kernel Hilbert Space

- Zhang, H., Y. Xu, and Q. Zhang (2012): Refinement of Operator-valued Reproducing Kernels *Journal of Machine Learning Research* **13** 91-136

- Zynda, T. L. (2020): On Weights which admit Reproducing Kernel of Szego Type *Journal of Contemporary Mathematical Analysis* **55** 320-327

# Representer Theorem

**Overview**

In statistical learning theory, a *representer theorem* is any of several related results stating that a minimizer $f^*$ of a regularized empirical risk functional defined over a reproducing kernel Hilbert space can be represented as a finite linear combination of kernel products evaluated on the input points in the training set data (Wikipedia (2022)).

## Formal Statement

1. <u>Scholkopf, Herbrich, and Smola Treatise</u>: The following Representer Theorem and its proof are due to Scholkopf, Herbrich, and Smola (2001).
2. <u>Setup of the Representer Theorem</u>: Consider a positive-definite real-valued kernel

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

   on a non-empty set $\mathcal{X}$ with a corresponding kernel Hilbert space $H_k$. The following are assumed to be given.
3. <u>Training Sample</u>:

$$(x_1, y_1), \cdots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$$

4. <u>Strictly Increasing Real-valued Function</u>:

$$g : [0, \infty) \rightarrow \mathbb{R}$$

5. <u>Arbitrary Error Function</u>:

$$E : (\mathcal{X} \times \mathbb{R}^2)^n \rightarrow \mathbb{R} \cup \{\infty\}$$

6. <u>Regularized Empirical Risk Functional</u>: The above taken together define the following regularized empirical risk functional on $H_k$:

$$f \mapsto E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|)$$

7. <u>Minimizer of the Empirical Risk</u>: Then, any minimizer of the empirical risk

$$f^* = \underset{f \in H_k}{\arg \min} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|)\}$$

admits a representation of the form:

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

where

$$\alpha_i \in \mathbb{R}$$

for all

$$1 \leq i < n$$

8. <u>Proof Step #1 - Mapping Functional</u>: Define a mapping

$$\varphi : \mathcal{X} \rightarrow H_k$$

$$\varphi(x) = k(\cdot, x)$$

so that

$$\varphi(x) = k(\cdot, x)$$

is itself

$$\mathcal{X} \to \mathbb{R}$$

9. <u>Proof Step #2 - Inner Product</u>: Since $k$ is a reproducing kernel, then

$$\varphi(x)(x') = k(x', x) = \langle \varphi(x'), \varphi(x) \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $H_k$.

10. <u>Proof Step #3 - Component Decomposition</u>: Given any $x_1, \cdots, x_n$ one can use orthogonal projection to decompose any

$$f \in H_k$$

into a sum of two functions, one lying in the span $\{\varphi(x_1), \cdots, \varphi(x_n)\}$, and the other lying in the orthogonal complement:

$$f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) + v$$

where

$$\langle v, \varphi(x_i) \rangle = 0$$

for all $i$.

11. <u>Applying $f$ on the Training Points</u>: The above orthogonal decomposition and the reproducing property together show that applying $f$ to any training point $x_j$ produces

$$f(x_j) = \langle \sum_{i=1}^{n} \alpha_i \varphi(x_i) + v, \varphi(x_j) \rangle = \sum_{i=1}^{n} \alpha_i \langle \varphi(x_i), \varphi(x_j) \rangle$$

which can be seen to be independent of $v$.

12. <u>Independence of $E$ from $v$</u>: Consequently, the value of the error function $E$ in

$$f^* = \begin{array}{c} \arg min \\ f \in H_k \end{array} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|)\}$$

is likewise independent of $v$

13. <u>Minimization of the Regularization Function</u>: For the second term, i.e., the regularization term, since $v$ is orthogonal to

$$\sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

and $g$ is strictly monotonic, one has

$$g(f) = g\left(\sum_{i=1}^{n} \alpha_i \varphi(x_i) + v\right) = g\left(\sqrt{\left\|\sum_{i=1}^{n} \alpha_i \varphi(x_i)\right\|^2 + \|v\|^2}\right) \geq$$

$$= g\left(\left\|\sum_{i=1}^{n} \alpha_i \varphi(x_i)\right\|\right)$$

14. <u>Setting Orthogonal Component to Zero</u>: Therefore, setting

$$v = 0$$

does not affect the first term of

$$f^* = \begin{array}{c} \arg min \\ f \in H_k \end{array} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|)\}$$

while it strictly decreases the second term.

15. <u>Proof Step #3 - Finale</u>: Consequently, any minimizer

$$f^* = \begin{array}{c} \arg min \\ f \in H_k \end{array} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|)\}$$

must have

$$v = 0$$

i.e., it must be of the form

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i \varphi(x_i) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

which is the desired result.

## Generalizations

1. <u>The Family of Representer Theorems</u>: The theorem stated above is a particular example of a family of results that are collectively referred to a "representer theorems": This section describes several such.

2. <u>Version of Kimeldorf and Wahba</u>: The first statement of a representer theorem was due to Kimeldorf and Wahba (1970) for the special case in which

$$E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i) - y_i]^2$$

$$g(\|f\|) = \lambda \|f\|^2$$

for

$$\lambda > 0$$

3. <u>Generalization by Scholkopf, Herbrich, Smola</u>: Scholkopf, Herbrich, and Smola (2001) generalized this result by relaxing the assumption of the squared-loss and allowing the regularizer to be any strictly monotonically increasing function $g(\cdot)$ of the Hilbert space norm.

4. <u>Augmenting the Regularized Risk Function</u>: It is possible to generalize further by augmenting the regularized and empirical risk functional through the addition of unpenalized offset terms.

5. <u>Extension using Unpenalized Function Family</u>: For example, Scholkopf, Herbrich, and Smola (2001) also consider the minimization

$$\tilde{f}^* = \arg\min\{E([x_1, y_1, \tilde{f}(x_1)], \cdots, [x_n, y_n, \tilde{f}(x_n)]) + g(\|f\|) \mid \tilde{f} = f + h$$
$$\in H_k \oplus span(\psi_p \mid 1 \leq p \leq M)\}$$

i.e., they consider functions of the form

$$\tilde{f} = f + h$$

where

$$f \in H_k$$

and $h$ is an unpenalized function lying in the span of a finite set of real-valued functions

$$\{\psi_p : \mathcal{X} \to \mathbb{R} \mid 1 \le p \le M\}$$

6. <u>Representation of the Corresponding Minimizer</u>: Under the assumption that the $n \times M$ matrix $\left[\psi_p(x_i)\right]_{ip}$ has the rank $M$ they show that the minimizer $\tilde{f}^*$ in

$$\tilde{f}^* = \arg\min\{E\left([x_1, y_1, \tilde{f}(x_1)], \cdots, [x_n, y_n, \tilde{f}(x_n)]\right) + g(\|f\|) \mid \tilde{f} = f + h$$
$$\in H_k \oplus span(\psi_p \mid 1 \le p \le M)\}$$

admits a representation of the form

$$\tilde{f}^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + \sum_{p=1}^{M} \beta_p \psi_p(\cdot)$$

where

$$\alpha_i, \beta_p \in \mathbb{R}$$

and the $\beta_p$ are all uniquely determined.

7. <u>Existence of the Representer Theorem</u>: The conditions under which a representer theorem exist were investigated by Argyriou, Micchelli, and Pontil (2009), who proved the following.

8. <u>Axioms behind Representer Theorem Existence</u>: Let $\mathcal{X}$ be a non-empty set, $k$ a positive-definite real-valued kernel on $\mathcal{X} \times \mathcal{X}$ with corresponding reproducing kernel Hilbert space $H_k$, and let

$$R : H_K \to \mathbb{R}$$

be a differentiable regularization function.

9. <u>Statement of Argyriou, Micchelli, and Pontil</u>: Then, given a training sample

$$(x_1, y_1), \cdots, (x_n, y_n) \in \mathcal{X} \to \mathbb{R}$$

and an arbitrary error function

$$E : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{\infty\}$$

a minimizer

$$f^* = \mathop{\arg\min}_{f \in H_k} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + R(f)\}$$

of the regularized empirical risk admits a representation of the form

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

where

$$\alpha_i \in \mathbb{R}$$

for all

$$1 \leq i \leq n$$

if and only if there exists a non-decreasing function

$$h : [0, \infty) \to \mathbb{R}$$

for which

$$R(f) = h(\|f\|)$$

10. Necessary and Sufficient Condition: Effectively, this result provides a necessary and sufficient condition on a differential regularizer $R(\cdot)$ under which the corresponding regularized empirical risk minimization

$$f^* = \begin{array}{c} \arg\min \\ f \in H_k \end{array} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + R(f)\}$$

will have a representer theorem.

11. Broad Class of Regularized Risk Minimizations: In particular, this shows that a broad class of regularized risk minimizations – much broader than those originally considered by Kimeldorf and Wahba (1970) – have representer theorems.

## Applications

1. Simplification of Empirical Risk Minimization: Representer theorems are useful from a practical standpoint because they dramatically simplify the regularized empirical risk minimization problem

$$f^* = \begin{array}{c} \arg \min \\ f \in H_k \end{array} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + R(f)\}$$

2. <u>Infinite dimensional Search Space</u>: In most interesting applications, the search domain $H_k$ for the minimization will be an infinite-dimensional subspace $L_2(\mathcal{X})$ and therefore the search – as written – does not admit implementation on finite-memory and finite-precision computers.

3. <u>Dimensionality Reduction offered by Representation Theorem</u>: In contrast, the representation of $f^*(\cdot)$ afforded by a representer theorem reduces the original infinite-dimensional minimization problem to a search for the optimal $n$-dimensional vector of coefficients

$$\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$$

$\alpha$ can then be obtained by applying any standard function minimization algorithm.

4. <u>Theoretical Basis for Feasible Solutions</u>: Consequently, representer theorems provide the theoretical basis for the reduction of the general machine learning problems to algorithms that can actually be implemented on computers in practice.

5. <u>Illustrative Development of Feasible Solution</u>: The following provides an example of how to solve for the minimizer whose existence is guaranteed by the representer theorem.

6. <u>Reduction to a Finite Linear System</u>: This method works for any positive definite kernel $K$ and allows one to transform a complicated – possibly infinite-dimensional – optimization problem into a simple linear system that can be solved numerically.

7. <u>Forms of Empirical Loss/Regularizer</u>: Assume that one uses a least-squares error function

$$E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) := \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

and a regularization function

$$g(x) = \lambda x^2$$

for some

$$\lambda > 0$$

8. Form of the Minimizer Function: By the representer theorem, the minimizer

$$f^* = \underset{f \in \mathcal{H}}{\arg\min} \{E([x_1, y_1, f(x_1)], \cdots, [x_n, y_n, f(x_n)]) + g(\|f\|_{\mathcal{H}})\}$$

$$= \underset{f \in \mathcal{H}}{\arg\min} \left\{ \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

has the form

$$f^*(x) = \sum_{i=1}^{n} \alpha_i^* k(x, x_i)$$

for some

$$\alpha^* = (\alpha_1^*, \cdots, \alpha_n^*) \in \mathbb{R}^n$$

9. Formulation of the Kernel Coefficients: Noting that

$$\|f\|_{\mathcal{H}}^2 = \langle \sum_{i=1}^{n} \alpha_i^* k(\cdot, x_i), \sum_{j=1}^{n} \alpha_j^* k(\cdot, x_j) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i^* \alpha_j^* k(x_i, x_j)$$

one sees that $\alpha^*$ has the form

372

$$\alpha^* = \begin{matrix} \arg\min \\ \alpha \in \mathbb{R}^n \end{matrix} \left\{ \sum_{i=1}^{n} \left[ y_i - \sum_{j=1}^{n} \alpha_j k(x_i, x_j) \right]^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

$$= \begin{matrix} \arg\min \\ \alpha \in \mathbb{R}^n \end{matrix} \{ \|y - A\alpha\|^2 + \lambda \alpha^T A\alpha \}$$

where

$$A_{ij} = k(x_i, x_j)$$

and

$$y = (y_1, \cdots, y_n)$$

10. <u>Explicit Form for the Coefficients</u>: This can be factored out and simplified to

$$\alpha^* = \begin{matrix} \arg\min \\ \alpha \in \mathbb{R}^n \end{matrix} \{ \alpha^T (A^T A + \lambda A)\alpha - 2\alpha^T Ay \}$$

11. <u>Existence of Single Global Minima</u>: Since $A^T A + \lambda A$ is positive-definite, there is indeed a single global minimum for this expression.

12. <u>Convex Nature of Optimization Function</u>: Let

$$F(\alpha) = \alpha^T (A^T A + \lambda A)\alpha - 2\alpha^T Ay$$

and note that $F$ is convex.

13. <u>Gradient of the Optimization Function</u>: Then $\alpha^*$, the global minima, can be solved by setting

$$\nabla_\alpha F = 0$$

14. <u>Application of a Linear Solver</u>: Recalling that all positive definite matrices are invertible, one can see that

$$\nabla_\alpha F = \alpha^T(A^T A + \lambda A)\alpha - 2\alpha^T Ay = 0 \Rightarrow \alpha^* = (A^T A + \lambda A)^{-1}Ay$$

so the minimizer may be found via a linear space.

# References

- Argyriou, A., C. A. Micchelli, and M. Pontil (2009): When is there a Representer Theorem? Vector versus Matrix Regularizers *Journal of Machine Learning Research* **10** 2507-2529
- Kimeldorf, G. S. and G. Wahba (1970): A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines *Annals of Mathematical Statistics* **41 (2)** 495-502
- Scholkopf, B., R. Herbrich, and A. J. Smola (2001): [A Generalized Representer Theorem](#)
- Wikipedia (2022): [Representer Theorem](#)

# Kernel Method

## Overview

1. <u>Class of Algorithms for Pattern Analysis</u>: *Kernel machines* are a class of algorithms for pattern analysis, whose best-known member is the support vector machine- SVM (Wikipedia (2022)).

2. <u>Less Commonly Known Algorithms</u>: These include the Importance Vector Machine, and Kernel PCA (Theodoridis (2008)).

3. <u>General Task of Pattern Analysis</u>: The focus is to find and study general types of relations – for example clusters, rankings, principal components, correlations, classifications – in datasets.

4. <u>Data Transformation in Feature Space</u>: For many algorithms that solve these tasks, the data in raw representation have to be explicitly transformed into feature vector representations using user-specified *feature map*; in contrast, kernel methods require only a user-specified *kernel*, i.e., a similarity function over all pairs of data points computed using inner products.

5. <u>Dimensionality Reduction using Represeter Theorem</u>: The feature map in kernel machines is infinite-dimensional, but only requires a finite-dimensional matrix from user-input according to the Represeter theorem.

6. <u>Kernel Machine Sample Size Performance</u>: Kernel machines are slow to compute for datasets larger than a couple of thousand examples without parallel processing.

7. <u>High dimensional, *Implicit* Feature Space</u>: Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional *implicit* feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in feature space.

8. <u>Motivation behind the Kernel Trick</u>: This operation is often computationally cheaper than the explicit computation on the feature map using the coordinates. This approach is called the *kernel trick* (Theodoridis (2008)).

9. <u>Applications of Kernel Functions</u>: Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

10. <u>Algorithms that use Kernel Methods</u>: Algorithms capable of operating with kernels include the kernel perceptron, support-vector machines SVM, Gaussian processes, principal component analysis PCA, canonical correlation analysis, ridge regression, spectral clustering, linear adaptive filter, and many others.

11. <u>Theoretical Foundations of Kernel Applications</u>: Most kernel algorithms are based on convex optimization or eigen-problems are statistically well-founded.

12. <u>Analysis of Kernel Algorithms' Complexity</u>: Typically, their statistical properties are analyzed using statistical learning theory – for example, using Rademacher complexity.

## Motivation and Informal Expression

1. <u>Instance-based Learning Methods</u>: Kernel methods can be thought of as instance-based learners; rather than learning some fixed set of parameters corresponding to features of their inputs, they instead "remember" the ith training example $(x_i, y_i)$ and learn for it a corresponding weight $w_i$.

2. <u>Similarity Method Based Outcome Prediction</u>: Prediction for unlabeled inputs, i.e., those not in the training set, is treated by applying the similarity function $k$, called a *kernel*, between the unlabeled input $x'$ and each of the training inputs $x_i$.

3. <u>Example - A Kernelized Binary Classifier</u>: For instance, a kernelized binary classifier typically computes a weighted sum of similarities

$$y = sgn \left( \sum_{i=1}^{n} w_i y_i k(x_i, x') \right)$$

where:

4. Kernelized Binary Classifier's Predictor Label:

$$\hat{y} = \{-1, +1\}$$

is the kernelized binary classifier's predicted label for the unlabeled input $x'$ whose hidden true label $y$ is of interest.

5. Kernel Function that Measures Similarity:

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

is the kernel function that measures similarity between any pair of inputs

$$x', x \in \mathcal{X}$$

6. Sum Ranges over Labeled Examples: The sum ranges over the $n$ labeled examples

$$\{(x_i, y_i)\}_{i=1}^{n}$$

in the classifier's training set, with

$$y_i \in \{-1, +1\}$$

7. <u>Weights for the Training Examples</u>: The weights

$$w_i \in \mathbb{R}$$

are the weights for the training examples, as determined by the learning algorithm.

8. <u>Sign of the Classification Outcome</u>: The sign function $sgn$ determines whether the predicted classification $\hat{y}$ comes out positive or negative.

9. <u>Performance of the Kernel Classifiers</u>: Kernel classifiers rose to great prominence with the popularity of the support-vector machine - SVM, when the SVM was found to be competitive with neural networks on tasks such as handwriting recognition.

**Mathematics – The Kernel Trick**

1. Advantage of the Kernel Trick: The kernel trick avoids the explicit mapping that is needed to get linear learning algorithms to learn a non-linear function or decision boundary.

2. Kernel Function as an Inner Product: For all $x$ and $x'$ in the input space $\mathcal{X}$, certain functions $k(x, x')$ can be expressed as an inner product in another space $v$.

3. Definition of a Kernel Function: The function

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

is referred to as a *kernel* or a *kernel function*.

4. Use of "Kernel" in Mathematics: The word "kernel" is used in mathematics to denote a weighting function for a weighted sum or integral.

5. Example - Illustration of the SVM Kernel:
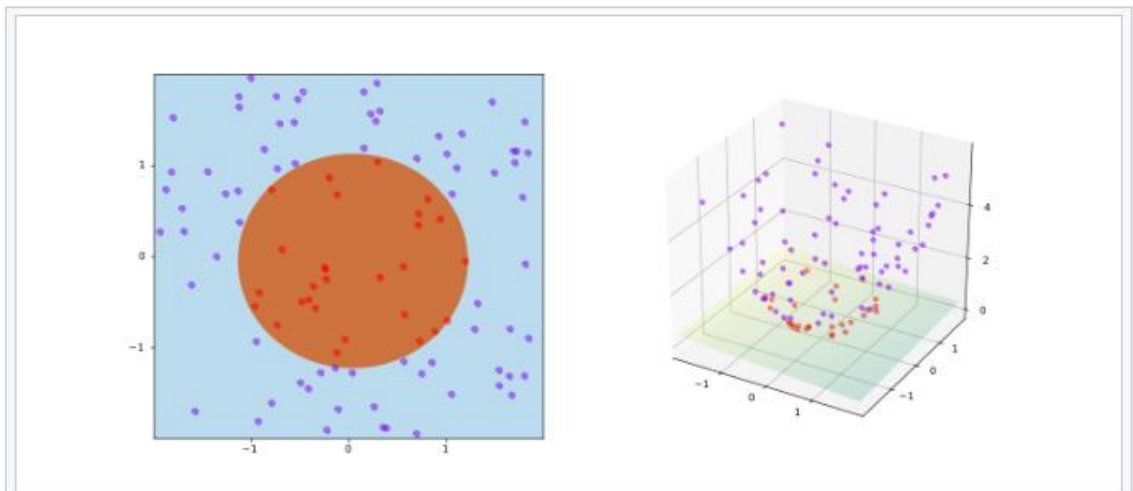


Above is an SVM with kernel given by

$$\varphi(a, b) = (a, b, a^2 + b^2)$$

and thus

$$K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}.\boldsymbol{y} + \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2$$

The training points are mapped to a 3-dimensional space where a separating hyperplane can be easily found.

6. <u>Structure behind Machine Learning Problems</u>: Certain problems in machine learning have more structure than an arbitrary weighting function $k$.

7. <u>Dot Product using Feature Maps</u>: The computation is made much simpler of the kernel can be written in the form of a "feature map"

$$\varphi : \mathcal{X} \to \nu$$

satisfies

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_\nu$$

8. <u>Explicit Feature Map Function Form</u>: The key restriction is that $\langle \cdot, \cdot \rangle_\nu$ must be a proper inner product. On the other hand, an explicit representation for $\varphi$ is not necessary, as long as $\nu$ is an inner product space.

9. <u>Mercer's Theorem - Existence of $\varphi$</u>: The alternative follows from Mercer's theorem: an implicitly defined $\varphi$ exists whenever the space $\mathcal{X}$ can be equipped with a suitable measure ensuring that function $k$ satisfies Mercer's condition.

10. <u>Parallels with Linear Algebra</u>: Mercer's theorem is similar to the generalization of the result from linear algebra that associates an inner product to any positive-definite matrix. In fact, Mercer's condition can be reduced to this simpler case.

11. <u>Explicit Use of Counting Measure</u>: If one chooses as a measure the counting measure

$$\mu(T) = |T|$$

for all

$$T \subset \mathcal{X}$$

which counts the number of points inside the set $T$, then the integral in the Mercer's theorem reduces to a summation

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

12. <u>Fulfillment of the Mercer's Condition</u>: If this summation holds for all finite sequences of points

$$(x_1, \cdots, x_n) \in \mathcal{X}$$

and all choices of $n$ real-valued coefficients $(c_1, \cdots, c_n)$ – cf. positive definite kernel – then the function $k$ satisfies Mercer's condition.

13. <u>Gram Matrix over Training Set</u>: Theoretically, a Gram matrix

$$K \in \mathbb{R}^{n \times n}$$

with respect to $\{x_1, \cdots, x_n\}$ – sometimes also called a "kernel method" (Hofman, Scholkopf, and Smola (2008)) – where

$$K_{ij} = k(x_i, x_j)$$

must be positive-definite (Mohri, Rostamizadeh, and Talwalkar (2012)).

14. <u>Approximation to Notion of Similarity</u>: Similarly, for machine learning heuristics, choices of a function $k$ that do not satisfy Mercer's condition may still perform reasonably if $k$ at least approximates the intuitive idea of similarity. Regardless of whether $k$ is a Mercer kernel, $k$ may still be referred to as a "kernel".

15. Kernel Functions as Covariance Metrics: If the kernel function $k$ is also a covariance function as used in Gaussian processes, then the Gram matrix $K$ can also be called a covariance matrix (Rasmussen and Williams (2005)).

## Applications

Application areas of kernel methods are diverse and include geo-statistics (Honarkhah Caers (2010)), kriging, inverse distance weighting, 3D reconstruction, bio-informatics, chemo-informatics, information extraction, and hand-writing recognition.

## References

- Hofmann, T., B. Scholkopf, and A. J. Smola (2008): Kernel Methods in Machine Learning *Annals of Statistics* **36 (3)** 1171-1220
- Honarkhah, M., and J. Caers (2010): Stochastic Simulation of Patterns using Distance-based Pattern Modeling *Mathematical Geosciences* **42 (5)** 487-517
- Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012): *Foundations of Machine Learning* **MIT Press** Cambridge, MA
- Rasmussen, C. E., and C. K. I. Williams (2005): *Gaussian Processes for Machine Learning* **MIT Press** Cambridge, MA
- Theodoridis, S. (2008): *Pattern Recognition* **Elsevier** Amsterdam
- Wikipedia (2022): Kernel Method

# Householder Transformation

## Overview

1. Definition of the Householder Transformation: In linear algebra, a *Householder transformation* – also known as a *Household reflection* or *elementary reflector* – is a linear transformation that describes a reflection about a plane or a hyperplane containing the origin (Householder (1958), Wikipedia (2022)).
2. Analogue over General Inner Product: Its analogue over the general inner product spaces is the Householder operator.

## Definition - Transformation

1. Normal Vector Defining Reflection Hyperplane: The reflection hyperplane can be defined by its *normal* vector, a unit vector $v$ – a vector with a length $1$ – that is orthogonal to the hyperplane.
2. Reflection of $x$ about Hyperplane: The reflection of the point $x$ about this hyperplane is the linear transformation

$$x - 2\langle x, v \rangle v = x - 2v(v^H x)$$

where $v$ is given as a column unit vector with Hermitian transpose $v^H$.

## Definition – Householder Matrix

The matrix constructed from this transformation can be expressed in terms of an outer product

$$P = I - 2vv^H$$

known as the Householder matrix, where $I$ is the identity matrix.

## Definition – Properties

1. <u>Hermitian</u>: The Householder matrix has the following properties. It is Hermitian:

$$P = P^H$$

2. <u>Unitary</u>: It is unitary:

$$P^{-1} = P^H$$

3. <u>Involutary</u>: Hence, it is involutary:

$$P = P^{-1}$$

4. <u>Eigenvalues $\pm 1$</u>: The Householder matrix has eigenvalues $\pm 1$. To see this, notice that if $u$ is orthogonal to the vector $v$ which was used to create the reflector, then

$$Pu = u$$

i.e., 1 is an eigenvalue with multiplicity $n - 1$, since there are $n - 1$ independent vectors orthogonal to $v$. Also, notice

$$Pv = -v$$

and so $-1$ is an eigenvalue with multiplicity $+1$.

5.  Determinant: The determinant of a Householder reflection is -1, since the determinant of a matrix is the product of the eigenvalues, in this case one of which is -1 with the remainder being +1 – as in the previous point.

## Applications – Numerical Linear Algebra

1.  Extensive Use in Linear Algebra: Householder transformations are widely used in numerical linear algebra, for example, to annihilate the entries below the main diagonal of the matrix, to perform QR decomposition, and in the first step of the QR decomposition algorithm.
2.  Use in Hessenberg/Hermitian Matrices: They are also widely used for transformation to a Hessenberg form. For symmetric or Hermitian matrices, the symmetry can be preserved, resulting in triangularization (Schabauer, Pacher, Sunderland, and Gansterer (2010)).

## Applications – Numerical Linear Algebra – QR Decomposition

Householder reflections can be used to calculate QR decompositions by reflecting one first column of a matrix onto a multiple of a standard basis vector, calculating the transformation matrix, multiplying it with the original matrix and then recursing down the $(i, i)$ minors of that product.

## Applications – Numerical Linear Algebra – Tridiagonalization

1. <u>Treatment in Burden, Faires, and Burden</u>: This procedure is presented Burden, Faires, and Burden (2015) – it just uses a slightly altered $sgn$ function with

$$sgn(0) = 1$$

2. <u>Step #1 - Determining $\alpha/r$</u>: In the first step, to form the Householder matrix in each step one needs to determine $\alpha$ and $r$, which are:

$$\alpha = -sgn(a_{21}) \sqrt{\sum_{j=2}^{n} a_{j1}^2}$$

$$r = \sqrt{\frac{1}{2}(\alpha^2 - a_{21}\alpha)}$$

3. <u>Step #2 – Construction of $v$</u>: In $\alpha$ and $r$, construct vector $v$:

$$v(1) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

where

$$v_1 = 0$$

$$v_2 = \frac{a_{21} - \alpha}{2r}$$

and

$$v_k = \frac{a_{k1}}{2r}$$

for each

$$k = 3, 4, \cdots, n$$

4. Step #3 - Initial $P/A$: Then compute

$$P(1) = I - 2v(1)v^T(1)$$

$$A(2) = P(1) \cdot A \cdot P(1)$$

5. Step #4 - Iterative $P/A$: Having found $P(1)$ and computing $A(2)$ the process is repeated for

$$k = 2, \cdots, n - 2$$

as follows:

$$\alpha = -sgn\left(a_{k+1,k}(k)\right)\sqrt{\sum_{j=k+1}^{n} a_{jk}^2(k)}$$

$$r = \sqrt{\frac{1}{2}\left(\alpha^2 - a_{k+1,k}(k)\alpha\right)}$$

$$v_1(k) = v_2(k) = \cdots = v_k(k) = 0$$

$$v_{k+1}(k) = \frac{a_{k+1,k}(k) - \alpha}{2r}$$

$$v_j(k) = \frac{a_{jk}(k)}{2r}$$

for each

$$j = k + 2, k + 3, \cdots, n$$

$$P(k) = I - 2v(k)v^T(k)$$

$$A(k + 1) = P(k) \cdot A(k) \cdot P(k)$$

6. <u>Computing Tridiagonal and Symmetric Matrix</u>: Continuing in this manner, the tridiagonal and the symmetric matrix is formed.

## Applications – Numerical Linear Algebra – Examples

1. <u>Applying Householder Scheme to a 4 × 4 Matrix</u>: In this example, taken from Burden, Faires, and Burden (2016), the given matrix is transformed to the similar tridiagonal matrix $A_3$ by using the Householder method.

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}$$

2. First Householder Iteration: The first Householder matrix is generated as shown below:

$$Q_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

$$A_2 = Q_1 A Q_1 = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & 1 & \frac{4}{3} \\ 0 & 1 & \frac{5}{3} & -\frac{4}{3} \\ 0 & \frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix}$$

3. Second Householder Iteration: $A_2$ generated above is used as shown below:

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix},$$

$$A_3 = Q_2 A_2 Q_2 = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & -\frac{5}{3} & 0 \\ 0 & -\frac{5}{3} & -\frac{33}{25} & \frac{68}{75} \\ 0 & 0 & \frac{68}{75} & \frac{149}{75} \end{bmatrix}$$

4.  Final Result - Tridiagonal Symmetric Matrix: As can be seen, the final result is a tridiagonal symmetric matrix similar to the original one. The process here finished in two steps.

## Computational and Theoretical Relationship to Other Unitary Transformations

1.  Recap – Constructing a Householder Transform: The householder transformation is a reflection about a hyperplane with a unit normal vector $v$, as stated earlier. An $N \times N$ unitary transformation $U$ satisfies

$$UU^H = I$$

2.  Householder Transform Unit Determinant - Statement: Taking the determinant - $N^{th}$ power of the geometric mean, and trace – proportional to the arithmetic mean, of a unitary matrix reveals that its eigenvalues $\lambda_i$ has unit modulus.

3.  Householder Transform Unit Determinant - Proof: This can be directly:

$$\frac{Trace(UU^T)}{N} = \frac{\sum_{j=1}^{N}|\lambda_j|^2}{N} = 1$$

$$\det(UU^T) = \prod_{j=1}^{N}|\lambda_j|^2 = 1$$

4.  Equality of Arithmetic/Geometric Means: Since arithmetic and geometric means are equal if the variables are constant, the claim of unit modulus is therefore established.

5.  Givens Rotations and Householder Reflections: For the case of real-valued unitary matrices, one obtains orthogonal decompositions

$$UU^T = I$$

It may be shown that any orthogonal matrix may be decomposed into a product of $2 \times 2$ rotations called Givens Rotations, and Householder Reflections.

6. Length Invariance under Rotation/Reflection: This has an intuitive appeal since multiplication of a vector by an orthogonal matrix preserves the length of that vector, and rotations and reflections exhaust the set of real-valued geometric operations hat render invariant a vector's length.

7. Coset Decomposition of Unitary Matrices: The Householder transformation has been shown to have a one-to-one relationship with canonical coset decomposition of unitary matrices defined in the group theory, which can be used to parametrize unitary operators in a very efficient manner (Cabrera, Strohecker, and Rabitz (2010)).

8. Advantage of Householder over Givens: Finally, it may be noted that a single Householder transform, unlike a solitary Givens transform, can act on all columns of a matrix, and as such exhibits the lowest computational cost for QR decomposition and triangularization.

9. Drawback of Householder over Givens: The penalty for this "computational optimality" is, of course, that Householder operations cannot be as deeply or as efficiently parallelized.

10. Suitability of Householder vs. Givens: As such, Householder is preferred over dense matrices on sequential machines, while Givens is preferred on sparse matrices, and/or parallel machines.

## References

- Burden, R. L., J. D. Faires, and A. M. Burden (2015): *Numerical Analysis 10$^{th}$ Edition* **Cengage Learning** Boston, MA

- Cabrera, R., T. Strohecker, and H. Rabitz (2010): The Canonical Coset Decomposition of Unitary Matrices through Householder Transformations *Journal of Mathematical Physics* **51 (8)** 082101

- Householder, A. S. (1958): Unitary Transformation of a Non-singular Matrix *Journal of the ACM* **5 (4)** 339-342

- Schabauer, H., C. Pacher, A. G. Sunderland, and W. N. Gansterer (2010): Toward a Parallel Solver for Generalized Complex Symmetric Eigenvalue Problems *Procedia Computer Science* **1 (1)** 437-445

- Wikipedia (2022): [Householder Transformation](#)

# The Householder Transformation in Numerical Linear Algebra

## Abstract

1. <u>Use of the Householder Transformation</u>: This chapter defines the Householder transformation, then puts it to work in several ways (Kerl (2008)).

2. <u>Geometrical Illustration of Algebraic Properties</u>: To illustrate the usefulness of geometry to elegantly derive and prove seemingly algebraic properties if the transform;

3. <u>Use in Determinants and Inverses</u>: To demonstrate an application to numerical linear algebra – specifically, for matrix determinants and inverses;

4. <u>Analysis of Roundoff Errors</u>: To show how geometric notions of determinant and matrix norm can be used to easily understand round-off error in Householder and Gaussian-elimination methods.

## Linear Algebra

1. <u>Gaussian Elimination vs. Householder Method</u>: This chapter compares and contrasts two techniques for computation of determinants and inverses of square matrices: the more familiar Gaussian elimination method, and the less familiar Householder method. It primarily uses geometric methods to do so.

2. <u>Review of the Requisite Background</u>: This requires some preliminaries from linear algebra, including geometric interpretations of determinant, matrix norm, and error propagation.

## Geometric Meanings of Determinant and Matrix Norm

1.  <u>Linear Transformation - Definition of Determinant</u>: Let

$$A : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

be a linear transformation – which may be thought of as a matrix with respect to a standard basis. The *determinant* of $A$ can be thought of as the *signed volume* of the image of the *unit cube*.



det($A$) = volume of image of unit cube
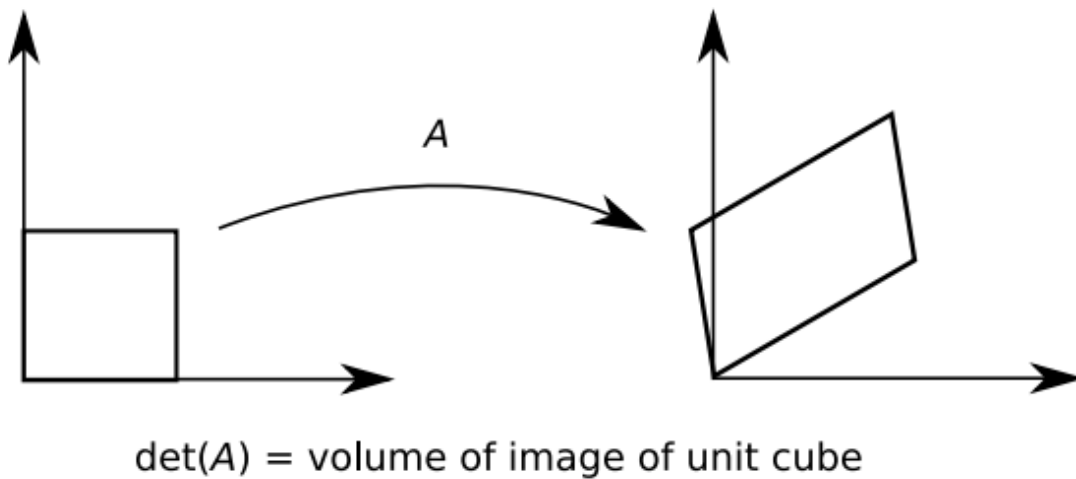
2.  <u>Geometric Definition of Matrix Norm</u>: The *matrix norm* is, by definition, the *maximum extent* of the image of the *unit ball*.
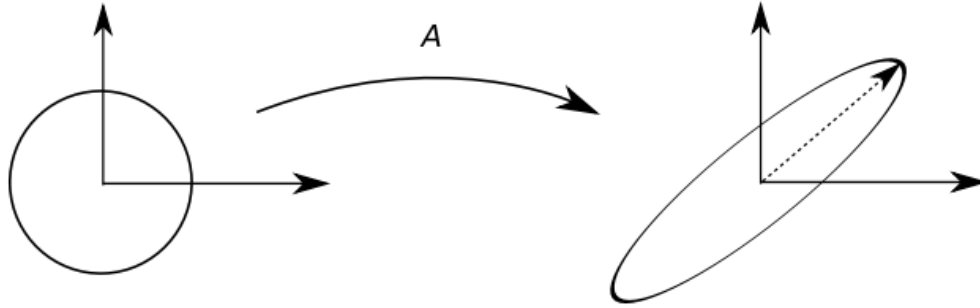
||A|| = maximum extent of image of unit ball

3. <u>Mathematical Formulation of Matrix Norm</u>: The matrix norm of $A$ is defined either by

$$\|A\| = \sup_{\|x\| = 1} \|Ax\|$$

or equivalently by

$$\|A\| = \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}$$

4. <u>Intuition behind the Matrix Norm</u>: That is, either one considers the vectors in the unit ball and finds the maximum extent of their images, or we consider *all* non-zero vectors and find the maximum amount by the which they are scaled. The equivalence of these two characterizations is due to the linearity of $A$.

5. <u>Intuition behind the Matrix Norm</u>: That is, either one considers the vectors in the unit ball and finds the maximum extent of their images, or we consider *all* non-zero vectors and find the maximum amount by which they are scaled. The equivalence of these two characterizations is due to the linearity of $A$.

6. <u>Implications of Determinant Value $\pm 1$</u>: The matrix with determinant $\pm 1$ preserves unsigned volume, but does not necessarily preserve norm. Of particular interest for this chapter are three kinds of matrices.

7.  Rotation Matrix: A $2 \times 2$ *notation matrix* is of the form

$$A = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}$$
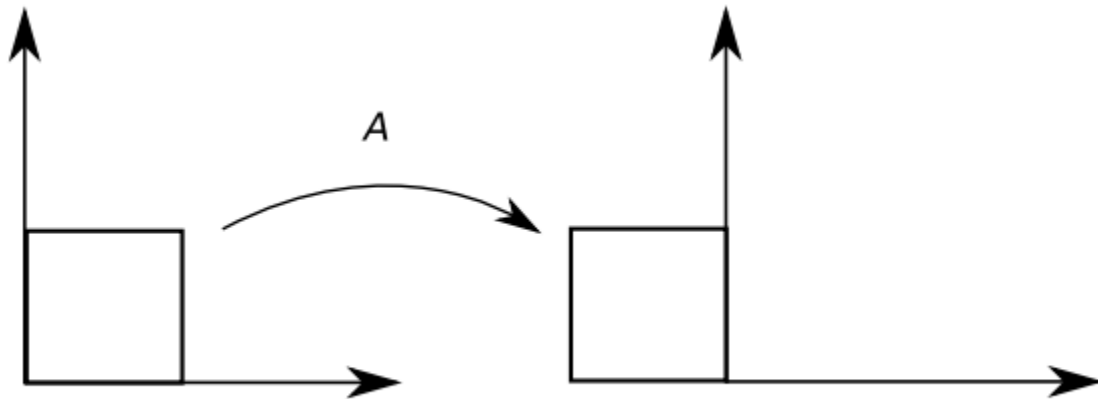
and has determinant 1
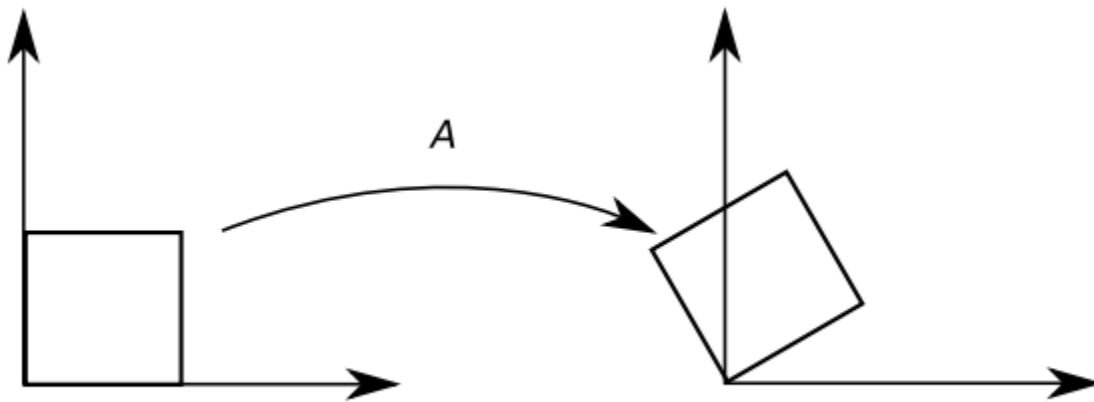


Rotation has determinant 1 and norm 1

8.  Reflection Matrix: An example of a $2 \times 2$ *reflection matrix*, reflecting about the $y$-axis

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is which has determined $-1$:
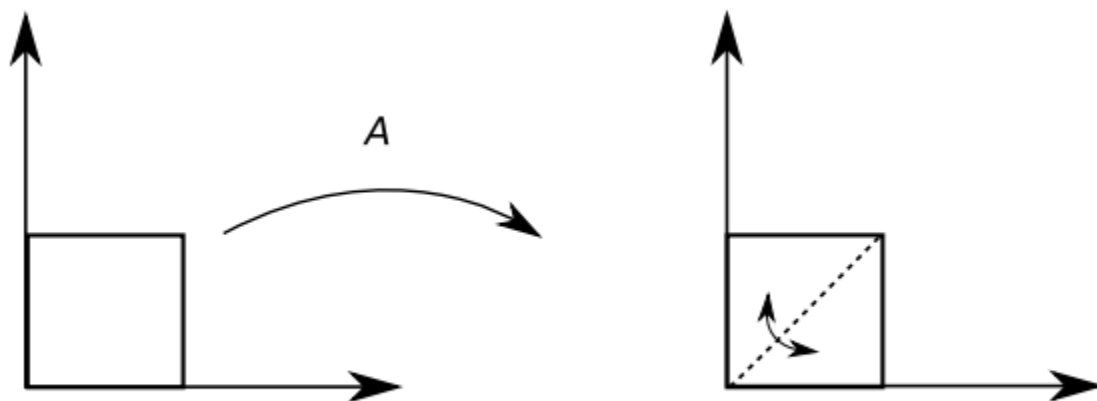
Reflection has determinant -1 and norm 1

9. <u>Permutation Matrix</u>: Another example of a reflection is a *permutation matrix*:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

which has determinant $-1$:



Permutation has determinant -1 and norm 1
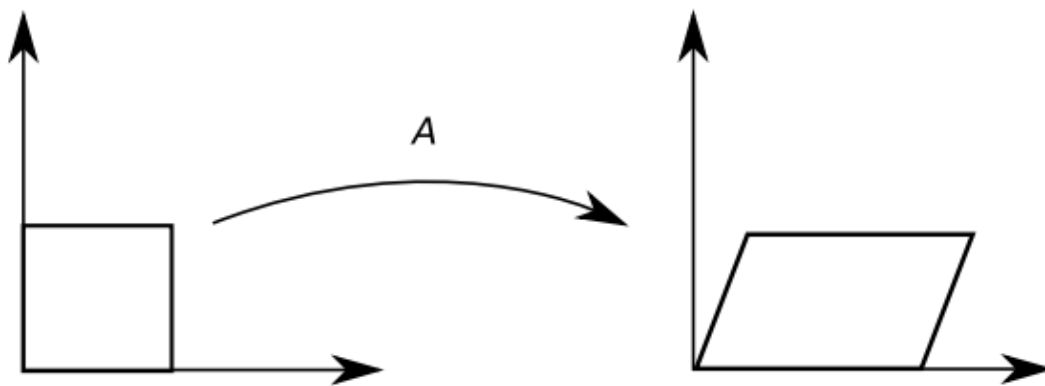
This reflection is about the $45^0$ line

$$x = y$$

10. <u>Reflection about an Arbitrary Axis</u>: Construction of a reflection matrix about an arbitrary axis is accomplished using Householder transformations, as discussed later.

11. <u>Shear Matrix</u>: An example of a $2 \times 2$ *shear matrix* is

$$A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$$

which has determinant $+1$



Shear has determinant 1 but not norm 1

## Computation of Determinants

1. <u>Computing Determinants using Cofactor Expansion</u>: In elementary linear algebra (Friedberg, Insel, and Spence (2002)), one is, first taught to compute determinants using *cofactor expansion*. This is fine for computing determinants of $2 \times 2$'s or $3 \times 3$'s. However, it is ineffective for larger matrices.

2. <u>Determinant of a 2 × 2 Matrix</u>: The determinant of a 2 × 2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $ad - bc$; uses two multiplies and an add.

3. <u>Determinant of a 3 × 3 Matrix</u>: To compute the determinant of a 3 × 3 matrix using cofactor expansion, one works down a new column or a row. There are 3 2 × 2 determinants.

4. <u>Determinant of a 4 × 4 Matrix</u>: To do a 4 × 4, there are 4 determinants of 3 × 3's, each of which recursively takes 3 determinants of 2 × 2's.

5. <u>Determinant of an $n \times n$ Matrix</u>: Continuing this pattern, it can be seen that there are on the order of $n!$ multiplies and adds for an $n \times n$ matrix. For example,

$$25! \approx 10^{25}$$

Even at a billion operations a second, this requires $10^{16}$ seconds, which is hundreds of millions of years. A better approach is needed.

6. <u>Upper and Lower Triangular Matrices</u>: If a matrix is *upper-triangular* or *lower-triangular*, then its determinant is the product of its diagonal entries. There are $n$ of these and one only needs to multiply them, so there are $n$ multipliers.

7. <u>Transformation into Upper Triangular Form</u>: The next step is to see how efficiently one can rework a given square matrix into upper-triangular form, and how this modification affects the determinant.

8. <u>The Process of Upper Triangularization</u>: Upper-triangularization is the process of converting certain elements of a matrix to zero, while modifying the others.

9. <u>Pre-multiplication of $Q$ by $A$</u>: Recall that when a matric $Q$ acts by pre-multiplication on a matrix $A$, one may think of $Q$ acting on each column vector of $A$.

10. <u>Zeroing out Locations in Matrix</u>: That is, the j<sup>th</sup> column of $QA$ is simply the j<sup>th</sup> column of $A$. Thus, the column vectors can be transformed to place zeros in various locations.

11. <u>Impact of Transformation on Determinant</u>: If one applies transformations $M_1$ through $M_m$, then

$$\det(M_m \cdots M_2 M_1 A) = \det(M_m) \cdots \det(M_2) \det(M_1) \det(A)$$

i.e.,

$$\det(A) = \frac{\det(M_m \cdots M_2 M_1 A)}{\det(M_m) \cdots \det(M_2) \det(M_1)}$$
$$= \frac{product\ along\ diagonal\ of\ upper\ triangular\ matrix}{\det(M_m) \cdots \det(M_2) \det(M_1)}$$

12. <u>Tracking the Determinants of Transforms</u>: That is, all one needs to do is to keep track of the determinants of the transformations, multiply them up, and compute the determinant of the resulting upper-triangular matrix.

13. <u>$\mathcal{O}(n)$ for Gaussian and Householder</u>: It can be shown that Gaussian elimination and Householder methods for upper-triangularization take $\mathcal{O}(n^3)$. Thus, these methods are far more efficient than naïve cofactor expansion.

## Computation of Matrix Inverses

1. <u>Augmented Matrix vs. Cofactor Extraction</u>: A more efficient method than cofactor expansion is the *augmented matrix* method. Here, one places the $n \times n$ input matrix next to an identity matrix $[A \mid I]$ and converts the augmented matrix into row-reduced echelon form.

2. <u>Identity Matrix on the LHS</u>: If it is possible to get the identity matrix on the LHS, then the inverse will be found on the RHS: $[I \mid A^{-1}]$

3. <u>Row-Reduction of Augmented Matrix</u>: When the augmented matrix is row-reduced, one applies a sequence $M_1, \cdots, M_n$ of linear transformations to the augmented matrix. Let their product be $M$:

$$M = M_n \cdots M_1$$

4. <u>Proof of Inverse at RHS</u>: Then, if row reduction works, i.e., if one gets $I$ on the right-hand side, then

$$M[A \mid I] = [I \mid *]$$

i.e.,

$$MA = I$$

But this forces

$$M = A^{-1}$$

Thus,

$$* = MI = A^{-1}I = A^{-1}$$

5. <u>Two-Step Echelon Formation Process</u>: Note that placing a matrix into a row reducing echelon form is a two-step process.

6. <u>Generating the Upper Triangular Form</u>: Putting it into the row echelon form simply means putting it into the *upper-triangular* form and dividing each row by its leading coefficient:

$$\begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix} \mapsto \begin{bmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \end{bmatrix} \mapsto \begin{bmatrix} 1 & * & * & * & * & * \\ 0 & 1 & * & * & * & * \\ 0 & 0 & 1 & * & * & * \end{bmatrix}.$$

7. <u>Clearing the Above-diagonal Entries</u>:

$$\begin{bmatrix} 1 & * & * & * & * & * \\ 0 & 1 & * & * & * & * \\ 0 & 0 & 1 & * & * & * \end{bmatrix} \longmapsto \begin{bmatrix} 1 & 0 & 0 & * & * & * \\ 0 & 1 & 0 & * & * & * \\ 0 & 0 & 1 & * & * & * \end{bmatrix}.$$

8. <u>Sequence of $M$'s to apply</u>: The remaining question to be addressed below is what particular sequence of $M$'s to apply.

## Error Propagation

1. <u>Challenges with Poorly Designed Methods</u>: Whenever one uses numerical techniques, on needs to realize that a poorly designed algorithm can allow error to grow enormously, producing highly incorrect results.

2. <u>Errors Generated in Numerical Analysis</u>: The study error in numerical analysis is discussed in detail in, for example, Burden, Faires, and Burden (2015).

3. <u>Error Propagation under Linear Transform</u>: Consider a true matrix $A$ at the kth step of an upper-triangularization process along with an actual matrix $A + \varepsilon$, where $\varepsilon$ is a matrix.

4. <u>Transformation of a Perturbed Matrix</u>: Since one uses linear transformation, one has

$$Q(A + \varepsilon) = Q(A) + Q(\varepsilon)$$

5. <u>Transformation of a Perturbed Vector</u>: Also, since the jth column of $QA$ is $Q$ times the jth column of $A$; this further becomes

$$Q(v + \eta) = Q(v) + Q(\eta)$$

6. <u>Extracting Vector Norm from Matrix Norm</u>: The *vector norm* of $Q(\eta)$ is related to the vector norm of $\eta$ by the *matrix norm* of $Q$; since

$$\|Q\| \geq \frac{\|Q(\eta)\|}{\|\eta\|}$$

one has

$$\|Q(\eta)\| \leq \|Q\| \cdot \|\eta\|$$

7. <u>Norm-preserving Linear Transformation Matrix</u>: In particular, if a transformation matrix $Q$ is norm-preserving, it does not increase error.
8. <u>Norms of Gaussian/Householder Schemes</u>: The next sections consider the norms used in the Gaussian elimination and Householder methods.

## Gaussian Elimination

The next few sections discuss Gaussian elimination from the standpoint of concatenated linear transformations, with an emphasis on the geometrical of these transformations.

## Row Reduction using Gaussian Elimination

1. <u>Row Reduction of Sample Matrix</u>: For clarity, one considers matrices of height 2. The following matrix is row-reduced:

$$\begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This example illustrates the three matrices used in Gaussian elimination.

2. <u>Row-swap Matrix</u>: The *row-swap matrix* – a permutation matrix – has determinant -1 and norm 1:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

3. <u>Row-scaling Matrix</u>: The *row-scaling matrix* has determinant $m$ – this example

$$m = \frac{1}{3}$$

and norm $max(1, n)$

$$\begin{bmatrix} 1 & 0 \\ 0 & m \end{bmatrix}$$

4. <u>Row-update Matrix</u>: The *row-update matrix* – a shear matrix – has determinant 1 and a norm which can most easily visualize using the diagrams above:
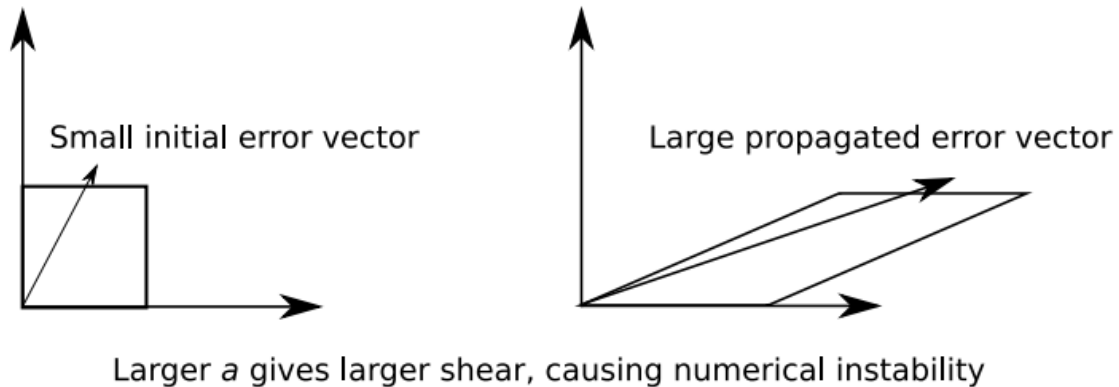
$$\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}.$$

In this example

$$a = -2$$

5. <u>Gaussian Elimination's Magnification of Error</u>: It is clear that Gaussian elimination magnifies error when a row-scaling matrix has a large $|m|$, and/or when a row-update matrix has a large $|a|$.



Larger *a* gives larger shear, causing numerical instability

## Gaussian Elimination without Pivoting

1. <u>Naïve Application of Gaussian Elimination</u>: Here is an example of the error that accumulates when one naively uses Gaussian elimination. The solution to the linear system

$$y = 1$$

$$x - y = 0$$

is, in the matrix form

$$\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

or, in the augmented form

$$\left[\begin{array}{cc|c} 0 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right].$$

2. <u>Solution to the Linear System</u>: Clearly, the solution is

$$\left[\begin{array}{c} x \\ y \end{array}\right] = \left[\begin{array}{c} 1 \\ 1 \end{array}\right]$$

3. <u>Perturbation of the Augmented Matrix</u>: Perturbation to the linear system above results in solving the system

$$\left[\begin{array}{cc|c} 0.001 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right]$$

4. <u>Gaussian Elimination with Rounding Applied</u>: One would expect a solution close to $(1, 1)$. The Gaussian elimination with no pivoting operations – rounded to 4 and 3 decimal places, respectively – is shown below.

5. <u>Intermediate Stages of the Matrix Transform</u>: Each step in the picture includes the transformation matrix to the next step, along with the norm of the transformation matrix

|  | 4 places | | 3 places | |
|---|---|---|---|---|

<div style="display:flex">

**4 places**

(norm 1000) $\begin{bmatrix} 1000 & 0 \\ 0 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 0.001 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right]$

(norm $\approx$ 1) $\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 1 & -1 & 0 \end{array}\right]$

(norm $=$ 1) $\begin{bmatrix} 1 & 0 \\ 0 & -1/1001 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 0 & -1001 & -1000 \end{array}\right]$

(norm $=$ 1000) $\begin{bmatrix} 1 & -1000 \\ 0 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 0 & 1 & 0.999 \end{array}\right]$

$\left[\begin{array}{cc|c} 1 & 0 & 0.999 \\ 0 & 1 & 0.999 \end{array}\right]$

**3 places**

(norm 1000) $\begin{bmatrix} 1000 & 0 \\ 0 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 0.001 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right]$

(norm $\approx$ 1) $\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 1 & -1 & 0 \end{array}\right]$

(norm $=$ 1) $\begin{bmatrix} 1 & 0 \\ 0 & -1/1000 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 0 & -1000 & -1000 \end{array}\right]$

(norm $=$ 1000) $\begin{bmatrix} 1 & -1000 \\ 0 & 1 \end{bmatrix}$ $\left[\begin{array}{cc|c} 1 & 1000 & 1000 \\ 0 & 1 & 1 \end{array}\right]$

$\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 1 \end{array}\right]$

</div>

6.  Exaggeration of the Round-off Effect: Here, the round-off error dominates; the right-hand solution is very wrong, and the left-hand solution is nearly correct. There were two norm-1000 operations, contributing to the error.

## Gaussian Elimination with Pivoting

1.  Pivoting Applied on Gaussian Elimination: This section re-does the sample, but with pivoting. This means that a permutation matrix is used to place the largest absolute-value column head at the top.
2.  3/4 Decimal Place Rounding: Again, use of 4 and 3 decimal places, respectively, produces the following figure.

| 4 places | | | 3 places | | |
|---|---|---|---|---|---|
| (norm 1) | $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 0.001 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right]$ | (norm 1) | $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 0.001 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right]$ |
| (norm $\approx 1$) | $\begin{bmatrix} 1 & 0 \\ -1/1000 & 1 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0.001 & 1 & 1 \end{array}\right]$ | (norm $\approx 1$) | $\begin{bmatrix} 1 & 0 \\ -1/1000 & 1 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0.001 & 1 & 1 \end{array}\right]$ |
| (norm $\approx 1$) | $\begin{bmatrix} 1 & 1/1.001 \\ 0 & 1 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 1.001 & 1 \end{array}\right]$ | (norm $\approx 1$) | $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 1 & 1 \end{array}\right]$ |
| (norm $\approx 1$) | $\begin{bmatrix} 1 & 0 \\ 0 & 1/1.001 \end{bmatrix}$ | $\left[\begin{array}{cc|c} 1 & 0 & 0.999 \\ 0 & 1.001 & 1 \end{array}\right]$ | | | $\left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 1 \end{array}\right]$ |
| | | $\left[\begin{array}{cc|c} 1 & 0 & 0.999 \\ 0 & 1 & 0.999 \end{array}\right]$ | | | |

3. <u>Solutions Approximately Equal and Correct</u>: Here, both solutions are approximately equal and approximately correct. All transformations have norm of the order of 1, which in turn is the case because of the pivoting operation.

4. <u>Pivoting - a Data-dependent Decision</u>: The point is that successful Gaussian elimination *requires* pivoting, which is a *data-dependent decision*. Also, the *permutations* must be stored. Neither of these is a serious detriment in a carefully implemented software system.

5. <u>Solution using Householder Transformation</u>: The next few sections examine Householder transformations not as a *necessity*, but as an *alternative*. It turns out that they will permit somewhat simpler software implementation.
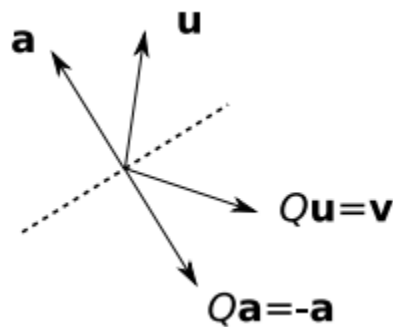
## Householder Transformations

The next few sections the Householder is derived, and then put to use in applications.

## Geometric Construction

1. Construction of the Reflection Matrix: One may construct a *reflection matrix* $Q$ acting on $\mathbb{R}^n$ which sends a chosen *axis vector*, $a$, to its negative, and reflects all other vectors through the hyperplane perpendicular to $a$:



2. Decomposition of $u$ into Components: A given matrix $u$ may be decomposed into its components which are parallel and perpendicular to $a$:
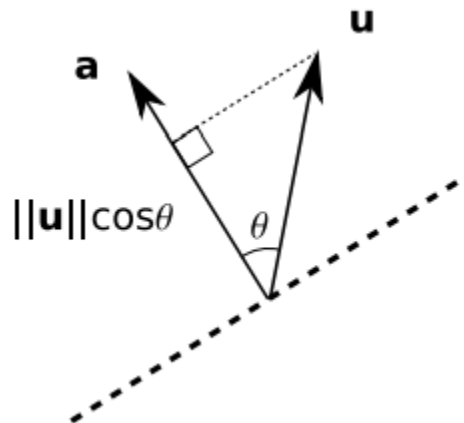


3. Geometric View of the Reflection: Then, if one can subtract off from $u$ twice its parallel component $u_{||}$, then the following reflection is obtained:
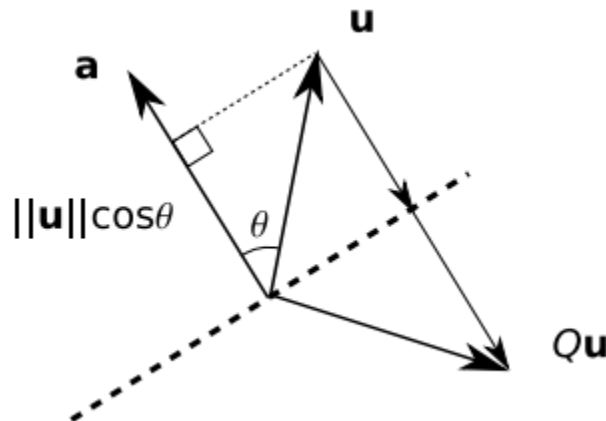
4. <u>Formulation of the Matrix $Q$</u>: That is, one wants

$$Qu = u - 2u_{||}$$

Moreover, one would like a single matrix $Q$ which would transform all $u$'s.

5. <u>Formulation of the Projection Vector</u>: The projection vector $u_{||}$ must be in the *direction* of $a$, and it must have the magnitude $\|u\| \cos \theta$:

$$u_{||} = \hat{a} \, \|u\| \cos \theta = \frac{a}{\|a\|} \|u\| \cos \theta$$

6. <u>Unit Vector of $a$</u>: The following notation is used:

$$\hat{a} = \frac{a}{\|a\|}$$

7. <u>Projection Vector using Dot Product #1</u>: It turns out – this is the central trick of the Householder transformation – that one can reformulate this expression in terms of dot products, eliminating $\cos \theta$. To do this, recall that

$$a \cdot u = \|a\| \, \|u\| \cos \theta$$

So

$$u_{\parallel} = \frac{a}{\|a\|^2} \|a\| \, \|u\| \cos\theta = \frac{a}{\|a\|^2} \, a \cdot u$$

8. <u>Square Modulus of $a/u$</u>: Also recall that

$$\|a\|^2 = a \cdot a$$

and

$$\|u\|^2 = u \cdot u$$

9. <u>Projection Vector using Dot Product #2</u>: Now, one has

$$u_{\parallel} = a \, \frac{a \cdot u}{a \cdot a}$$

10. <u>Reflection Matrix acting on $u$</u>: One can use this to re-write the reflection of $u$:

$$Qu = u - 2u_{\parallel} = u - 2a \frac{a \cdot u}{a \cdot a}$$

11. <u>Inner Product as a Matrix Product</u>: Recall that the dot-product – or *inner product* – can be written as a matrix product:

$$a \cdot a = a^T a$$

12. <u>Setting $Q$ using Matrix Transform</u>: One may think of a column vector as an $n \times 1$ matrix, and the row vector as an $1 \times n$ matrix, which is the transpose of the column vector. So

$$Qu = u - 2a\frac{a^T u}{a^T a}$$

13. <u>Outer Product as Matrix Multiplication</u>: Now that the dot products can be expressed in terms of matrix multiplication, one may use the associativity matrix multiplication:

$$a(a^T u) = (aa^T)u$$

14. <u>Statement of the Householder Transformation</u>: The product $aa^T$ is the *outer product* of $a$ with itself. It is an $n \times n$ matrix with ijth entry $a_i a_j$. One has

$$Qu = \left[I - 2\frac{a^T u}{a^T a}\right]u$$

from which

$$Q = I - 2\frac{a^T u}{a^T a}$$

## Construction with Specified Source and Destination

1. <u>Locating the Householder Reflection Axis</u>: The next step is, given $v$ and $u$, one wants to find the axis of reflection $a$ through which $v$ is $u$'s mirror image. Since Householder transformations are norm-preserving, $u$ and $v$ have same norm. As shown below, using the head-to-tail method

$$a = u - v$$

might be a candidate solution.
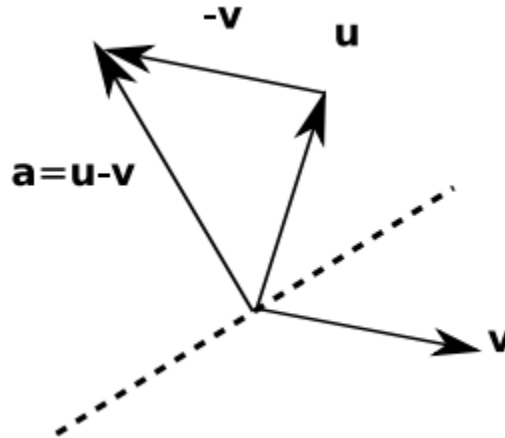


2. Verification of Householder Reflection Axis: It may be checked easily that this is correct. If $a$ is indeed that reflection axis, then it ought to get reflected to its own negative:

$$Q(a) = Q(u - v) = Q(u) - Q(v) = v - u = -a$$

## Properties of $Q$, Obtained Algebraically

1. Algebraic Proof of Householder Properties: This section attempts to prove algebraically certain properties of the Householder transform. The next section does the same geometrically.
2. Orthogonal: i.e.

$$QQ^T = I$$

Applying $I - 2\frac{a^T u}{a^T a}$ on itself proves this.

3. <u>Symmetric</u>: $Q$ is *symmetric*, i.e.

$$Q = Q^T$$

This is easily seen since

$$Q = I - 2\frac{a^T u}{a^T a}$$

This identity is symmetric, $aa^T$ has ij<sup>th</sup> entry

$$a_i a_j = a_j a_i$$

and so is symmetric as well.

4. <u>Involutary</u>: $Q$ is *involutary*, i.e.

$$Q^2 = I$$

Same as orthogonality, due to symmetry, since

$$Q = Q^T$$

5. <u>Determinant</u>: No easy way to find this algebraically.
6. <u>Matrix Norm</u>: No easy way to find this algebraically.

## Properties of Q, Obtained Geometrically

1. <u>Geometric Proof of Householder Properties</u>: In this section, certain properties of Householder matrix using geometric methods are proven. Remember that dot products, via

$$a \cdot u = \|a\| \|u\| \cos\theta$$

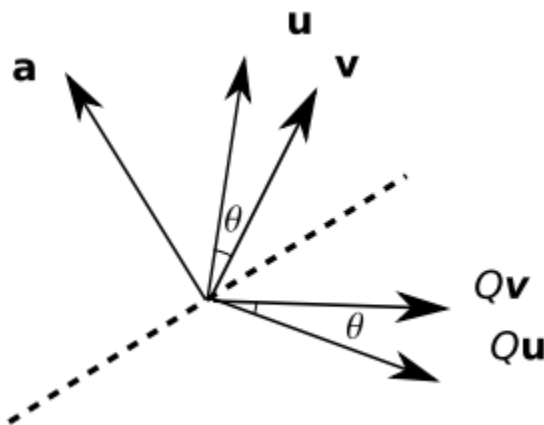provides with norms as well as angles.

2. $Q$ is *Involutary*: i.e.

$$Q^2 = I$$

$Q$ reflects $u$ to its mirror image $v$; a second application of $Q$ sends it back again.

3. $Q$ is *Orthogonal*: i.e.

$$(Qu \cdot Qv) = u \cdot v$$

for all vectors $u$, $v$. Since $Q$ is a reflection, it preserves norm; also, from the picture below, it's clear that it preserves angles.



4. $Q$ is *Symmetric*: i.e.

$$(Qu) \cdot v = u \cdot (Qv)$$

for all vectors $u$, $v$. This is the same geometric argument as for orthogonality, making use of involuntary.

5. <u>Determinant</u>: Since $Q$ is a reflection about the $a$ axis, leaving all the axes orthogonal to a fixed, $Q$ must have determinant $-1$. That is, it turns the unit cube inside out along one axis.

6. <u>Matrix Norm</u>: Since $Q$ is a reflection, it is length-preserving.

## Repeater Householders for Upper-Triangularization

1. <u>Purpose of Upper Triangular Transformation</u>: The goal of upper triangularization is to put a matrix of the form

$$\begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix}$$

into the form

$$\begin{bmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \end{bmatrix}$$

2. <u>One Step as a Time</u>: This may be one step at a time: from

$$\begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix}$$

to

$$\begin{bmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \end{bmatrix}$$

to

$$\begin{bmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \end{bmatrix}$$

3. <u>First Step in the Transformation</u>: The first step – for which Householder transformation will be used – is on all of the $m \times n$ input matrix entries, with an operation which transforms the first column.

4. <u>Second Step in the Transformation</u>: The second step is on the $(m - 1) \times (n - 1)$ submatrix obtained by omitting the top row and the left column:

$$\left[\begin{array}{c|ccccc} * & * & * & * & * & * \\ \hline 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \end{array}\right]$$

5. <u>Further Steps in the Transformation</u>: One can keep operating on the sub-matrices until there are no more of them left. So, the problem if upper-triangularization reduces to

that of modifying the first column of a matrix to have non-zero entries except the entries at the top of the column.

## Householders for Column-Zeroing

1. <u>Final Objective behind the Transform</u>: The goal is to put the matrix

$$A = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix}$$

into the form

$$QA = \begin{bmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \end{bmatrix}$$

2. <u>Application of $Q$ on $A$</u>: The key insight is that when $Q$ acts on $A$ by left multiplication, the j$^{th}$ column of $QA$ is the matrix-times-vector product of $Q$ times the j$^{th}$ column of $A$.

3. <u>Zeroing of First $A$ Column</u>: Let $u$ be the first column of $A$, and $v$ be the first column of $QA$. At this point all we know is that we want $u$ to have all zeros except for the first entry.

4. <u>Use of the Norm-preserving Nature</u>: Since this section is going to use Householder transformations, which are norm-preserving, one knows that

$$\|u\| = \|v\|$$

This means

$$v_1 = \pm\|u\|$$

5.  <u>Determining the Householder Matrix $Q$</u>: One wants a Householder matrix $Q$ which sends $u$ to $v$ and which does what it does with the rest of the matrix as a side effect.
6.  <u>Establishing the Reflection Axis $A$</u>: As seen above the reflection axis $a$ is

$$a = u - v$$

Further, since

$$v_1 = \pm\|u\|$$

and since the task is to compute

$$a = u - v$$

one can choose $v_1$ to have the opposite sign of $u_1$ to avoid the cancellation error that happens when one subtracts two nearly equal numbers.

7.  <u>Error Analysis</u>: Since the $Q$'s are orthogonal, they are norm-preserving, so

$$Q(u + \varepsilon) = Q(u) + Q(\varepsilon)$$

but

$$\|Q(\varepsilon)\| \leq \|\varepsilon\|$$

as discussed earlier.

8. <u>No Pivoting/Data- dependent Decisions</u>: Also, there is no pivoting involved, and thus – other than the choice of the sign of $v_1$ – there are no *data-dependent decisions*.

## Computation of Determinants

1. <u>Recap - Determinant of a Householder Matrix</u>: Given a square matrix $A$, one can use repeated Householder transformations to turn it into an upper-triangular matrix $U$. As discussed in the previous sections, $\det A$ is the product of the diagonal entries of $U$ divided by the product of the determinants of the transformation matrix.

2. <u>Sign of the Composite Determinant</u>: However, as seen in the earlier sections, Householder matrices have a determinant $-1$. Thus, one just has to track whether the number of Householders are odd or even. It takes $n-1$ Householders to upper-triangularize an $n \times n$ matrix:

$$\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix}$$

Therefore

$$\det A = (-1)^{n-1} \det U$$

## Computation of Matrix Inverses

1. <u>Use of Householder Upper-Triangularization</u>: Given and invertible square matrix $A$, one can make an augmented matrix just as seen in the earlier section. The

Householder upper-triangularization can be used to put the augmented matrix into the upper-triangular form, i.e., row echelon form.

2. <u>Clearing out the Above-diagonal Entries</u>: The rest of the work is done in clearing out the above-diagonal entries of the left-hand side of the augmented matrix, for which the scale and the shear matrices are used, as seen above.

## Rotation Matrices

1. <u>Connected Components of Householder Matrices</u>: $Q(n)$ has two connected components: $SO(n)$ and $O^-(n)$. The former are rigid *rotations* with determinants $+1$; the latter are *reflections* with determinant $-1$. Simon (1995) and Frenkel (2011) contains more information on $SO(n)$.

2. <u>Simplified Technique to Compute $SO(n)$</u>: The natural question is: If the Householder transformations provide easily understandable, readily computable elements of $O^-(n)$, then is there another equally pleasant technique to compute elements of $SO(n)$? In fact, there is. One can consult the literature for the Rodrigues formula, and modify that.

3. <u>Rotation as Product of Reflections</u>: Alternatively, one can use the fact that since determinants multiply, the product of two reflections, each having determinants -1, will have determinant +1. Thus, the product will be a rotation.

4. <u>Product of Two Identical Reflections</u>: If a given reflection is composed with itself, the product will be an identity, as seen earlier.

5. <u>Product of Two Distinct Reflections</u>: However, the product of two *distinct* reflections will be a non-trivial rotation.

6. <u>Geometric Visualization of the Reflection Sequence</u>: Let $u$ and $w$ be given. Since rotations are norm-preserving, one can only hope to rotate $u$ into $w$ if they have the same norm. This may be helped by computing $\hat{u}$ and $\hat{w}$. Note that $\hat{u} + \hat{w}$ will lie halfway between $\hat{u}$ and $\hat{w}$.
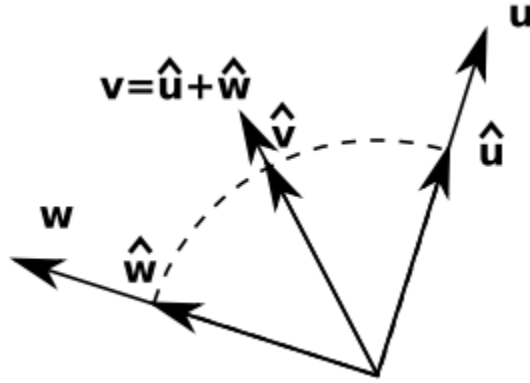
7. <u>Rotation Resulting from Two Householders</u>: Let

$$v = \hat{u} + \hat{w}$$

and compute $v$. The let $Q_1$ reflect $\hat{u}$ to $\hat{v}$. Just as in an earlier section, this is a reflection about the axis

$$a_1 = \hat{u} - \hat{v}$$

Likewise for the second reflection. One then has

$$R = Q_1 Q_2 = \left[ I - 2 \frac{a_2 a_2^T}{a_2^T a_2} \right] \left[ I - 2 \frac{a_1 a_1^T}{a_1^T a_1} \right]$$

**References**

- Burden, R. L., J. D. Faires, and A. M. Burden (2015): *Numerical Analysis 10th Edition* **Cengage Learning** Boston, MA
- Frankel, T. (2011): *The Geometry of Physics: An Introduction 3rd Edition* **Cambridge University Press** Cambridge, UK

- Friedberg, S. H., A. J. Insel, and L. E. Spence (2002): *Linear Algebra 4$^{th}$ Edition* **Pearson** London, UK

- Kerl, J. (2008): [The Householder Transformation in Numerical Linear Algebra](#)

- Simon, B. (1995): *Representations of Finite and Compact Groups* **American Mathematical Society** Providence, RI

# Successive Over-relaxation

## Introduction

The method of successive over-relaxation is a variant of the Gauss-Seidel method for solving a linear system of equations, resulting in faster convergence. A similar method can be used for any slowly converging iterative process (Wikipedia (2024)).

## Formulation

1. <u>Linear System LHS and RHS</u>: Given a square system of $n$ linear equations with unknown $\vec{x}$

$$A\vec{x} = \vec{b}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{21} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

2. <u>Decomposing $A$ into Strictly Upper/lower Triangular Matrices</u>: Then $A$ can be decomposed into a diagonal component $D$ and strictly lower and upper triangular components $L$ and $U$

$$A = D + L + U$$

where

$$D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{21} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

3. <u>Recasting the Linear System using a Relaxation Parameter</u>: The system of linear equations can be rewritten as

$$(D + \omega L)\vec{x} = \omega b - [\omega U + (\omega - 1)D]\vec{x}$$

for a constant

$$\omega > 1$$

called the *relaxation factor*.

4. <u>SOR Based Iterative LHS Solution</u>: The method of successive over-relaxation is an iterative technique that solves the LHS of the expression for $\vec{x}$ using the previous value for $\vec{x}$ on the RHS.

5. <u>Analytical Expression for the Iterator</u>: Analytically, this may be written as

$$\vec{x}_{(k+1)} = (D + \omega L)^{-1}\left\{\omega\vec{b} - [\omega U + (\omega - 1)D]\vec{x}_{(k)}\right\} = L_\omega \vec{x}_{(k)} + \vec{c}$$

where $\vec{x}_{(k)}$ is the $k^{th}$ approximation or iteration of $\vec{x}$ and $\vec{x}_{(k+1)}$ is the next or $k + 1$ iteration of $\vec{x}$.

6. <u>Forward Substitution for the Triangular Recast</u>: However, taking advantage of the triangular form of $D + \omega L$, the elements of $\vec{x}_{(k+1)}$ can be computed sequentially using forward substitution:

$$\vec{x}_{i,(k+1)} = (1 - \omega_i)\vec{x}_{i,(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}\vec{x}_{j,(k+1)} - \sum_{j>i} a_{ij}\vec{x}_{j,(k)}\right)$$

$$i = 1, 2, \cdots, n$$

7. <u>Compact Matrix-Vector Form Expression</u>: This can again be written analytically in matrix-vector form without need for inverting the matrix $D + \omega L$ (Tornig (1979)):

$$\vec{x}_{(k+1)} = (1 - \omega)\vec{x}_{(k)} + \omega D^{-1}\left(\vec{b} - L\vec{x}_{(k+1)} - U\vec{x}_{(k)}\right)$$

**Convergence**

1. Choice of the Relaxation Factor: The choice of the relaxation factor is not necessarily easy, and depends on the properties of the coefficient matrix.

2. Ostrowski Convergence for Positive Semi-definite $A$: In 1947, Ostrowski proved that if $A$ is symmetric and positive-definite then

$$\rho(L_\omega) < 1$$

for

$$0 < \omega < 2$$

Thus, convergence of the iteration process follows, but is generally more interested in faster convergence.

## Convergence Rate

1. Analytical Derivation of the Convergence: The convergence rate for the SOR method can be analytically derived. The following assumptions need to be made (Greenbaum (1997), Hackbusch (2016)).

2. Assumption #1 - Relaxation Parameter: The relaxation parameter is appropriate, i.e.,

$$\omega \in (0, 2)$$

3. Assumption #2 - Jacobi's Iteration Matix:

$$C_{Jacobi} := I - D^{-1}A$$

has only real eigen-values

4. Assumption #3 - Convergent Jacobi's Matrix: Jacobi's matrix is convergent:

$$\mu := \rho\left(C_{Jacobi}\right) < 1$$

5. Assumption #4 - Matrix Decomposition Criterion: The matrix decomposition

$$A = D + L + U$$

satisfies the property that

$$det\left(\lambda D + zL + \frac{1}{z}U\right) = det(\lambda D + L + U)$$
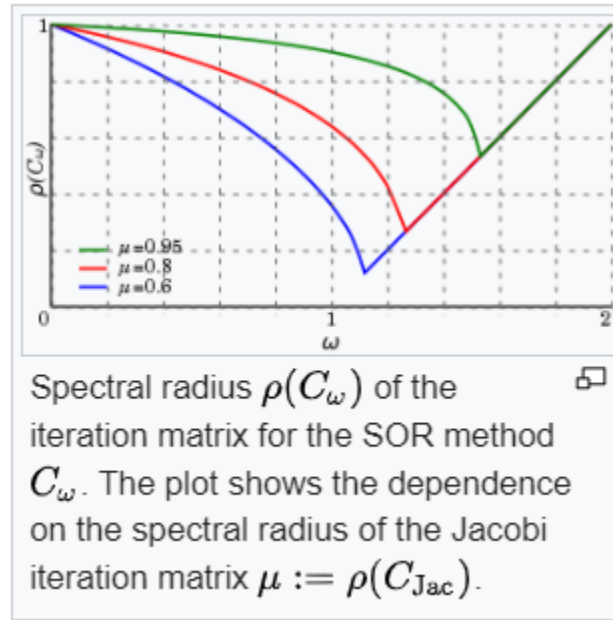
for any

$$z \in \mathbb{C}\backslash\{0\}$$

and

$$\lambda \in \mathbb{C}$$

6. Expression for the Convergence Rate:

Spectral radius $\rho(C_\omega)$ of the
iteration matrix for the SOR method
$C_\omega$. The plot shows the dependence
on the spectral radius of the Jacobi
iteration matrix $\mu := \rho(C_{\text{Jac}})$.

Then the convergence rate can be expressed as

$$\rho(C_\omega) = \begin{cases} \frac{1}{4}\left[\omega\mu - \sqrt{\omega^2\mu^2 - 4(\omega - 1)}\right]^2 & 0 < \omega \le \omega_{Opt} \\ \omega - 1 & \omega_{Opt} < \omega < 2 \end{cases}$$

where the optimal relaxation parameter is given by

$$\omega_{Opt} := 1 + \left(\frac{\mu}{1 + \sqrt{1 - \mu^2}}\right)^2 = 1 + \frac{\mu^2}{4} + \mathcal{O}(\mu^3)$$

7. Jacobi Spectral Radius for Gauss-Seidel Scheme: In particular, for

$$\omega = 1$$

– Gauss-Seidel – it holds that

$$\rho(C_\omega) = \mu^2 = \rho^2(C_{Jacobi})$$

430

8. Optimal $\omega$ Jacobi Spectral Radius: For the optimal $\omega$, one gets

$$\rho(C_\omega) = \frac{1 - \sqrt{1 - \mu^2}}{1 + \sqrt{1 - \mu^2}} = \frac{\mu^2}{4} + \mathcal{O}(\mu^3)$$

which shows that SOR is roughly 4 times more efficient than Gauss-Seidel.

9. Fulfillment of the Matrix Decomposition Criterion: The final assumption is satisfied for tridiagonal matrices since

$$Z(\lambda D + L + U)Z^{-1} = \lambda D + ZL + \frac{1}{Z}U$$

for diagonal $Z$ with entries

$$Z_{ii} = z^{i-1}$$

and

$$det(\lambda D + L + U) = det[Z(\lambda D + L + U)Z^{-1}]$$

## Algorithm

1. Optimization Inside of Iteration Loop: Note that $(1 - \omega)\phi_i + \frac{\omega}{a_{ii}}(b_i - \sigma)$ can also be written as $\phi_i + \omega\left(\frac{b_i - \sigma}{a_{ii}} - \phi_i\right)$ thus saving one multiplication is each iteration of the outer for loop.

2. <u>Pseudocode for the Algorithm</u>: Since elements can be over-written as they are computed in this algorithm, only one storage vector is needed, and vector indexing is omitted in the algorithm below.

```
Inputs: A, b, ω
Output: φ

Choose an initial guess φ to the solution
repeat until convergence
    for i from 1 until n do
        set σ to 0
        for j from 1 until n do
            if j ≠ i then
                set σ to σ + aᵢⱼ φⱼ
            end if
        end (j-loop)
        set φᵢ to (1 − ω)φᵢ + ω(bᵢ − σ) / aᵢᵢ
    end (i-loop)
    check if convergence is reached
end (repeat)
```

## Symmetric Successive Over-relaxation

The SOR version for symmetric matrix $A$, in which

$$U = L^T$$

is referred to as Symmetric Successive Over-relaxation, or SSOR, in which

$$P = \left(\frac{D}{\omega} + L\right)\frac{\omega}{2 - \omega}D^{-1}\left(\frac{D}{\omega} + U\right)^{-1}$$

and the iterative method is

$$\vec{x}_{(k+1)} = \vec{x}_{(k)} - \gamma^k P^{-1}\left(A\vec{x}_{(k)} - \vec{b}\right)$$

$$k \geq 0$$

## Other Applications of the Method

1.  <u>Extension across Iterative Methods</u>: A similar technique can be used for any iterative method. If the original iteration has the form

$$x_{n+1} = f(x_n)$$

then the modified version would use

$$x_{n+1,SOR} = (1 - \omega)x_{n,SOR} + \omega f(x_{n,SOR})$$

2.  <u>Extension to Vector Unknowns</u>: However, the formulation presented above, used for solving systems of linear equations, is not a special case of this formulation if $x$ is considered to be a complete vector. If this formulation is used instead, the equation for calculating the next vector will look like

$$\vec{x}_{(k+1)} = (1 - \omega)\vec{x}_{(k)} + \omega L_*^{-1}\left(\vec{b} - U\vec{x}_{(k)}\right)$$

where

$$L_* = L + D$$

3.  <u>Outcome Control using Relaxation Parameter</u>: Values of

$$\omega > 1$$

are used to speed up convergence of a slow converging process, while values of

$$\omega < 1$$

are often used to establish convergence of a diverging iterative process or speed up convergence of an overshooting process.

4. <u>Adaptively Setting the Relaxation Parameter</u>: There are various methods that adaptively set the relaxation parameter $\omega$ based on the observed behavior of the converging process. Usually, they help to reach a super-linear convergence for some problems but fail for others.

## References

- Greenbaum, A. (1997): *Iterative Methods for Solving Linear Systems* **Society for Industrial and Applied Mathematics** Philadelphia, PA
- Hackbusch, W. (2016): *Iterative Solution of Large Sparse Systems of Equations* **Spring-Verlag** Berlin, Germany
- Tornig, W. (1979): *Numerische Mathematik fur Ingenieure und Physiker* **Spring-Verlag** Berlin, Germany
- Wikipedia (2024): <u>Successive Over-relaxation</u>

# Symmetric Successive Over-relaxation

1. <u>Constructing the SSOR Pre-conditioner Matrix</u>: The *Symmetric Successive Over-relaxation* (*SSOR*) step is used to generate the pre-conditioner. If the original matrix can be split into diagonal, lower, and upper triangular matrices as

$$A = D + L + L^T$$

   then the SSOR pre-conditioner matrix is defined as

$$M = (D + L)D^{-1}(D + L)^T$$

   (Wikipedia (2023)).

2. <u>Parametrization of the Pre-conditioner</u>: It can also be parametrized by $\omega$ as

$$M(\omega) = \frac{\omega}{2 - \omega}\left(\frac{D}{\omega} + L\right)D^{-1}\left(\frac{D}{\omega} + L\right)^T$$

## References

- Wikipedia (2023): [Symmetric Successive Over-relaxation](Symmetric Successive Over-relaxation)

# Tridiagonal Matrix Algorithm

## Introduction

1. <u>Thomas Algorithm for Tridiagonal Systems</u>: The *tridiagonal matrix algorithm*, also known as the *Thomas algorithm*, is a simplified form Gaussian elimination that can be used to solve tridiagonal system of equations (Wikipedia (2024)).

2. <u>Representation of a Tridiagonal System</u>: A tridiagonal system for $n$ unknowns may be written as

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i$$

where

$$a_1 = 0$$

and

$$c_n = 0$$

$$\begin{bmatrix} b_1 & c_1 & \square & \square & 0 \\ a_2 & b_2 & c_2 & \square & \square \\ \square & a_3 & b_3 & \ddots & \square \\ \square & \square & \ddots & \ddots & c_{n-1} \\ \square & \square & \square & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n \end{bmatrix}$$

3. <u>Algorithm Efficiency over Gaussian Elimination</u>: For such systems, the solution can be derived in $\mathcal{O}(n)$ operations instead of $\mathcal{O}(n^3)$ required by Gaussian elimination.

The first sweep eliminated the $a_i$'s, then an abbreviated backward substitution produces the solution.

4. <u>Examples of Triangular Matrix Formulations</u>: Examples of such matrices commonly arise from the discretization of the 1D Poisson equation and natural cubic spline interpolation.

5. <u>Stability of the Tridiagonal Algorithm</u>: The algorithm is not stable in general, but is so in several special cases, such as when the matrix is diagonally dominant either by rows or columns, or symmetric positive definite (Niyogi (2006), Datta (2010)); for a more precise characterization of the stability of Thomas' algorithm, see Higham (2002).

6. <u>General Stability using GEPP Technique</u>: If stability is required in the general case, Gaussian elimination with Partial Pivoting GEPP is recommended instead (Datta (2010)).


## Method

1. <u>Re-sketched Coefficients during Forward Sweep</u>: The forward sweep consists of computation of new coefficients as follows, denoting new coefficients with primes:

$$
c_i' = \begin{cases} \dfrac{c_i}{b_i} & i = 1 \\ \dfrac{c_i}{b_i - a_i c_{i-1}'} & i = 2, 3, \dots, n-1 \end{cases}
$$

and

$$
d_i' = \begin{cases} \dfrac{d_i}{b_i} & i = 1 \\ \dfrac{d_i - a_i d_{i-1}'}{b_i - a_i c_{i-1}'} & i = 2, 3, \dots, n-1 \end{cases}
$$

2. <u>Solution by Back Substitution</u>: The solution is then obtained by back-substitution as:

$$x_n = d_n{}'$$

$$x_i = d_i' - c_i' x_{i+1}$$

$$i = n - 1, n - 2, \dots, 1$$

3. <u>Reduction of Inherent Storage Duplication</u>: The method above does not modify the original coefficients, but must also keep track of new coefficients. If the coefficient vectors may be modified, then an algorithm with less book-keeping can be used.

4. <u>Back substitution Algorithm for Space Optimization</u>: For

$$i = 2, 3, \dots, n - 1$$

do

$$w = \frac{a_i}{b_{i-1}}$$

$$b_i := b_i - w c_{i-1}$$

$$d_i := d_i - w d_{i-1}$$

followed by back substitution

$$x_n = \frac{d_n}{b_n}$$

$$x_i = \frac{d_i - c_i x_{i+1}}{b_i}$$

438

for

$$i = n - 1, \quad n - 2, \ldots, 1$$

5. <u>Illustrative C Routine with Optimal Space</u>: The implementation as a C function, which uses scratch space to avoid modifying its inputs $a - c$, allowing them to be re-used:

```c
void thomas(const int X, double x[restrict X],
            const double a[restrict X], const double b[restrict X],
            const double c[restrict X], double scratch[restrict X]) {
    /*
    solves Ax = d, where A is a tridiagonal matrix consisting of vectors a, b, c
    X = number of equations
    x[] = initially contains the input v, and returns x. indexed from [0, ..., X - 1]
    a[] = subdiagonal, indexed from [1, ..., X - 1]
    b[] = main diagonal, indexed from [0, ..., X - 1]
    c[] = superdiagonal, indexed from [0, ..., X - 2]
    scratch[] = scratch space of length X, provided by caller, allowing a, b, c to be const
    not performed in this example: manual expensive common subexpression elimination
    */
    scratch[0] = c[0] / b[0];
    x[0] = x[0] / b[0];

    /* loop from 1 to X - 1 inclusive */
    for (int ix = 1; ix < X; ix++) {
        if (ix < X-1){
        scratch[ix] = c[ix] / (b[ix] - a[ix] * scratch[ix - 1]);
        }
        x[ix] = (x[ix] - a[ix] * x[ix - 1]) / (b[ix] - a[ix] * scratch[ix - 1]);
    }

    /* loop from X - 2 to 0 inclusive */
    for (int ix = X - 2; ix >= 0; ix--)
        x[ix] -= scratch[ix] * x[ix + 1];
}
```

## Variants

1. <u>Altered Form of the Tridiagonal System</u>: In some situations, particularly those involving periodic boundary conditions, a slightly perturbed form of the tridiagonal

system may need to be solved: $a_i x_n + b_1 x_1 + c_1 x_2 = d_i$  $a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i$  $i = 2, 3, \ldots, n-1$  $a_n x_{n-1} + b_n x_n + c_n x_i = d_n$

2. <u>Using the Sherman Morrison Formula</u>: In this case, one can make use of the Sherman-Morrison formula to avoid additional operations of Gaussian elimination and still use the Thomas algorithm.

3. <u>Solving Modified Non-cyclic Systems</u>: The method requires solving a modified non-cyclic version of the system for both the input and the sparse corrective vector, and then combining the solutions.

4. <u>Simultaneous Efficient Solution of the Systems</u>: This can be done efficiently is both the solutions can be done at once, as the forward portion of the pure matrix algorithm can be shared.

5. <u>Setup of the Linear System</u>:
$$
\begin{bmatrix}
b_1 & c_1 & \square & \square & a_1 \\
a_2 & b_2 & c_2 & \square & \square \\
\square & a_3 & b_3 & \ddots & \square \\
\square & \square & \ddots & \ddots & c_{n-1} \\
c_n & \square & \square & a_n & b_n
\end{bmatrix}
\begin{bmatrix}
x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n
\end{bmatrix}
=
\begin{bmatrix}
d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n
\end{bmatrix}
$$
the system to be solved becomes $Ax = d$

6. <u>Incorporation of "Periodic" Boundary Condition Terms</u>: In this case, the coefficient $a_1$ and $c_n$ are, generally speaking, non-zero, so their presence does not allow application of the Thomas' algorithm directly.

7. <u>Re-constitution using Intermediate Vectors</u>: One therefore considers $B \in \mathbb{R}^{n \times n}$ and $u, v \in \mathbb{R}^n$ as follows:  where $B = \begin{bmatrix} b_1 - \gamma & c_1 & \square & \square & a_1 \\ a_2 & b_2 & c_2 & \square & \square \\ \square & a_3 & b_3 & \ddots & \square \\ \square & \square & \ddots & \ddots & c_{n-1} \\ c_n & \square & \square & a_n & b_n - \frac{c_n a_1}{\gamma} \end{bmatrix}$  $u = \begin{bmatrix} \gamma \\ 0 \\ 0 \\ \vdots \\ c_n \end{bmatrix}$  and

$u = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \frac{a_1}{\gamma} \end{bmatrix}$  where $\gamma \in \mathbb{R}$ is a parameter to be chosen. The matrix $A$ can be re-constructed as $A = B + uv^T$

8. <u>Re formulation of the Solution</u>: The solution is then obtained as follows (Batista and Ibrahim-Karawia (2009)); first, the two tridiagonal systems of equations are solved using the Thomas algorithm: $By = d \quad Bq = u$

9. <u>Synthesizing Solution using Sherman-Morrison Formula</u>: Finally, the solution $x$ is re-constructed using the Sherman-Morrison formula: $x = A^{-1}d = (B + uv^T)^{-1}d =$
$$B^{-1}d - \frac{B^{-1}uv^T B^{-1}}{1+v^T B^{-1}u}d = y - \frac{qv^T y}{1+v^T q}$$

10. <u>C Illustrative Implementation of Above</u>: The implementation as a C function, which uses scratch-space to avoid modifying its inputs $a - c$, allowing them to be used:

```c
void cyclic_thomas(const int X, double x[restrict X], const double a[restrict X], const double
b[restrict X], const double c[restrict X], double cmod[restrict X], double u[restrict X]) {
    /*
    solves Ax = v, where A is a cyclic tridiagonal matrix consisting of vectors a, b, c
    X = number of equations
    x[] = initially contains the input v, and returns x. indexed from [0, ..., X - 1]
    a[] = subdiagonal, regularly indexed from [1, ..., X - 1], a[0] is lower left corner
    b[] = main diagonal, indexed from [0, ..., X - 1]
    c[] = superdiagonal, regularly indexed from [0, ..., X - 2], c[X - 1] is upper right
    cmod[], u[] = scratch vectors each of length X
    */

    /* lower left and upper right corners of the cyclic tridiagonal system respectively */
    const double alpha = a[0];
    const double beta = c[X - 1];

    /* arbitrary, but chosen such that division by zero is avoided */
    const double gamma = -b[0];

    cmod[0] = alpha / (b[0] - gamma);
    u[0] = gamma / (b[0] - gamma);
    x[0] /= (b[0] - gamma);

    /* loop from 1 to X - 2 inclusive */
    for (int ix = 1; ix + 1 < X; ix++) {
        const double m = 1.0 / (b[ix] - a[ix] * cmod[ix - 1]);
        cmod[ix] = c[ix] * m;
        u[ix] = (0.0f  - a[ix] * u[ix - 1]) * m;
        x[ix] = (x[ix] - a[ix] * x[ix - 1]) * m;
    }

    /* handle X - 1 */
    const double m = 1.0 / (b[X - 1] - alpha * beta / gamma - beta * cmod[X - 2]);
    u[X - 1] = (alpha    - a[X - 1] * u[X - 2]) * m;
    x[X - 1] = (x[X - 1] - a[X - 1] * x[X - 2]) * m;

    /* loop from X - 2 to 0 inclusive */
    for (int ix = X - 2; ix >= 0; ix--) {
        u[ix] -= cmod[ix] * u[ix + 1];
        x[ix] -= cmod[ix] * x[ix + 1];
    }

    const double fact = (x[0] + x[X - 1] * beta / gamma) / (1.0 + u[0] + u[X - 1] * beta / gamma);

    /* loop from 0 to X - 1 inclusive */
    for (int ix = 0; ix < X; ix++)
        x[ix] -= fact * u[ix];
}
```

11. <u>Alternative Perturbation of Tridiagonal System</u>: There is also another way to solve the slightly perturbed form of the tridiagonal system considered above (Ryaben'kii and Tsynkov (2006)). Consider the two auxiliary linear systems of dimensions $(n-1) \times (n-1)$: $b_2 x_2 + c_2 x_3 = d_2$ $a_3 u_2 + b_3 u_3 + c_3 u_4 = d_3$ $a_i u_{i-1} + b_i u_i + c_i u_{i+1} = d_i$ for $i = 4, \dots, n-1$ $\cdots$ $a_n u_{n-1} + b_n u_n = d_n$ $b_2 v_2 + c_2 v_3 = -a_2$

$$a_3 v_2 + b_3 v_3 + c_3 v_4 = 0 \quad a_i v_{i-1} + b_i v_i + c_i v_{i+1} = 0 \quad \text{for } i = 4, \ldots, n-1 \quad \cdots$$
$$a_n v_{n-1} + b_n v_n = -c_n$$

12. <u>Solution to the above System</u>: For convenience, one additionally defines $u_1 = 0$ and $v_1 = 1$ The solution $\{u_2, u_3, \cdots, u_n\}$ and $\{v_2, v_3, \cdots, v_n\}$ may be found by applying the Thomas' algorithm to the two auxiliary tridiagonal systems.

13. <u>Form of the Solution Vector</u>: The solution $\{u_2, u_3, \cdots, u_n\}$ can then be represented in the form: $x_i = u_i + x_1 v_i \quad i = 1, 2, \ldots, n$

14. <u>Recovering Original Equations 2 through $n$</u>: Indeed, multiplying each equation of second auxiliary system by $x_1$, adding with the corresponding equation of the first auxiliary system and using the representation $x_i = u_i + x_1 v_i$ one immediately sees that equations 2 through $n$ of the original system are satisfied: it only remains to satisfy equation 1.

15. <u>Recovering Original Equation 1</u>: To do so, consider formula for $i = 2$ and $i = n$ and substitute $x_2 = u_{i2} + x_1 v_2$ and $x_n = u_n + x_1 v_n$ into the first equation of the original system. This yields one scalar equation for $x_1$: $b_1 x_1 + c_1 (u_2 + x_1 v_2) + a_1 (u_n + x_1 v_n) = d_1$

16. <u>Explicit Form for $x_1$</u>: As such, one finds that $x_1 = \frac{d_1 - a_1 u_n - c_1 u_n}{b_1 + a_1 v_n + c_1 v_n}$

17. <u>Sample Implementation using C</u>: The implementation as a C function, which uses scratch-space to avoid modifying its inputs for $a - c$, allowing them to be re-used:

```
void cyclic_thomas(const int X, double x[restrict X], const double a[restrict X], const double
b[restrict X], const double c[restrict X], double cmod[restrict X], double v[restrict X]) {
    /* first solve a system of length X - 1 for two right hand sides, ignoring ix == 0 */
    cmod[1] = c[1] / b[1];
    v[1] = -a[1] / b[1];
    x[1] = x[1] / b[1];

    /* loop from 2 to X - 1 inclusive */
    for (int ix = 2; ix < X - 1; ix++) {
        const double m = 1.0 / (b[ix] - a[ix] * cmod[ix - 1]);
        cmod[ix] = c[ix] * m;
        v[ix] = (0.0f  - a[ix] * v[ix - 1]) * m;
        x[ix] = (x[ix] - a[ix] * x[ix - 1]) * m;
    }

    /* handle X - 1 */
    const double m = 1.0 / (b[X - 1] - a[X - 1] * cmod[X - 2]);
    cmod[X - 1] = c[X - 1] * m;
    v[X - 1] = (-c[0]    - a[X - 1] * v[X - 2]) * m;
    x[X - 1] = (x[X - 1] - a[X - 1] * x[X - 2]) * m;

    /* loop from X - 2 to 1 inclusive */
    for (int ix = X - 2; ix >= 1; ix--) {
        v[ix] -= cmod[ix] * v[ix + 1];
        x[ix] -= cmod[ix] * x[ix + 1];
    }

    x[0] = (x[0] - a[0] * x[X - 1] - c[0] * x[1]) / (b[0] + a[0] * v[X - 1] + c[0] * v[1]);

    /* loop from 1 to X - 1 inclusive */
    for (int ix = 1; ix < X; ix++)
        x[ix] += x[0] * v[ix];
```

18. <u>Computational Complexity of the Tridiagonal Structure</u>: In both cases, the auxiliary systems to be solved are tridiagonal, so the overall computational complexity of solving $Ax = d$ remains linear with respect to the dimension of the system $n$, that is $\mathcal{O}(n)$ arithmetic operations.

19. <u>Block Tridiagonal Form</u>: In other situations, the systems of equations may be *block-diagonal* with smaller sub-matrices arranged as individual elements in the above matrix system, e.g., 2D Poisson problem. Simplified forms of Gaussian elimination have developed for these problems (Quarteroni, Sacco, and Saleri (2007)).

20. <u>Version that avoids Divisions</u>: Quarteroni, Sacco, and Saleri (2007) list a modified version of the algorithm that avoids some the divisions – using multiplications instead – which is beneficial on some computer architectures.

21. <u>Literature Treatment of Tridiagonal Algorithm</u>: Parallel tridiagonal solvers have been published for many vector and parallel architectures, including GOUs (Change and Hwu (2014)), Venetis, Kouris, Sobczyk, Gallopoulos, and Sameh (2015). For an extensive treatment of parallel tridiagonal and block diagonal solvers, see Gallopoulos, Phillippe, and Sameh (2016).
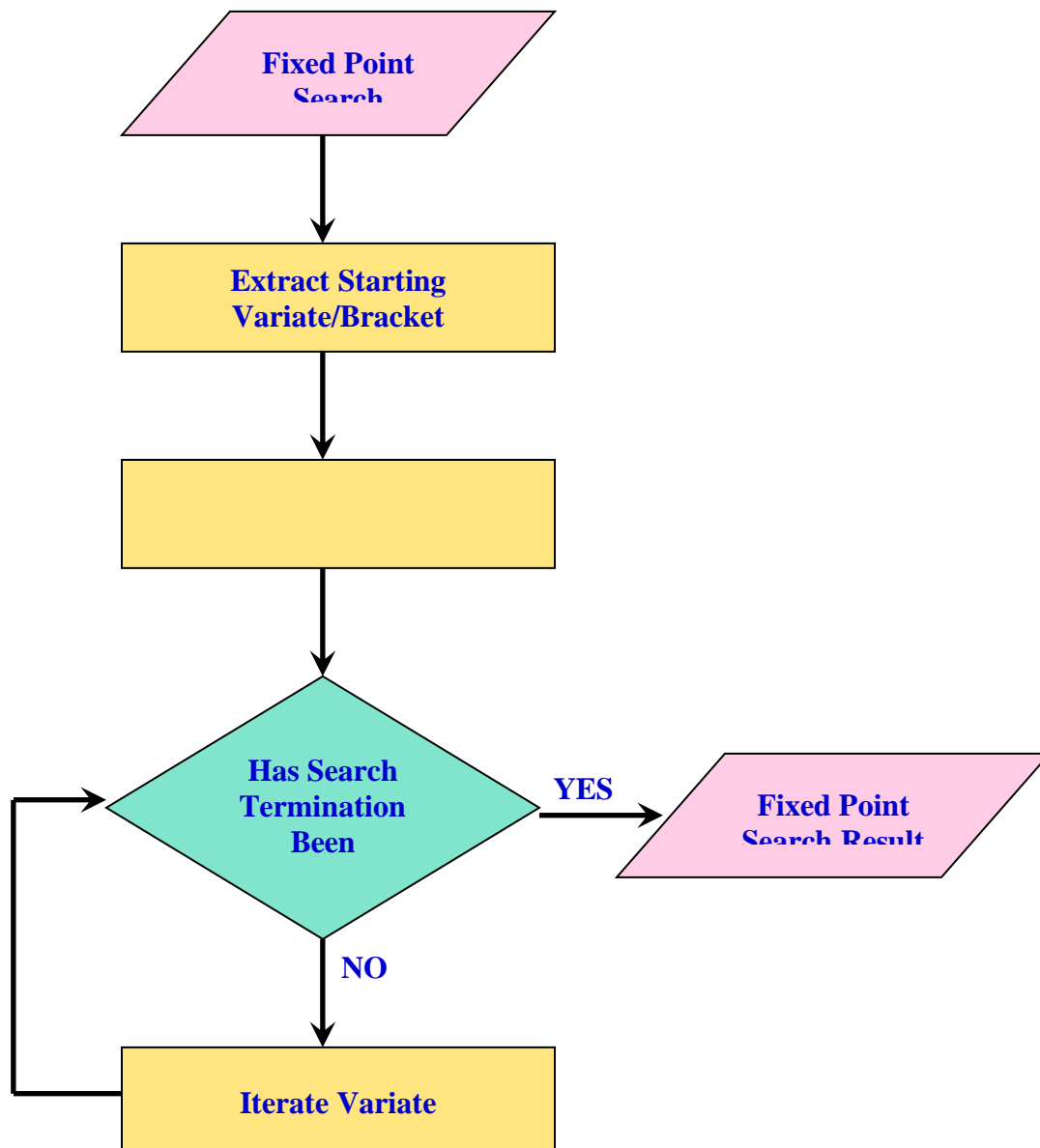
## References

- Batista, M., and A. R. A. Ibrahim-Karawia (2009): The use of Sherman-Morrison-Woodbury formula to solve cyclic block tridiagonal and cyclic block penta-diagonal linear systems of equations *Applied Mathematics of Computation* **210 (2)** 558-563

- Change, L. W., and W. M. Hwu (2014): "A Guide for Implementing Tridiagonal Solvers in GPUs", in: *Numerical Computations in CPUs (Editor - V. Kidratenko)* **Springer** Berlin, Germany

- Datta, B. N. (2010): *Numerical Linear Algebra and Applications 2$^{nd}$ Edition* **SIAM** Philadelphia, PA

- Gallopoulos, E., B. Phillippe, and A. H. Sameh (2016): *Parallelism in Matrix Computations* **Springer** Berlin, Germany

- Higham, N. J. (2002): *Accuracy and Stability of Numerical Algorithms 2$^{nd}$ Edition* **SIAM** Philadelphia, PA

- Niyogi, P. (2006): *Introduction to Computation Fluid Dynamics* **Pearson** London, UK

- Quarteroni, A., R. Sacco, and F. Saleri (2007): *Numerical Mathematics* **Springer** New York, NY

- Ryaben'kii, V. S., and S. V. Tsynkov (2006): *Theoretical Introduction to Numerical Analysis* **Wolters Kluwer** Aalphen aan den Rijn, Netherlands

- Venetis, I. E., A. Kouris, A. Sobczyk, E. Gallopoulos, and A. Sameh (2015): A direct tridiagonal solver based on Givens rotations for GPU architectures *Parallel Computing* **49 (C)** 101-116

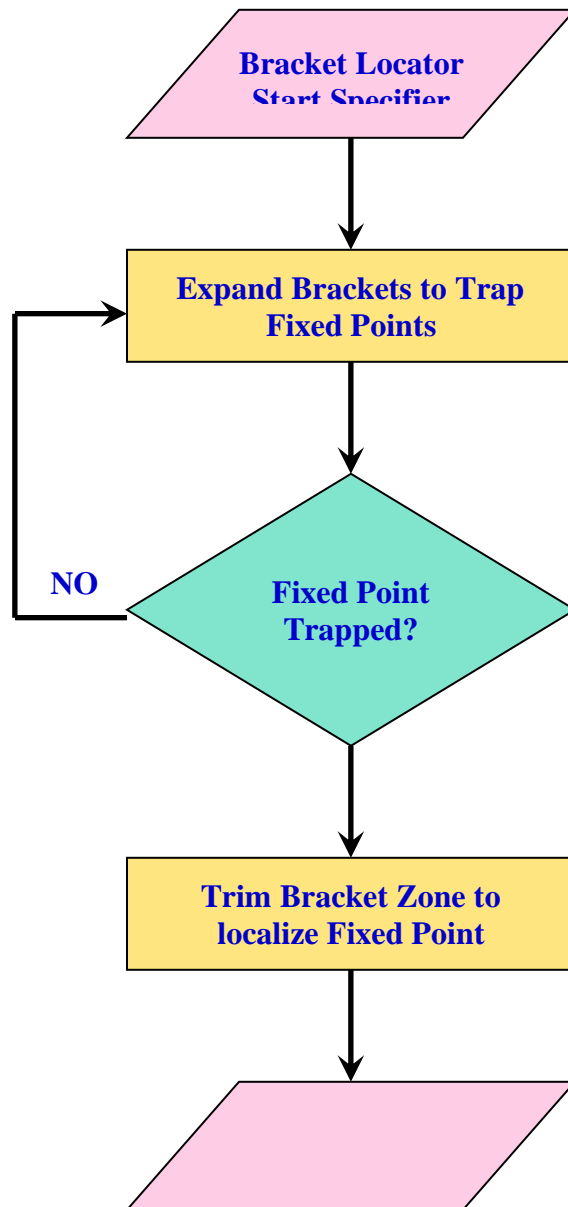- Wikipedia (2024): [Tridiagonal Matrix Algorithm](#)
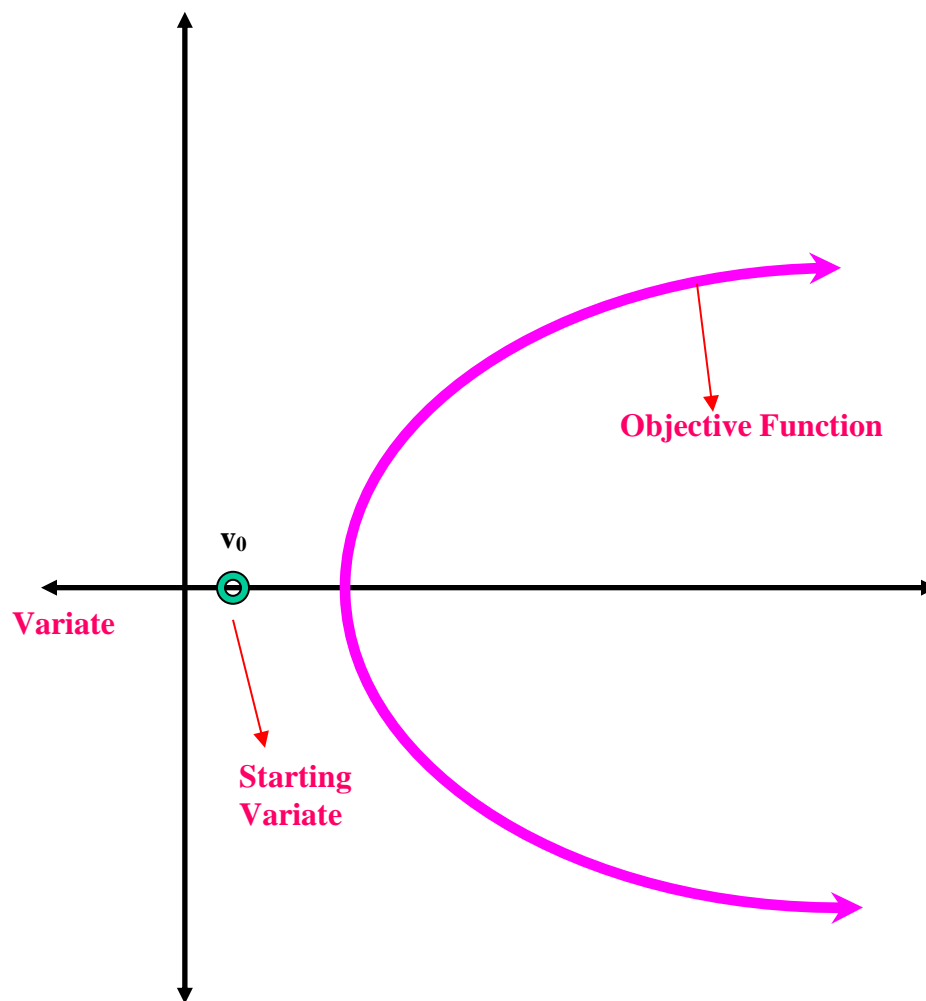
## Figure #1
## Fixed Point Search SKU Flow

Fixed Point
Search

Extract Starting
Variate/Bracket

Has Search
Termination
Been

**YES**

Fixed Point
Search Result

**NO**

Iterate Variate

## Figure #2
## Bracketing SKU Flow

Bracket Locator
Start Specifier

Expand Brackets to Trap
Fixed Points

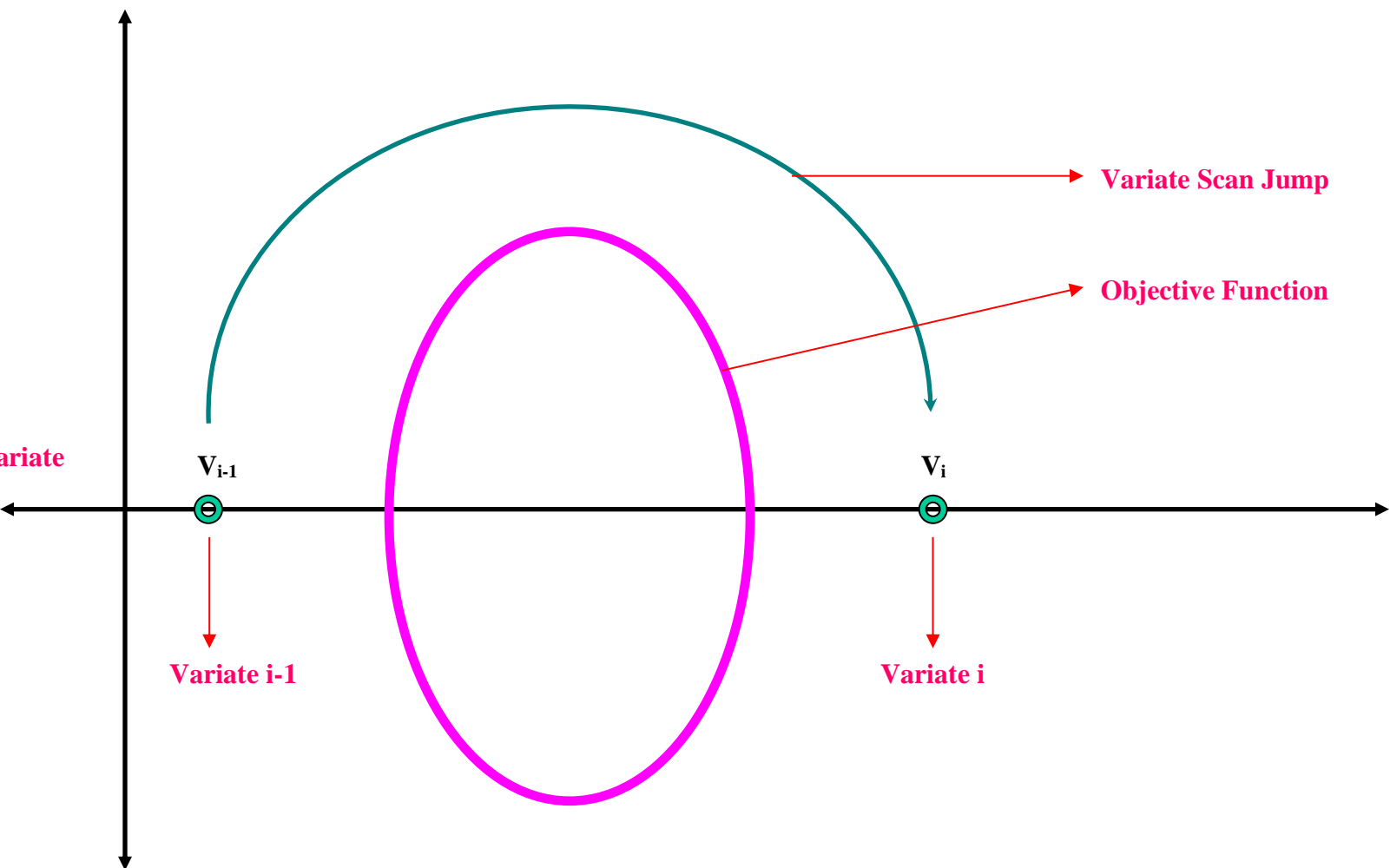Fixed Point
Trapped?

**NO**

Trim Bracket Zone to
localize Fixed Point

**Figure #3**
**Objective Function Undefined at the Starting Variate**

Figure #4
Objective Function Undefined at any of
the Candidate Variates

Variate Scan Jump
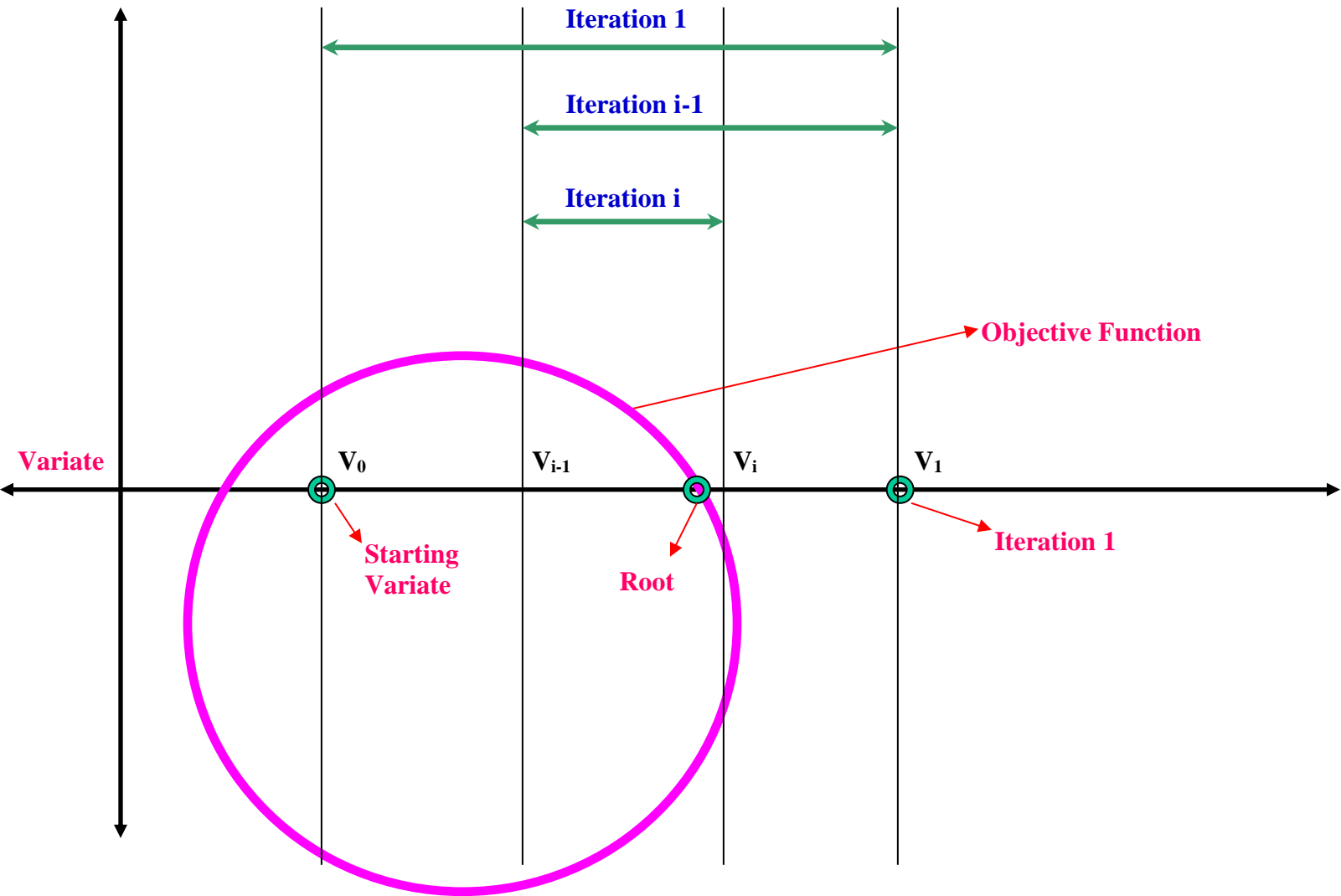
Objective Function

$V_{i-1}$

$V_i$

Variate i-1

Variate i

**Figure #5**
**General Purpose Bracket Start Locator**

Bracket Start Locator

Span the Left Variate

Is the Objective Function

NO

YES

Span the Right Variate

Is the Objective Function

YES

NO

Found Starting Point
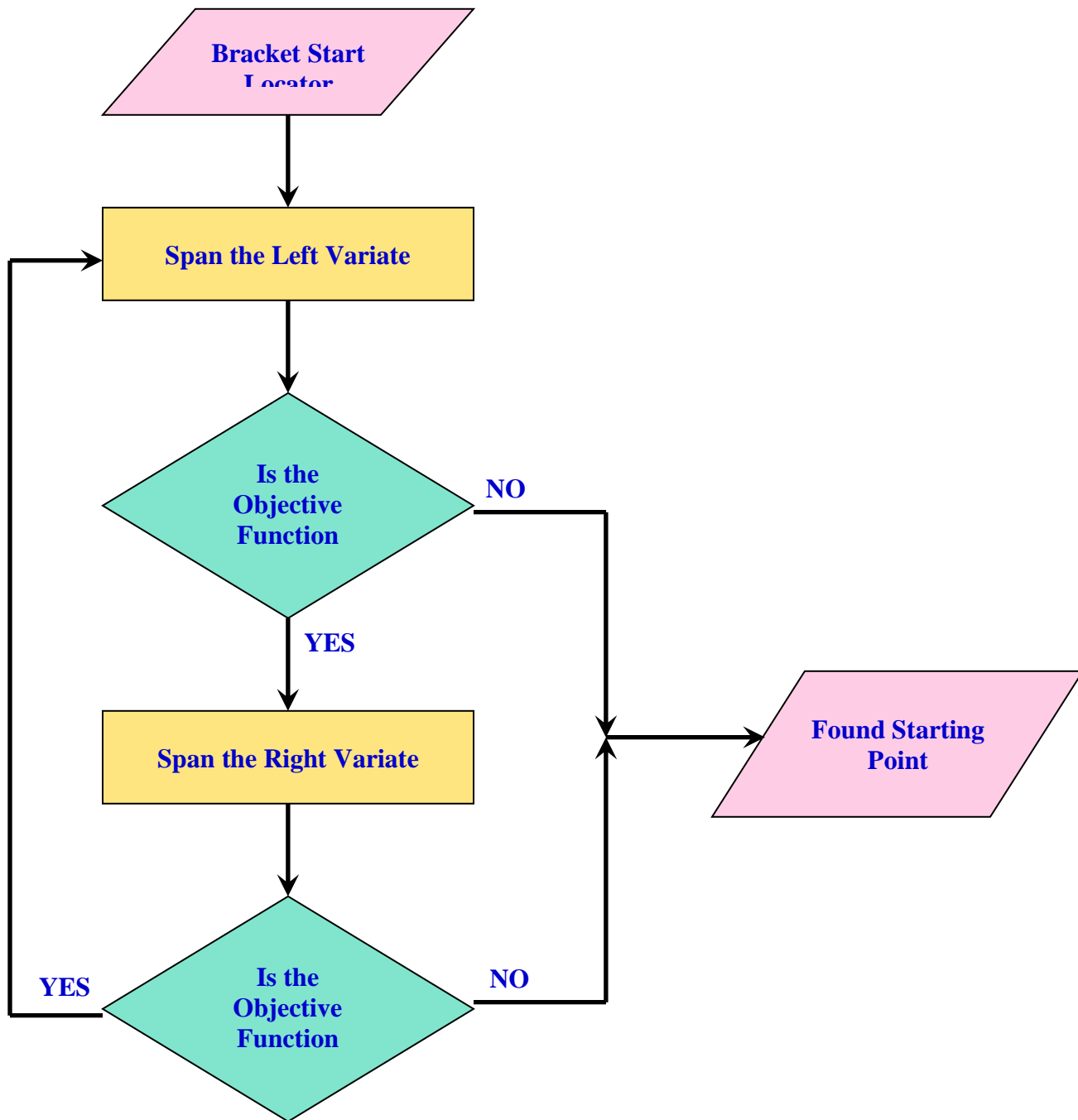
**Figure #6**
**Bracketing when Objective Function**
**Validity is Range-bound**

**Figure #7**
**Objective Function Fixed Point Bracketing**

Objective Function

Bracketing Iteration #1

Bracketing Iteration #2

Bracketing Iteration #3

Bracketing Iteration #4

Bracketing Iteration #5

Final Brackets

Variate

$-v_3$ $\quad$ $-v_2$ $\quad$ $-v_1$ $\quad$ $v_1$ $\quad$ $v_2$ $\quad$ $v_3$